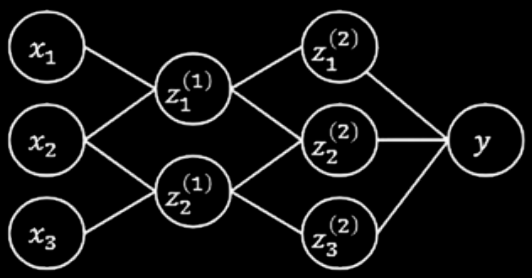
Neural Networks, Mixture Models, and Gibbs Sampling

Razeen Wasif

May 11th, 2025

1 Section 1: Back-propagation

Figure 1 shows a neural network with two hidden layers. The first hidden layer has two hidden nodes, $z_1^{(1)}$ and $z_2^{(1)}$, and the second hidden layer has three hidden nodes, $z_1^{(2)}, z_2^{(2)}$ and $z_3^{(2)}$. Note that this neural network is not fully connected.



Suppose that we are using this neural network for a binary classification problem. The output node y uses the sigmoid function shown below as its activation function.

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{1}$$

Let $t \in \{0, 1\}$ be a binary variable that denotes the final class label. The output value y given input values x_1, x_2, x_3 is interpreted as the probability $p(t = 1 \mid x_1, x_2, x_3)$. Let us further assume that the two hidden layers use the tanh function as their activation functions, which is defined below:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{2}$$

Let $w_{jk}^{(l)}$ denote the weight connection between the nodes $z_j^{(l-1)}$ (node j of previous layer) and z_k^l (node k of current layer), where the node $z_j^{(0)}$ is defined to be x_j and the node $z_1^{(3)}$ is y.

1.1 Question 1.1: Parameters of the model

How many parameters (including both weights and bias parameters) does this neural network model have? Show your final answer as an integer and explain what these parameters are, i.e., how you have arrived at your final answer
Solution:

- Parameters between Input and Hidden Layer 1:
  - Weights:
    - $z_1^{(1)} \rightarrow x_1$ & x_2 so 2 weights
    - $z_2^{(1)} \rightarrow x_2$ & x_3 so 2 weights
  - Biases:
    - Each node in hidden layer 1 has one bias param so 2 biases
- Parameters between Hidden Layer 1 and Hidden Layer 2:

  - Weights:
    * $z_1^{(1)} \rightarrow z_1^{(2)}1$ & $z_2^{(2)}$ so 2 weights
    * $z_2^{(1)} \rightarrow z_2^{(2)}$ & $z_3^{(2)}$ so 2 weights
  - Biases:
    * Each node in hidden layer 2 has one bias param so 3 biases
- Parameters between Hidden Layer 2 and Output:
  - Weights:
    * $y \rightarrow z_1^{(2)}1$ & $z_2^{(2)}$ & $z_3^{(2)}$ so 3 weights
  - Biases:
    * The output node has 1 bias

$\therefore$ The total number of parameters are $2+2+2+2+2+3+3+1 = 17$ parameters

1.2 Question 1.2: $\frac{\partial \mathcal{L}}{\partial w_{11}^{(3)}}$

[1] Let us assume that we have a set of N training data denoted as $\{x_{i1}, x_{i2}, x_{i3}, t_i\}_{i=1}^N$, where $t_i \in \{0, 1\}$. Let us use the cross-entropy loss over the entire training set as the loss function:

$$\mathcal{L} = -\sum_{i=1}^{N} \Big(t_i \log\, p(t=1 \mid x_{i1}, x_{i2}, x_{i3}) + (1-t_i) \log\, p(t=0 \mid x_{i1}, x_{i2}, x_{i3}) \Big) \tag{3}$$

you can assume that a forward pass has taken place and all the nodes have their output values computed and stored. You can use $z_{ij}^{(l)}$ to denote the output value of the node $z_j^{(l)}$ for the i-th data point. Similarly, you can use y_i to denote the output value of the node y for the i-th data point.
With the definition of $\mathcal{L}$ in Eqn. 3, show how $\frac{\partial \mathcal{L}}{\partial w_{11}^{(3)}}$ is expressed as a function in terms of the t_i's, the output values of the various nodes (y_i's, $z_{i1}^{(2)}$'s, etc.), and the values of the weight parameters (i.e., $w_{jk}^{(l)}$'s).
Solution:

$$\frac{\partial \mathcal{L}}{\partial w_{11}^{(3)}} = \partial - \sum_{i=1}^{N} \Big(t_i \log\, y_i + (1-t_i) \log\, (1-y_i) \Big) \tag{4}$$

$$\tag{5}$$

$w_{11}^{(3)}$ signifies the weight between node 1 in layer 2 and node 1 in layer 3 which is $z_1^{(2)}$ and y respectively.
Suppose the total input y receives is $\alpha_i^{(3)}$. When $w_{11}^{(3)}$ changes, $\alpha_i^{(3)}$ changes by exactly $z_{i1}^{(2)}$ which would mean

$$\partial \alpha_i^{(3)} / \partial w_{11}^{(3)} = z_{i1}^{(2)}. \tag{6}$$

The y neuron takes the input $\alpha_i^{(3)}$ and passes it through the sigmoid activation function so $y_i = \sigma(\alpha_i^{(3)})$. To find out how much y_i changes when $\alpha_i^{(3)}$ changes, we can make use of the derivative of the sigmoid function ($\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x))$) such that we have

$$\frac{\partial y_i}{\partial \alpha_i^{(3)}} = y_i \cdot (1 - y_i) \tag{7}$$

Since y_i affects the loss $\mathcal{L}$, we want to know how much $\mathcal{L}$ changes when y_i changes.

$$\mathcal{L}_i = -t_i \log(y_i) - (1-t_i) \log(1-y_i) \tag{8}$$

$$\frac{\partial \mathcal{L}_i}{\partial y_i} = -t_i \cdot \frac{1}{y_i} - (1-t_i) \cdot \frac{1}{1-y_i} \cdot -1 \tag{9}$$

$$\frac{\partial \mathcal{L}_i}{\partial y_i} = -\frac{t_i}{y_i} + \frac{1-t_i}{1-y_i} \tag{10}$$

$$= \frac{-t_i(1-y_i) + y_i(1-t_i)}{y_i(1-y_i)} \tag{11}$$

$$= \frac{-t_i + t_i y_i + y_i - t_i y_i}{y_i(1-y_i)} \tag{12}$$

$$= \frac{y_i - t_i}{y_i(1-y_i)} \tag{13}$$

Thus we have determined that the weight $w_{11}^{(3)}$ influences $\mathcal{L}$ in the following way for each data point i:
$w_{11}^{(3)} \to \alpha_i^{(3)} \to y_i \to \mathcal{L}$
Therefore, over all N data points, $\partial\mathcal{L}/\partial w_{11}^{(3)} = \sum_i^N (\partial\mathcal{L}/\partial y_i) \times (\partial y_i/\partial \alpha_i^{(3)}) \times (\partial \alpha_i^{(3)}/\partial w_{11}^{(3)})$
Calculating each part we get:

$$\sum_i^N \frac{y_i - t_i}{y_i(1-y_i)} \times y_i \cdot (1-y_i) \times z_{i1}^{(2)}. \tag{14}$$

$$= \sum_i^N (y_i - t_i) \times z_{i1}^{(2)} \tag{15}$$

$(y_i - t_i)$ is the error which is multiplied by the output of the neuron that fed into this weight which being $z_{i1}^{(2)}$. The sum tells us the overall sensitivity of the total loss $\mathcal{L}$ to changes in the weight $w_{11}^{(3)}$. If the value is large and positive, increasing $w_{11}^{(3)}$ increases the loss. If it's large and negative, increasing $w_{11}^{(3)}$ decreases the loss and this is what is needed for gradient descent.

1.3 Question 1.3: $\frac{\partial\mathcal{L}}{\partial w_{22}^{(2)}}$

Solution: $w_{22}^{(2)}$ signifies the weight between node 2 in layer 1 and node 2 in layer 2 which is $z_2^{(1)}$ and $z_2^{(2)}$ respectively.
The weight $w_{22}^{(2)}$ influences the loss $\mathcal{L}$ through the following path for each data point i:
$w_{22}^{(2)} \to \alpha_{2i}^{(2)} \to z_{2i}^{(2)} \to \alpha_i^{(3)} \to y_i \to \mathcal{L}$ (where $\alpha_{2i}^{(2)}$ is the input to $z_2^{(2)}$)
The derivative $\frac{\partial\mathcal{L}}{\partial w_{22}^{(2)}}$ will be a sum over all N data points as such:

$$\frac{\partial\mathcal{L}}{\partial w_{22}^{(2)}} = \sum_i^N \frac{\partial\mathcal{L}}{\partial y_i} \times \frac{\partial y_i}{\partial \alpha_i^{(3)}} \times \frac{\partial \alpha_i^{(3)}}{\partial z_{2i}^{(2)}} \times \frac{\partial z_{2i}^{(2)}}{\partial \alpha_{2i}^{(2)}} \times \frac{\partial \alpha_{2i}^{(2)}}{\partial w_{22}^{(2)}} \tag{16}$$

Now we can calculate each part:

Note the derivative of tanh(x) w.r.t x is $1 - tanh^2(x)$.

Note α_i^3 is the input to y and y is connected to z_1^2,z_2^2,z_3^2.

$$\frac{\partial\mathcal{L}}{\partial y_i} = \frac{y_i - t_i}{y_i(1-y_i)} \tag{17}$$

$$\frac{\partial y_i}{\partial \alpha_i^{(3)}} = y_i(1-y_i) \tag{18}$$

$$\frac{\partial \alpha_i^{(3)}}{\partial z_{2i}^{(2)}} = \frac{\partial(w_{11}^{(3)} z_{1i}^{(2)} + w_{21}^{(3)} z_{2i}^{(2)} + w_{31}^{(3)} z_{3i}^{(2)} + b_1^{(3)})}{\partial z_{2i}^{(2)}} = w_{21}^{(3)} \tag{19}$$

$$\frac{\partial z_{2i}^{(2)}}{\partial \alpha_{2i}^{(2)}} = 1 - (z_{2i}^{(2)})^2 \tag{20}$$

$$\frac{\partial \alpha_{2i}^{(2)}}{\partial w_{22}^{(2)}} = \frac{\partial(w_{12}^{(2)} z_{1i}^{(1)} + w_{22}^{(2)} z_{2i}^{(1)} + b_2^{(2)})}{\partial z_{2i}^{(2)}} = z_{2i}^{(1)} \tag{21}$$

Now we can combine all the parts:

$$\frac{\partial\mathcal{L}}{\partial w_{22}^{(2)}} = \sum_i^N \frac{y_i - t_i}{y_i(1-y_i)} \times y_i(1-y_i) \times w_{21}^{(3)} \times 1 - (z_{2i}^{(2)})^2 \times z_{2i}^{(1)} \tag{22}$$

$$= \sum_i^N y_i - t_i \times w_{21}^{(3)} \times 1 - (z_{2i}^{(2)})^2 \times z_{2i}^{(1)} \tag{23}$$

1.4 Question 1.4: $\frac{\partial\mathcal{L}}{\partial w_{21}^{(1)}}$

Solution:
$w_{21}^{(1)}$ is the weight connecting the input x_2 to the node $z_1^{(1)}$.
The weight $w_{21}^{(1)}$ influences the loss L for each data point i through the following path:

$w_{21}^{(1)} \to a_{1i}^{(1)} \to z_{1i}^{(1)}$ where $a_{1i}^{(1)}$ is the input to $z_1^{(1)}$ and $z_{1i}^{(1)}$ is the output of $z_1^{(1)}$.
From $z_{1i}^{(1)}$, the influence branches out because $z_1^{(1)}$ connects to multiple nodes in the second hidden layer ($z_1^{(2)}$ and $z_2^{(2)}$). Both of these paths then converge towards y_i and then $\mathcal{L}$.

The derivative $\frac{\partial \mathcal{L}}{\partial w_{21}^{(1)}}$ will be a sum over all N data points. For each data point i:

$$\frac{\partial \mathcal{L}_i}{\partial w_{21}^{(1)}} = \sum \left(\frac{\partial \mathcal{L}_i}{\partial z_{1i}^{(1)}} \right) \times \left(\frac{\partial z_{1i}^{(1)}}{\partial a_{1i}^{(1)}} \right) \times \left(\frac{\partial a_{1i}^{(1)}}{\partial w_{21}^{(1)}} \right) \tag{24}$$

$\frac{\partial a_{1i}^{(1)}}{\partial w_{21}^{(1)}}$:

$$a_{1i}^{(1)} = w_{11}^{(1)} x_{1i} + w_{21}^{(2)} x_{2i} + b_1^{(1)} \tag{25}$$

$$\therefore \frac{\partial a_{1i}^{(1)}}{\partial w_{21}^{(1)}} = x_{2i} \tag{26}$$

$\frac{\partial z_{1i}^{(1)}}{\partial a_{1i}^{(1)}}$:

$$z_{1i}^{(1)} = \tanh(a_{1i}^{(1)}) \tag{27}$$

$$\therefore \frac{\partial z_{1i}^{(1)}}{\partial a_{1i}^{(1)}} = 1 - \tanh^2(a_{1i}^{(1)}) = 1 - (z_{1i}^{(1)})^2 \tag{28}$$

$\frac{\partial L_i}{\partial z_{1i}^{(1)}}$: ($z_1^{(1)}$ connects to $z_1^{(2)}$ and $z_2^{(2)}$)

$$\frac{\partial L_i}{\partial z_{1i}^{(1)}} = \frac{\partial \mathcal{L}_i}{\partial a_{1i}^{(2)}} \times \frac{\partial a_{1i}^{(2)}}{\partial z_{1i}^{(1)}} + \frac{\partial \mathcal{L}_i}{\partial a_{2i}^{(2)}} \times \frac{\partial a_{2i}^{(2)}}{\partial z_{1i}^{(1)}} \tag{29}$$

$$= \frac{\partial \mathcal{L}_i}{\partial a_{1i}^{(2)}} \times w_{11}^{(2)} + \frac{\partial \mathcal{L}_i}{\partial a_{2i}^{(2)}} \times w_{12}^{(2)} \tag{30}$$

$\frac{\partial \mathcal{L}_i}{\partial a_{1i}^{(2)}}$:

$$\frac{\partial \mathcal{L}_i}{\partial a_{1i}^{(2)}} = \frac{\partial \mathcal{L}_i}{\partial z_{1i}^{(2)}} \times \frac{\partial z_{1i}^{(2)}}{\partial a_{1i}^{(2)}} \tag{31}$$

$$\frac{\partial z_{1i}^{(2)}}{\partial a_{1i}^{(2)}} = 1 - (z_{1i}^{(2)})^2 \tag{32}$$

$$\frac{\partial \mathcal{L}_i}{\partial z_{1i}^{(2)}} = \frac{\partial \mathcal{L}_i}{\partial a_i^{(3)}} \times \frac{\partial a_i^{(3)}}{\partial z_{1i}^{(2)}} \tag{33}$$

from previously, we know that (34)

$$\frac{\partial \mathcal{L}_i}{\partial a_i^{(3)}} = y_i - t_i \tag{35}$$

$$\frac{\partial a_i^{(3)}}{\partial z_{1i}^{(2)}} = w_{11}^{(3)} \tag{36}$$

$$\therefore \frac{\partial \mathcal{L}_i}{\partial a_{1i}^{(2)}} = (y_i - t_i) \times w_{11}^{(3)} \times (1 - (z_{1i}^{(2)})^2) \tag{37}$$

$\frac{\partial \mathcal{L}_i}{\partial a_{2i}^{(2)}}$:

$$\frac{\partial \mathcal{L}_i}{\partial a_{2i}^{(2)}} = \frac{\partial \mathcal{L}_i}{\partial z_{2i}^{(2)}} \times \frac{\partial z_{2i}^{(2)}}{\partial a_{2i}^{(2)}} \tag{38}$$

$$\frac{\partial z_{2i}^{(2)}}{\partial a_{2i}^{(2)}} = 1 - (z_{2i}^{(2)})^2 \tag{39}$$

$$\frac{\partial \mathcal{L}_i}{\partial z_{2i}^{(2)}} = \frac{\partial \mathcal{L}_i}{\partial a_i^{(3)}} \times \frac{\partial a_i^{(3)}}{\partial z_{2i}^{(2)}} = (y_i - t_i) \times w_{21}^{(3)} \times (1 - (z_{2i}^{(2)})^2) \tag{40}$$

$$\tag{41}$$

Now we can substitute $\frac{\partial \mathcal{L}_i}{\partial a_{1i}^{(2)}}$ and $\frac{\partial \mathcal{L}_i}{\partial a_{2i}^{(2)}}$ back into $\frac{\partial L_i}{\partial z_{1i}^{(1)}}$:

$$\frac{\partial L_i}{\partial z_{1i}^{(1)}} = (y_i - t_i) \times w_{11}^{(3)} \times (1 - (z_{1i}^{(2)})^2) \times w_{11}^{(2)} + (y_i - t_i) \times w_{21}^{(3)} \times (1 - (z_{2i}^{(2)})^2) \times w_{12}^{(2)} \tag{42}$$

$$= (y_i - t_i) \times [w_{11}^3 w_{11}^2 \cdot (1 - (z_{1i}^{(2)})^2) + w_{21}^3 w_{12}^2 (1 - (z_{2i}^{(2)})^2)] \tag{43}$$

Finally substitute everything into $\frac{\partial \mathcal{L}_i}{\partial w_{21}^{(1)}}$:

$$\frac{\partial \mathcal{L}_i}{\partial w_{21}^{(1)}} = \sum_i^N (y_i - t_i) \times [w_{11}^3 w_{11}^2 \cdot (1 - (z_{1i}^{(2)})^2) + w_{21}^3 w_{12}^2 (1 - (z_{2i}^{(2)})^2)] \times (1 - (z_{1i}^1)^2) \times x_{2i} \tag{44}$$

$$\tag{45}$$

1.5 Question 1.5: Backpropagation

Using the computation of $\frac{\partial \mathcal{L}_i}{\partial w_{21}^{(1)}}$ in the network above as an example, explain why in general the computation of $\frac{\partial \mathcal{L}}{\partial w_{jk}^{(l)}}$ should start from the last layer rather than the first layer to save computational costs. **Solution:**
By calculating the error signal from the last layer to the front, the algorithm avoids the redundancy of re-evaluating long, shared chains of partial derivatives that would occur if we tried to compute each weight's gradient independently from the first layer instead.

2 Section 2: Neural Network for Regression

In this question, we will build a simple fully-connected neural network for two simple regression problems. Besides σ and tanh, another commonly used activation function is the Rectified Linear Unit (ReLU) function, as defined below: $\text{ReLU}(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$

2.1 Question 2.1: ReLU

Implemented ReLU class

2.2 Question 2.2: Squared error loss function

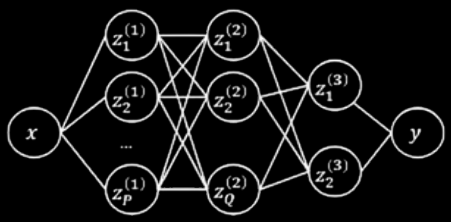
For our regression problems, let us use the mean squared error as our loss function, which is defined as follows:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} (t_i - y_i)^2, \tag{46}$$

where t_i is the observed output value given input x_i and y_i is the predicted output value based on the neural network, and we assume that there are in total N data points in the training dataset.

2.3 Question 2.3: Building the neural network

We will use a fully-connected neural network as shown in Figure 2 for this question. This network has three hidden layers, where the first hidden layer has P hidden nodes, the second hidden layer has Q hidden nodes, and the third (last) hidden layer has exactly 2 hidden nodes. The activation function to be used is ReLU for all hidden nodes. The input layer has a single node, and the output layer also has a single node. The output node does not use any activation function.

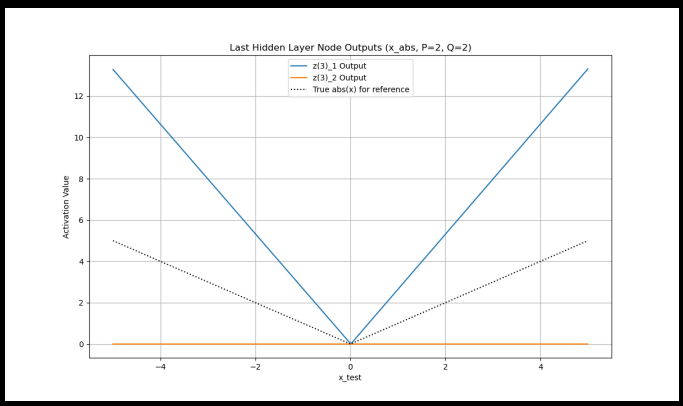


2.4 Question 2.4: Training the neural network

Next, let us consider two regression tasks. In the first regression task, we try to use the neural network to fit the curve $y = x^2$, whereas in the second one, we try to use the neural network to fit the curve $y = \text{abs}(x)$.
Plots can be found in the Appendix

2.5 Question 2.5: Visualising the hidden nodes in the last hidden layer

The last hidden layer of your network always has exactly 2 hidden nodes. Pick one of the configurations where your network has fit the curve $y = \text{abs}(x)$ well. Plot the output values of the two hidden nodes z_1^3 and z_2^3 over the input value x. Label the two curves as z(3) 1 and z(3) 2, respectively. Include this plot in your report. Based on your plot, what do you expect the two weights connecting z_1^3 to y and connecting z_2^3 to y to be? Print out the values of the two weights and discuss whether your guess roughly matches the actual values of these two weights.
Solution:
Based on the ideal decomposition abs(x) = ReLU(x) + ReLU(-x), we would expect one hidden node ($z(3)_1$) to approximate ReLU(x) and the other hidden node ($z(3)_2$) to approximate ReLU(-x) and the weights connecting these to the output y to be approximately +1 each.



Although the abs(x) function can be decomposed into ReLU(x) + ReLU(-x), my trained network with Q=2 instead learned to approximate the full abs(x) behaviour using a single ReLU node (z(3)_1). This node outputs a V-shaped curve resembling abs(x), while the other node (z(3)_2) remained inactive.
Despite the second weight being nonzero, its contribution is nullified due to the zero output of z(3)_2. This highlights the network's flexibility: instead of combining two opposing ReLU paths, it optimized a more efficient representation by shaping a single ReLU neuron to fit the entire abs(x) curve.
Actual weights from the last hidden layer (z(3)) to the output node y:
Weight for $z(3)_1 \rightarrow$ y: 0.3714

Weight for $z(3)_2 \rightarrow$ y: -0.8778

3 Section 3: Mixture Models and the EM Algorithm

[2] You are given a data file called `articles.txt` that contains a set of news articles. Each line of the file contains a single news article, where the first column is the body of the article, the second column is the headline of the article, and the third column is the category of the article. The three columns are separated by tabs. For this question, we will use only the first two columns of the file, i.e., the body and the headline of each article.

We will represent each news article as two sequences of words: $(w^b_{i,1}, w^b_{i,2}, \ldots, w^b_{i,M_i})$ are the words found in the body of the i-th article, where M_i is the number of words in the i-th article's body, and $(w^h_{i,1}, w^h_{i,2}, \ldots, w^h_{i,L_i})$ are the words found in the headline of the i-th article, where L_i is the number of words in the i-th article's headline. We ignore words that appear frequently in many articles but are usually not representative of any topic, such as "a", "an", "the", "and", "or", "to", and "of". These words are called "stopwords" in language processing. For convenience, after stopwords are removed, each remaining word is mapped to an ID (which is an an integer) ranging from 1 to V, where V is the vocabulary size (i.e., the number of unique words in the dataset). Note that we have provided you with the pre-processing code, so you do not need to perform stopword removal or the mapping from words to the word IDs. For the rest of the discussion, we will assume that $w^b_{i,j}$'s and $w^h_{i,j}$'s are word IDs, i.e., integers in $[1, V]$.

We assume that the words in the body of an article are generated from a mixture of K categorical distributions (a.k.a. generalised Bernoulli distributions or multinoulli distributions). Similarly, the words in the headline of an article are also generated from a mixture of (a different set of) K categorical distributions. Let us use $\{\theta^b_k\}^K_{k=1}$ to denote the parameters of the K categorical distributions for the article bodies, where θ^b_k is a categorical distribution over V outcomes. Specifically, θ^b_k is a V-dimensional vector where $\theta^b_{k,v} \in [0, 1]$ is the probability of selecting the word with the ID v ($1 \leq v \leq V$) from this categorical distribution. Similarly, we will use $\{\theta^h_k\}^K_{k=1}$ to denote the parameters of the K categorical distributions for the article headlines. These categorical distributions over the vocabulary are commonly referred to as "topics". For the rest of this question, we will refer to $\{\theta^b_k\}^K_{k=1}$ as the K "body topics" and $\{\theta^h_k\}^K_{k=1}$ as the K "headline topics".

The mixing coefficients are different for different articles. Specifically, we will use π_i, a K-dimensional vector, to denote the mixing coefficients for the i-th article, where $\pi_{i,k} \in [0, 1]$ denotes the probability to select the k-th body (or headline) topic among the K body (or headline) topics. π_i's are essentially the parameters of a categorical distribution over K outcomes.

We assume that for each word $w^b_{i,j}$ in the body of the i-th article, there is a hidden variable $z^b_{i,j}$ associated with it. $z^b_{i,j}$ is a discrete variable that takes a value between 1 and K, denoting the topic chosen for that word. $z^b_{i,j}$ follows the categorical distribution parameterised by π_i. Given a value of $z^b_{i,j}$, the word $w^b_{i,j}$ follows the categorical distribution parameterised by $\theta^b_{z^b_{i,j}}$.

Similarly, for each word $w^h_{i,j}$ in the headline of the i-th article, there is a corresponding hidden variable $z^h_{i,j}$, which also follows the categorical distribution parameterised by π_i. Given a value of $z^h_{i,j}$, the word $w^h_{i,j}$ follows the categorical distribution parameterised by $\theta^h_{z^h_{i,j}}$.

3.1 Question 3.1: The log likelihood function

Based on the description of the mixture model, give the formula of the log likelihood of the N observed news articles $\{((w^b_{i,1}, w^b_{i,2}, \ldots, w^b_{i,M_i},), (w^h_{i,1}, w^h_{i,2}, \ldots, w^h_{i,L_i},))\}^N_{i=1}$. Show the derivation of your formula.

Solution:

The overall likelihood $L(\theta|X)$ where $\theta = (\pi_i, \theta^b_k, \theta^h_k)$ is the probability of observing all the data given the parameters. Assuming articles are independent:

$$L(\theta|X) = P(X|\theta) = \prod_{i=1}^{N} logP(W^b_i, W^h_i|\pi_i, \theta^b, \theta^h) \tag{47}$$

The log likelihood is:

$$\text{log-likelihood}(\theta|X) = \text{log}P(X|\theta) = \sum_{i=1}^{N} logP(W^b_i, W^h_i|\pi_i, \theta^b, \theta^h) \tag{48}$$

A signle article i is:

$$P(W^b_i, W^h_i|\pi_i, \theta^b, \theta^h) \tag{49}$$

the question states that the topic for each word ($z^b_{i,j}$ and $z^h_{i,l}$) are drawn according to π_i which means the generation of body words and headline words are independent processes.

$$P(W^b_i, W^h_i|\pi_i, \theta^b, \theta^h) = P(W^b_i|\pi_i, \theta^b) \times P(W^h_i|\pi_i, \theta^h) \tag{50}$$

Words within the body are conditionally independent given their individual topic choices

$$P(W^b_i|\pi_i, \theta^b) = \prod_{j=1}^{M_i} P(w^b_{i,j}|\pi_i, \theta^b) \tag{51}$$

Considering a single body word $w^b_{i,j}$, we don't observe it's topic $z^b_{i,j}$. To get the probability of observing $w^b_{i,j}$, we must sum over all possible topic k that could have generated it:

$$P(w^b_{i,j}|\pi_i, \theta^b) = \sum_{k=1}^{K} P(w^b_{i,j}, z^b_{i,j} = k|\pi_i, \theta^b) \tag{52}$$

Here we can use the chain rule of probability $P(A, B) = P(A|B)P(B)$ to get:

$$P(w^b_{i,j}|\pi_i, \theta^b) = \sum_{k=1}^{K} P(w^b_{i,j}|z^b_{i,j} = k, \pi_i, \theta^b) \times P(z^b_{i,j} = k|\pi_i, \theta^b) \tag{53}$$

Given $z^b_{i,j} = k$, the word generation only depends on θ^b_k and $P(z^b_{i,j} = k|\pi_i, \theta^b)$ only depends on π_i:

$$P(w^b_{i,j}|\pi_i, \theta^b) = \sum_{k=1}^{K} P(w^b_{i,j}|z^b_{i,j} = k, \theta^b_k) \times P(z^b_{i,j} = k|\pi_i) \tag{54}$$

$$= \sum_{k=1}^{K} \theta^b_{k,w^b_{i,j}} \times \pi_{i,k} \tag{55}$$

So for all body words in article i:

$$P(W^b_i|\pi_i, \theta^b) = \prod_{j=1}^{M_i} \left[\sum_{k=1}^{K} \pi_{i,k} \theta^b_{k,w^b_{i,j}} \right] \tag{56}$$

Similarly, for all headline words in article i (using l as the index for headline words):

$$P(W^h_i|\pi_i, \theta^h) = \prod_{l=1}^{L_i} \left[\sum_{k=1}^{K} \pi_{i,k} \theta^h_{k,w^h_{i,l}} \right] \tag{57}$$

Substituting these back into the log-likelihood equation for a single article:

$$\log P(W^b_i, W^h_i|\pi_i, \theta^b, \theta^h) = \log \left[\left(\prod_{j=1}^{M_i} \left[\sum_{k=1}^{K} \pi_{i,k} \theta^b_{k,w^b_{i,j}} \right] \right) \times \left(\prod_{l=1}^{L_i} \left[\sum_{k=1}^{K} \pi_{i,k} \theta^h_{k,w^h_{i,l}} \right] \right) \right] \tag{58}$$

$$\log \left(\prod_{j=1}^{M_i} \left[\sum_{k=1}^{K} \pi_{i,k} \theta^b_{k,w^b_{i,j}} \right] \right) + \log \left(\prod_{l=1}^{L_i} \left[\sum_{k=1}^{K} \pi_{i,k} \theta^h_{k,w^h_{i,l}} \right] \right) \tag{59}$$

$$\sum_{j=1}^{M_i} \log \left[\sum_{k=1}^{K} \pi_{i,k} \theta^b_{k,w^b_{i,j}} \right] + \sum_{l=1}^{L_i} \log \left[\sum_{k=1}^{K} \pi_{i,k} \theta^h_{k,w^h_{i,l}} \right] \tag{60}$$

Summing over all N articles for the total log-likelihood:

$$\text{log-likelihood}(\theta|D) = \sum_{i=1}^{N} \left\{ \sum_{j=1}^{M_i} log \left[\sum_{k=1}^{K} \pi_{i,k} \theta^b_{k,w^b_{i,j}} \right] + \sum_{l=1}^{L_i} log \left[\sum_{k=1}^{K} \pi_{i,k} \theta^h_{k,w^h_{i,l}} \right] \right\} \tag{61}$$

3.2 Question 3.2: The E-Step and M-step

Derive the EM algorithm for the above mixture model.

- Complete Data Log-Likelihood $LL(\theta|X,Z)$
  If we knew the topic assignments $Z = (z^b_{i,j}, z^h_{i,l})$, the likelihood of the data and these assignments would be:

$$P(X,Z|\theta) = \prod_{i=1}^{N}\prod_{j=1}^{M_i} P(w^b_{i,j}, z^b_{i,j}|\pi_i, \theta^b) \times \prod_{i=1}^{N}\prod_{l=1}^{L_i} P(w^h_{i,l}, z^h_{i,l}|\pi_i, \theta^h) \tag{62}$$

$$P(D,Z|\theta) = \prod_{i=1}^{N}\prod_{j=1}^{M_i} [P(w^b_{i,j}|z^b_{i,j}, \theta^b)P(z^b_{i,j}|\pi_i)] \times \prod_{i=1}^{N}\prod_{l=1}^{L_i} [P(w^h_{i,l}|z^h_{i,l}, \theta^h)P(z^h_{i,l}|\pi_i)] \tag{63}$$

$$P(D,Z|\theta) = \prod_{i=1}^{N}\prod_{j=1}^{M_i} [\theta^b_{z^b_{i,j}, w^b_{i,j}} \times \pi_{i,z^b_{i,j}}] \times \prod_{i=1}^{N}\prod_{l=1}^{L_i} [\theta^h_{z^h_{i,l}, w^h_{i,l}} \times \pi_{i,z^h_{i,l}}] \tag{64}$$

  The complete data log likelihood using $\delta(z = k)$ as an indicator function:

$$LL(\theta|X,Z) = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\sum_{k=1}^{K} \delta(z^b_{i,j} = k)[log\pi_{i,k} + log\theta^b_{k,w^b_{i,j}}] + \sum_{i=1}^{N}\sum_{l=1}^{L_i}\sum_{k=1}^{K} \delta(z^h_{i,l} = k)[log\pi_{i,k} + log\theta^h_{k,w^h_{i,l}}] \tag{65}$$

- E-Step: Calculate expectation of LL
  The expectation involves replacing the indicator functions $\delta(z = k)$ with their posterior probabilities given the observed data D and the current parameters $\theta^{(old)}$

$$Q(\theta|\theta^{(old)}) = E_{Z|X,\theta^{(old)}}[LL(\theta|X,Z)] \tag{66}$$

  Let $\gamma^b_{i,j}(k)^{(old)} = P(z^b_{i,j} = k|w^b_{i,j}, \pi_i^{(old)}, \theta^{(b,old)})$ be the posterior probability that body word $w^b_{i,j}$ in article i was generated by topic k, given current parameters:

$$\gamma^b_{i,j}(k)^{(old)} = \frac{\pi^{(old)}_{i,k} \times \theta^{(b,old)}_{k,w^b_{i,j}}}{\sum_{s=1}^{K} \pi^{(old)}_{i,s} \times \theta^{(b,old)}_{s,w^b_{i,j}}} \tag{67}$$

  Similarly for headline words:
  Let $\gamma^h_{i,l}(k)^{(old)} = P(z^h_{i,l} = k|w^h_{i,l}, \pi_i^{(old)}, \theta^{(h,old)})$

$$\gamma^h_{i,l}(k)^{(old)} = \frac{\pi^{(old)}_{i,k} \times \theta^{(h,old)}_{k,w^h_{i,l}}}{\sum_{s=1}^{K} \pi^{(old)}_{i,s} \times \theta^{(h,old)}_{s,w^h_{i,l}}} \tag{68}$$

  The Q function becomes:

$$Q(\theta|\theta^{(old)}) = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\sum_{k=1}^{K} \gamma^b_{i,j}(k)^{(old)}[log\ \pi_{i,k} + log\ \theta^b_{k,w^b_{i,j}}] + \sum_{i=1}^{N}\sum_{l=1}^{L_i}\sum_{k=1}^{K} \gamma^h_{i,l}(k)^{(old)}[log\ \pi_{i,k} + log\ \theta^h_{k,w^h_{i,l}}] \tag{69}$$

- M-Step: Maximize $Q(\theta|\theta^{(old)})$ w.r.t to θ
  Maximize Q for $\pi_{i,k}, \theta^b_{k,v},$ and $\theta^h_{k,v}$.
  - a) Update for $\pi_{i,k}$:
    maximize the part of Q that depends on $\pi_{i,k}$ for a specific article i, subject to $\sum_k \pi_{i,k} = 1$.
    Terms involving $\pi_{i,k}$:

$$Q_\pi i = \sum_{j=1}^{M_i}\sum_{k=1}^{K} \gamma^b_{i,j}(k)^{(old)} log\pi_{i,k} + \sum_{l=1}^{L_i}\sum_{k=1}^{K} \gamma^h_{i,l}(k)^{(old)} log\pi_{i,k} \tag{70}$$

    Using Lagrange multipliers:

$$L_\pi i = Q_\pi i + \lambda_i(\sum_k \pi_{i,k} - 1) \tag{71}$$

Taking $\partial L_\pi i/\partial \pi_{i,k} = 0$:

$$\left[\sum_{j=1}^{M_i} \frac{\gamma_{i,j}^b(k)^{(old)}}{\pi_{i,k}}\right] + \left[\sum_{l=1}^{L_i} \frac{\gamma_{i,l}^h(k)^{(old)}}{\pi_{i,k}}\right] + \lambda_i = 0 \tag{72}$$

$$\pi_{i,k} = -(1/\lambda_i) \times \left[\sum_{j=1}^{M_i} \gamma_{i,j}^b(k)^{(old)} + \sum_{l=1}^{L_i} \gamma_{i,l}^h(k)^{(old)}\right] \tag{73}$$

Summing over k and using $\sum_k \pi_{i,k} = 1$:

$$1 = -(1/\lambda_i) \times \sum_k \left[\sum_{j=1}^{M_i} \gamma_{i,j}^b(k)^{(old)} + \sum_{l=1}^{L_i} \gamma_{i,l}^h(k)^{(old)}\right] \tag{74}$$

Since $\sum_k \gamma_{i,j}^b(k)^{(old)} = 1$ and similarly for γ^h:

$$1 = -(1/\lambda_i) \times [M_i \times 1 + L_i \times 1] = -(1/\lambda_i) \times (M_i + L_i) \tag{75}$$

$$\frac{-1}{\lambda_i} = \frac{1}{(M_i + L_i)} \tag{76}$$

Substituting this back:

$$\pi_{i,k}^{(new)} = \frac{\sum_{j=1}^{M_i} \gamma_{i,j}^b(k)^{(old)} + \sum_{l=1}^{L_i} \gamma_{i,l}^h(k)^{(old)}}{(M_i + L_i)} \tag{77}$$

This is the expected number of times topic k is assigned to words in article i, divided by the total number of words in article i.

– b) Update for $\theta_{k,v}^b$ (prob of word v in body topic k):
Maximize the part of Q that depends on $\theta_{k,v}^b$ for a specific topic k, subject to $\sum_v \theta_{k,v}^b = 1$.
Terms involving θ_k^b:

$$Q_\theta kb = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \gamma_{i,j}^b(k)^{(old)} log\theta_{k,w_{i,j}^b}^b \tag{78}$$

This can be rewritten by explicitly summing over vocabulary words v using an indicator $\delta(w_{i,j}^b = v)$:

$$Q_\theta kb = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \sum_{v'=1}^{V} \delta(w_{i,j}^b = v') \gamma_{i,j}^b(k)^{(old)} \log \theta_{k,v'}^b \tag{79}$$

Then using Lagrange multipliers:

$$L_\theta kb = Q_\theta kb + \mu_k (\sum_{v'} \theta_{k,v'}^b - 1) \tag{80}$$

Taking $\frac{\partial L_\theta kb}{\partial \theta_{k,v}^b} = 0$:

$$\frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} \delta(w_{i,j}^b = v) \gamma_{i,j}^b(k)^{(old)}}{\theta_{k,v}^b} + \mu_k = 0 \tag{81}$$

$$\theta_{k,v}^b = -(1/\mu_k) \times \sum_{i=1}^{N} \sum_{j=1}^{M_i} \delta(w_{i,j}^b = v) \gamma_{i,j}^b(k)^{(old)} \tag{82}$$

Summing over v and using $\sum_v \theta_{k,v}^b = 1$:

$$1 = -\frac{1}{\mu_k} \times \sum_v \sum_{i=1}^{N} \sum_{j=1}^{M_i} \delta(w_{i,j}^b = v) \gamma_{i,j}^b(k)^{(old)} \tag{83}$$

$$1 = -\frac{1}{\mu_k} \times \sum_{i=1}^{N} \sum_{j=1}^{M_i} \gamma_{i,j}^b(k)^{(old)} \cdot \left(\sum_v \delta(w_{i,j}^b = v)\right) \tag{84}$$

Since $\sum_v \delta(w^b_{i,j} = v) = 1$ (each word is one specific vocab item):

$$1 = -\frac{1}{\mu_k} \times \sum_{i=1}^{N} \sum_{j=1}^{M_i} \gamma^b_{i,j}(k)^{(old)} \tag{85}$$

$$\frac{-1}{\mu_k} = \frac{1}{(\sum_{i'=1}^{N} \sum_{j'=1}^{M_{i'}} \gamma^b_{i',j'}(k)^{(old)})} \tag{86}$$

Substituting this back:

$$\theta^{(b,new)}_{k,v} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} \delta(w^b_{i,j} = v)\gamma^b_{i,j}(k)^{(old)}}{\sum_{i'=1}^{N} \sum_{j'=1}^{M_{i'}} \gamma^b_{i',j'}(k)^{(old)}} \tag{87}$$

This is the expected count of body word v being assigned to topic k, divided by the total expected count of any body word being assigned to topic k.

– c) Update for $\theta^h_{k,v}$ (prob of word v in headline topic k): The derivation is the same as for $\theta^b_{k,v}$:

$$\theta^{(h,new)}_{k,v} = \frac{\sum_{i=1}^{N} \sum_{l=1}^{L_i} \delta(w^h_{i,l} = v)\gamma^h_{i,l}(k)^{(old)}}{\sum_{i'=1}^{N} \sum_{l'=1}^{L_{i'}} \gamma^h_{i',l'}(k)^{(old)}} \tag{88}$$

This is the expected count of headline word v being assigned to topic k, divided by the total expected count of any headline word being assigned to topic k.

3.3 Question 3.3 & 3.4: Implementation & Results

Set K to 5, 10, and 20. Run the EM algorithm. For each configuration, show the top-10 words with the highest probabilities in each θ_{bk} and each θ_{hk} .
Solution:

- K = 5
  Top 10 words for each BODY topic:
  Topic 1: people, women, time, make, work, men, life, things, show, world
  Topic 2: school, students, schools, education, people, public, teachers, time, make, years
  Topic 3: trump, president, told, news, company, donald, house, year, time, team
  Topic 4: art, climate, work, world, people, percent, american, years, artists, city
  Topic 5: sexual, women, people, told, police, abuse, harassment, church, assault, years

  Top 10 words for each HEADLINE topic:
  Topic 1: women, jimmy, book, day, love, kimmel, white, people, men, time
  Topic 2: school, education, facebook, schools, teachers, students, teacher, world, life, space
  Topic 3: trump, donald, news, trumps, house, fox, olympics, video, colbert, michael
  Topic 4: art, climate, artist, video, american, change, artists, world, photos, years
  Topic 5: sexual, photos, uber, week, women, abuse, harassment, assault, accused, police

- K = 10:
  Top 10 words for each BODY topic:
  Topic 1: school, students, schools, education, people, teachers, children, time, public, make
  Topic 2: percent, year, company, million, business, companies, pay, public, people, state
  Topic 3: climate, world, change, years, science, water, research, people, human, scientists
  Topic 4: uber, company, art, time, people, work, united, car, china, war
  Topic 5: news, facebook, trump, media, time, president, games, book, told, fox
  Topic 6: women, people, men, sexual, white, church, religious, god, world, american
  Topic 7: police, told, shooting, years, nassar, game, time, team, home, found
  Topic 8: trump, president, house, state, sexual, campaign, people, white, law, women
  Topic 9: trump, company, president, facebook, data, information, told, security, news, users
  Topic 10: show, time, people, art, work, life, love, film, make, back

Top 10 words for each HEADLINE topic: Topic 1: school, education, schools, students, teachers, teacher, gun, high, kids, teaching
Topic 2: amazon, devos, pay, betsy, tax, million, jobs, money, business, plan
Topic 3: photos, climate, week, animal, video, change, world, scientists, weather, science
Topic 4: uber, trevor, noah, art, ceo, artists, scott, cars, car, people
Topic 5: news, olympics, olympic, trump, winter, fox, fake, pope, facebook, book
Topic 6: women, sexual, white, men, muslim, harvey, american, womens, metoo, church
Topic 7: police, shooting, nassar, larry, video, man, dead, woman, found, victims
Topic 8: trump, sexual, house, bill, gop, trumps, assault, harassment, white, clinton
Topic 9: trump, donald, facebook, colbert, trumps, report, michael, stephen, seth, meyers
Topic 10: art, show, artist, jimmy, video, star, women, watch, love, james

- K = 20:
Top 10 words for each BODY topic: Topic 1: schools, public, school, education, charter, students, state, tax, devos, money
Topic 2: facebook, company, uber, data, apple, users, nassar, information, google, security
Topic 3: told, trial, time, man, case, cosby, video, black, death, jury
Topic 4: game, team, players, games, sports, nfl, athletes, olympic, olympics, league
Topic 5: people, church, religious, god, world, faith, christian, american, jesus, white
Topic 6: percent, million, company, companies, billion, year, pay, bill, deal, market
Topic 7: people, work, time, make, business, company, things, job, employees, great
Topic 8: news, election, senate, campaign, democratic, republican, political, democrats, sen, ad
Topic 9: trump, president, donald, white, news, house, told, campaign, fox, election
Topic 10: told, huffpost, company, statement, times, news, media, employees, allegations, work
Topic 11: climate, water, change, energy, global, oil, environmental, world, gas, year
Topic 12: women, sexual, men, harassment, assault, woman, female, people, gender, girls
Topic 13: art, work, artists, artist, museum, world, painting, works, city, gallery
Topic 14: post, day, shared, time, family, people, life, back, love, pdt
Topic 15: police, gun, told, shooting, school, people, violence, found, home, student
Topic 16: people, music, muslim, american, country, white, song, black, time, muslims
Topic 17: book, show, people, time, film, story, love, things, make, life
Topic 18: students, school, education, teachers, schools, children, percent, people, teacher, high
Topic 19: law, state, court, federal, department, people, states, legal, justice, case
Topic 20: space, scientists, earth, research, years, science, researchers, time, nasa, planet

Top 10 words for each HEADLINE topic: Topic 1: devos, school, schools, charter, betsy, education, public, choice, tax, college
Topic 2: facebook, uber, nassar, apple, larry, ceo, data, photos, google, iphone
Topic 3: death, guilty, photo, cosby, bill, man, christmas, westworld, theory, charged
Topic 4: photos, olympics, animal, olympic, week, winter, nfl, video, game, noah
Topic 5: pope, church, christians, god, francis, christian, faith, religious, people, white
Topic 6: colbert, stephen, deal, billion, amazon, bill, company, trump, pay, wall
Topic 7: work, make, ways, business, things, wolf, life, people, good, takes
Topic 8: news, gop, primary, senate, fake, ad, candidate, democratic, cia, sinclair
Topic 9: trump, donald, trumps, news, fox, white, michael, house, president, seth
Topic 10: sexual, accused, harassment, misconduct, editor, media, times, ny, schneiderman, amid
Topic 11: climate, change, kimmel, jimmy, global, time, water, chinese, oil, video
Topic 12: women, sexual, assault, harassment, men, harvey, weinstein, metoo, womens, woman
Topic 13: art, artist, artists, museum, world, global, street, amazon, search, painting
Topic 14: jimmy, super, bowl, reveals, day, john, instagram, week, show, kids
Topic 15: police, shooting, gun, student, school, man, florida, parkland, dead, violence
Topic 16: muslim, music, american, protest, hate, america, message, day, made, ariana
Topic 17: book, read, netflix, dont, story, star, tale, books, show, watch
Topic 18: school, students, education, teachers, teacher, schools, kids, teaching, college, california
Topic 19: law, texas, court, judge, federal, rules, immigrants, oliver, department, ban
Topic 20: scientists, space, video, nasa, earth, world, mars, science, spacex, elon

3.4 Question 3.5: Discussion

Discuss the results you have observed by answering the following three questions:

- 1. For each setting, do you observe meaningful topics?
- 2. Which setting (K = 5, 10, or 20) gives the most meaningful topics?
- 3. Did you expect to see alignment (i.e., similar top-ranked words) between θ_k^b and θ_k^h for each k?

Solution:

- K = 5
  Yes, several body topics appear quite meaningful and distinct and headline topics largely mirror the body topics in theme, indicating good discovery:
  - Body topic 1 theme: General human experience / daily life
  - Body topic 2 theme: Education related
  - Body topic 3: US politics/news
  - Body topic 4: Art and climate/environment
  - Body topic 5: Sexual Abuse/Harassment
  - Headline topic 1: General/Celebrity/Lifestyle
  - Headline topic 2: Education and Social media
  - Headline topic 3: US politics/news and olympics
  - Headline topic 4: Art media / climate
  - Headline topic 5: Sexual abuse/Harassment

and so on ... for K=10 and K=20.

- I would lean towards K=20 as providing the most meaningful topics in terms of capturing distinct, understandable themes with reasonable specificity.
  K=5 gives very broad but generally coherent topics. It provides a decent overview but misses some nuances.
  K=10 seems as good as K=20. It breaks down the broader K=5 topics into more specific, clearly identifiable and distinct themes. But K=20 offers the highest level of detail. But a higher value of K is not always ideal, as there is a risk that some topics might become too fragmented or represent noise. However in this case, the results seem meaningful / useful.

- Yes, a significant degree of alignment was to be expected, and is largely observed in the results as well. While the style and specific vocabulary might differ between body and headline, the core semantic concepts related to the topic should be present in both. For example, for a topic about climate, we can expect to see words like "global", "climate", "change" occur in both the body and head.

4 Section 4: Gibbs Sampling

For the model described in the previous question, let us place prior distributions over the model parameters $\pi's$ and $\theta's$. Specifically, we assume that each categorical distribution π_i is sampled from a symmetric Dirichlet distribution parameterised by α (denoted as $\text{Dir}(\alpha)$), each categorical distribution θ_k^b is sampled from a symmetric Dirichlet distribution parameterised by β_1 (denoted as $\text{Dir}(\beta_1)$), and each categorical distribution θ_k^h is sampled from a symmetric Dirichlet distribution parameterised by β_2 (denoted as $\text{Dir}(\beta_2)$).
Let w^b denote all the words in the bodies of the articles and w^h denote all the words in the headlines of the articles. Similarly, let z^b denote all the hidden variables associated with the words in the bodies of the articles, and z^h the hidden variables associated with the words in the headlines of the articles.
A common technique to estimate the model parameters $\pi's$ and $\theta's$ with the model described above is to first use collapsed Gibbs sampling to sample the hidden topics assigned to each word, i.e., z^b and z^h. Given the sampled z^b and z^h, the parameters $\pi's$ and $\theta's$ can be relatively easily estimated.
Collapsed Gibbs sampling attempts to obtain a sample of z^b and z^h based on the following posterior distribution:

$$p(z^b, z^h \mid w^b, w^h, \alpha, \beta_1, \beta_2). \tag{89}$$

4.1 Question 4.1: The collapsed Gibbs sampling formulas

Derive the collapsed Gibbs sampling formulas, i.e.,

$$p(z_{ij}^b = k \mid z_{\neg ij}^b, z^h, w^b, w^h, \alpha, \beta_1, \beta_2) \tag{90}$$

and

$$p(z_{ij}^h = k \mid z^b, z_{\neg ij}^h, w^b, w^h, \alpha, \beta_1, \beta_2) \tag{91}$$

Here $z_{\neg ij}^b$ denotes all z^b except z_{ij}^b. $z_{\neg ij}^h$ is similarly defined.
Solution:[4][5][3]
Collapsed Gibbs Sampling formula for $z_{i,j}^b = k$ (assigning topic k to the j-th word $w_{i,j}^b$ in the body of article i)
Let $w_{target} = w_{i,j}^b$.

$$P(z_{i,j}^b = k \mid \mathbf{z}_{\neg(i,j)}^b, \mathbf{z}^h, \mathbf{w}^b, \mathbf{w}^h, \alpha, \beta_1, \beta_2) \propto A \times B \tag{92}$$

where

- A is the word topic probability (for body)
  A represents how well word w_{target} fits into body topic k, based on other words already assigned to that body topic across the entire corpus.

$$A = \frac{n_{k,w_{target}}^{b,\neg(i,j)} + \beta_1}{n_k^{b,\neg(i,j)} + V\beta_1} \tag{93}$$

  where:
  - $n_{k,w_{target}}^{b,\neg(i,j)}$: Number of times the specific word w_{target} is assigned to body topic k in all articles, excluding the current instance (i, j).
  - $n_k^{b,\neg(i,j)}$: Total number of words (any word) assigned to body topic k in all articles, excluding the current instance (i, j)

- B is the topic article probability
  This term represents how prevalent topic k is in article i, based on the other topic assignments within that article.

$$B = \frac{n_{i,k}^{\neg(i,j)} + \alpha}{n_i^{\neg(i,j)} + K\alpha} \tag{94}$$

  where
  - $n_{i,k}^{\neg(i,j)}$: Number of times topic k is assigned to any other word in article i. This count includes:
    * Body words in article i other than $w_{i,j}^b$ (i.e., from $\mathbf{z}_{\neg(i,j)}^b$ restricted to article i).
    * All headline words in article i (i.e., from $\mathbf{z}_i^h$).
  - $n_i^{\neg(i,j)}$: Total number of such other words in article i. If article i has M_i body words and L_i headline words, then $n_i^{\neg(i,j)} = (M_i - 1) + L_i$.

  So the full formula $z_{i,j}^b = k$ is:

$$P(z_{i,j}^b = k | \dots) \propto \frac{n_{k,w_{i,j}^b}^{b,\neg(i,j)} + \beta_1}{n_k^{b,\neg(i,j)} + V\beta_1} \times \frac{n_{i,k}^{\neg(i,j)} + \alpha}{(M_i - 1) + L_i + K\alpha} \tag{95}$$

  Collapsed Gibbs Sampling formula for $z_{i,l}^h = k$ (assigning topic k to the l-th word $w_{i,l}^b$ in the body of article i)
Let $w_{target} = w_{i,l}^b$.
The formula is symmetric:

$$P(z_{i,l}^h = k | \mathbf{z}^b, \mathbf{z}_{\neg(i,l)}^h, \mathbf{w}^b, \mathbf{w}^h, \alpha, \beta_1, \beta_2) \propto C \times D \tag{96}$$

  where

- C is the word topic probability (for headline)
  A represents how well word w_{target} fits into headline topic k

$$C = \frac{n_{k,w_{target}}^{h,\neg(i,l)} + \beta_2}{n_k^{h,\neg(i,l)} + V\beta_2} \tag{97}$$

  where:

  - $n_{k,w_{target}}^{h,\neg(i,l)}$: Number of times the specific word w_{target} is assigned to headline topic k in all articles, excluding the current instance (i, l)
  - $n_k^{h,\neg(i,l)}$: Total number of words (any word) assigned to headline topic k in all articles, excluding the current instance (i, l)

- D is the topic article probability
  This term is pretty much the same as B except the counts are defined slightly differently because we are now excluding a headline word.

$$D = \frac{n_{i,k}^{\neg(i,l)} + \alpha}{n_i^{\neg(i,l)} + K\alpha} \tag{98}$$

  where

  - $n_{i,k}^{\neg(i,l)}$: Number of times topic k is assigned to any other word in article i. This count includes:
    - All body words in article i (i.e., from $\mathbf{z}_i^b$).
    - Headline words in article i other than $w_{i,l}^h$ (i.e., from $\mathbf{z}_{\neg(i,l)}^h$ restricted to article i).
  - $n_i^{\neg(i,l)}$: Total number of such other words in article i. If article i has M_i body words and L_i headline words, then $n_i^{\neg(i,l)} = M_i + (L_i - 1)$.

So, the full formula for $z_{i,j}^b = k$ is:

$$P(z_{i,l}^h = k| \dots) \propto \frac{n_{k,w_{i,l}^h}^{h,\neg(i,l)} + \beta_2}{n_k^{h,\neg(i,l)} + V\beta_2} \times \frac{n_{i,k}^{\neg(i,l)} + \alpha}{M_i + (L_i - 1) + K\alpha} \tag{99}$$

4.2 Question 4.2: Estimation of $\pi's, \theta^b s$ and $\theta^h s$

Given a particular sample of z^b and z^h, derive the formulas for estimating the model parameters $\pi's$, $\theta^b s$ and $\theta^h s$.
Solution:
1. Estimating π_i - topic mixture for article i

- Prior: $\pi_i \sim Dir(\alpha, \alpha, \dots, \alpha)$ (K times, since it's a symmetric Dirichlet with scalar α)
- Observed data (given z^b, z^h): for article i, we have a set of topic assignments for its words. Let $n_{i,k}$ be the number of words in article i that are assigned to topic k based on our sample z_i^b and z_i^h.

$$n_{i,k} = \sum_{j=1}^{M_i} \delta(z_{i,j}^b = k) + \sum_{l=1}^{L_i} \delta(z_{i,l}^h = k) \quad | \ \delta \text{ is the indicator function} \tag{100}$$

- Posterior: due to Dirichlet-Multinomial conjugacy, the posterior distribution for π_i is also a Dirichlet distribution:

$$\pi_i \mid z_i^b, z_i^h, w_i^b, w_i^h, \alpha \sim Dir(n_{i,1} + \alpha, n_{i,2} + \alpha, \dots, n_{i,K} + \alpha) \tag{101}$$

- Estimation (using the mean of the posterior Dirichlet distribution):
  The mean of a Dirichlet distribution $Dir(\alpha_1, \dots, \alpha_K)$ for a parameter p_k is $\frac{\alpha_k}{\sum_j \alpha_j}$.
  So the estimate for $\pi_{i,k}$ (the probability of topic k in article i) must be:

$$\pi_{i,k} = \frac{n_{i,k} + \alpha}{\sum_{s=1}^{K}(n_{i,s} + \alpha)} \tag{102}$$

$$\pi_{i,k} = \frac{n_{i,k} + \alpha}{(\sum_{s=1}^{K} n_{i,s}) + K\alpha} \tag{103}$$

Since $\sum_{s=1}^{K} n_{i,s}$ is simply the total number of words in article i, which is $M_i + L_i$, the formula becomes:

$$\hat{\pi}_{i,k} = \frac{n_{i,k} + \alpha}{M_i + L_i + K\alpha} \tag{104}$$

2. Estimating θ_k^b - word distribution for body topic k

- Prior: $\theta_k^b \sim Dir(\beta_1, \beta_1, \ldots, \beta_1)$ (V times, for each word in the vocabulary).
- Observed data (given z^b): for a specific body topic k, we look at all body words across all articles that were assigned to this topic k by the Gibbs sample z^b. Let $n_{k,v}^b$ be the number of times word v appears in the body of any article and is assigned to topic k.

$$n_{k,v}^b = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \delta(w_{i,j}^b = v \text{ and } z_{i,j}^b = k) \tag{105}$$

- Posterior: The posterior distribution for θ_k^b is:

$$\theta_k^b \mid z^b, w^b, \beta_1 \sim Dir(n_{k,1}^b + \beta_1, n_{k,2}^b + \beta_1, \ldots, n_{k,V}^b + \beta_1) \tag{106}$$

- Estimation:
  The estimate for $\theta_{k,v}^b$ (the probability of word v in body topic k) is:

$$\hat{\theta}_{k,v}^b = \frac{n_{k,v}^b + \beta_1}{\sum_{v'=1}^{V} (n_{k,v'}^b + \beta_1)} \tag{107}$$

$$\hat{\theta}_{k,v}^b = \frac{n_{k,v}^b + \beta_1}{(\sum_{v'=1}^{V} (n_{k,v'}^b) + V\beta_1} \tag{108}$$

Let $n_k^b = \sum_{v'=1}^{V} (n_{k,v'}^b)$ be the total number of body words assigned to topic k across all articles. Then:

$$\hat{\theta}_{k,v}^b = \frac{n_{k,v}^b + \beta_1}{n_k^b + V\beta_1} \tag{109}$$

3. Estimating θ_k^h - word distribution for headline topic k

- Prior: $\theta_k^h \sim Dir(\beta_2, \beta_2, ..., \beta_2)$ (V times).
- Observed data (given z^h): For a specific headline topic k, let $n_{k,v}^h$ be the number of times word v appears in the headline of any article and is assigned to topic k

$$n_{k,v}^h = \sum_{i=1}^{N} \sum_{l=1}^{L_i} \delta(w_{i,l}^h = v \text{ and } z_{i,l}^h = k) \tag{110}$$

- Posterior: The posterior distribution for θ_k^h is:

$$\theta_k^h | z^h, w^h, \beta_2 \sim Dir(n_{k,1}^h + \beta_2, n_{k,2}^h + \beta_2, ..., n_{k,V}^h + \beta_2) \tag{111}$$

- Estimation
  The estimate for $\theta_{k,v}^h$ (the probability of word v in headline topic k) is

$$\hat{\theta}_{k,v}^h = \frac{n_{k,v}^h + \beta_2}{\sum_{v'=1}^{V} (n_{k,v'}^h + \beta_2)} \tag{112}$$

$$\hat{\theta}_{k,v}^h = \frac{n_{k,v}^h + \beta_2}{(\sum_{v'=1}^{V} n_{k,v'}^h) + V\beta_2} \tag{113}$$

Let $n_k^h = \sum_{v'=1}^{V} n_{k,v'}^h$ be the total number of headline words assigned to topic k across all articles. Then:

$$\hat{\theta}_{k,v}^h = \frac{n_{k,v}^h + \beta_2}{n_k^h + V\beta_2} \tag{114}$$

4.3 Question 4.3: Implementing the Gibb sampler

Create a file called a2 sampling.py. You can re-use the relevant code from a2 mixture.py to pre-process the data. Implement the collapsed Gibbs sampler as defined above

4.4 Question 4.4: Results

Set $K = 10, \alpha = 1, \beta_1 = 0.01,$ and $\beta_2 = 0.01$. Run your implemented Gibbs sampler to collect a sample of z^b and z^h. Use it to estimate the $\pi's, \theta^b s$ and $\theta^h s$. Show the top-10 words with the highest probabilities in each θ_k^b and each θ_k^h. Discuss whether these topics are meaningful.

Solution:

Results:

Top 10 words for each Body topic (from Gibbs sampler estimate):

Topic 1: school, students, schools, education, public, state, percent, teachers, student, children
Topic 2: show, game, film, play, season, music, book, series, time, team
Topic 3: police, told, city, gun, shooting, people, home, video, found, car
Topic 4: trump, president, donald, house, news, campaign, national, white, election, washington
Topic 5: people, white, american, church, religious, god, world, community, political, america
Topic 6: art, climate, space, world, artists, science, years, work, water, artist
Topic 7: women, people, time, work, make, feel, things, men, life, thing
Topic 8: time, make, people, good, find, home, food, day, small, work
Topic 9: company, facebook, companies, business, million, data, percent, information, billion, uber
Topic 10: sexual, told, women, statement, harassment, abuse, court, assault, investigation, huffpost

Top 10 words for each Headline topic (from Gibbs sample estimates):

Topic 1: school, education, students, schools, american, teachers, student, college, public, high
Topic 2: book, olympics, watch, jimmy, olympic, game, winter, james, reveals, star
Topic 3: police, man, shooting, video, york, dead, gun, city, death, texas
Topic 4: trump, donald, news, trumps, house, fox, bill, gop, michael, seth
Topic 5: colbert, stephen, white, muslim, women, pope, trevor, noah, church, cohen
Topic 6: art, climate, change, scientists, artists, world, space, years, science, video
Topic 7: show, time, day, artist, people, good, men, dont, world, love
Topic 8: photos, week, make, baby, animal, heres, video, work, life, women
Topic 9: facebook, uber, million, apple, company, ceo, google, data, amazon, iphone
Topic 10: sexual, women, abuse, harassment, accused, assault, report, sex, nassar, larry

Overall, the discovered topics with K=10 are very meaningful and interpretable.

- Coherence: Most topics exhibit strong internal coherence, with the top words clearly pointing to a distinct real-world theme.
- Alignment: There is a strong and logical alignment between the top words in the body topics (θ^b) and the corresponding headline topics (θ^h). This is expected, as headlines should reflect the article's content. The differences are also logical – headlines might use more prominent names, stronger verbs, or more concise phrasing related to the same theme.
- Distinctiveness: Most of the 10 topics are well-separated from each other, though some general "lifestyle" topics (like 7 and 8) might have some overlap.

References

[1] Backpropagation. *Wikipedia.*

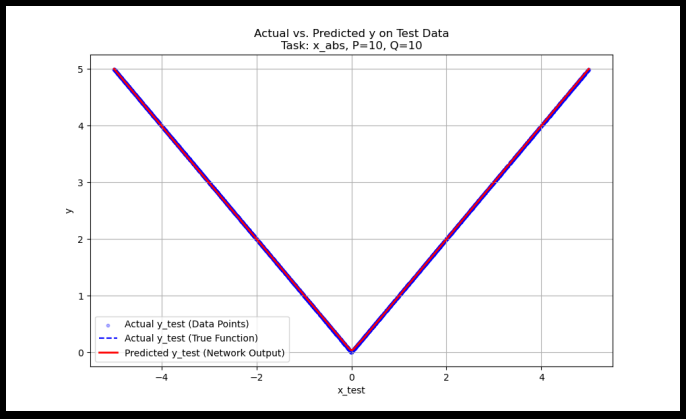
[2] Em algorithm and gmm model. *Wikipedia.*

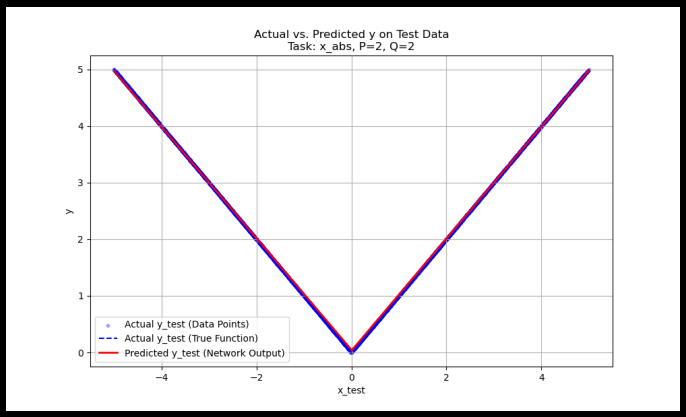
[3] Gibbs sampling. *Wikipedia.*

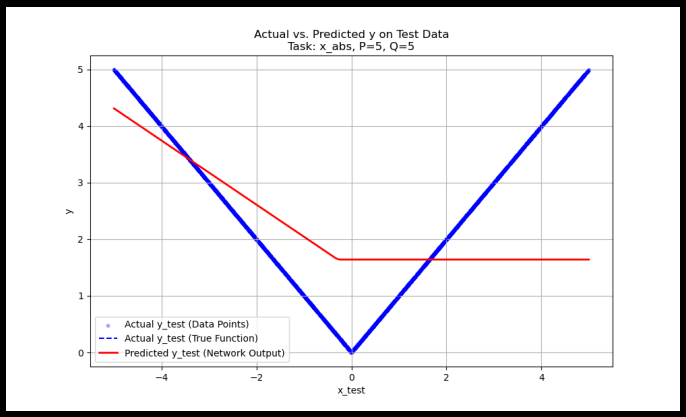
[4] Christopher M. Bishop. Pattern recognition and machine learning. *New York :Springer*, 2006.

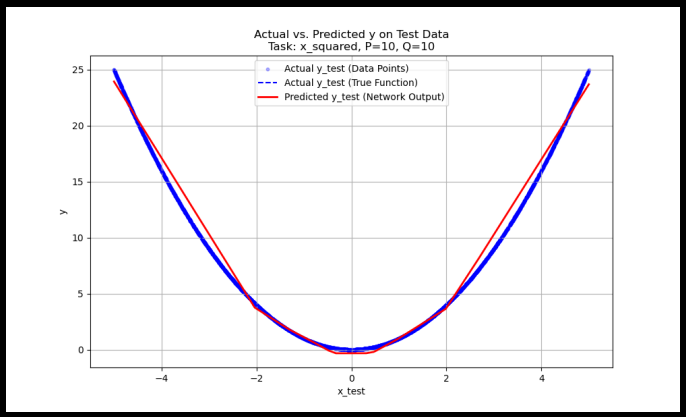
[5] William M. Darling. A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. *School of Computer Science. University of Guelph*, 2011.

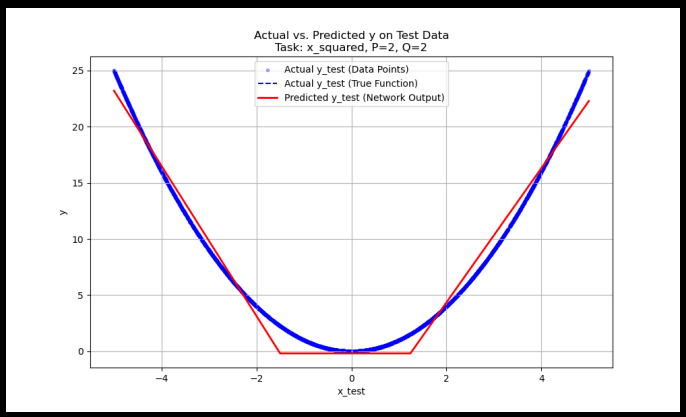
4.5 Appendix

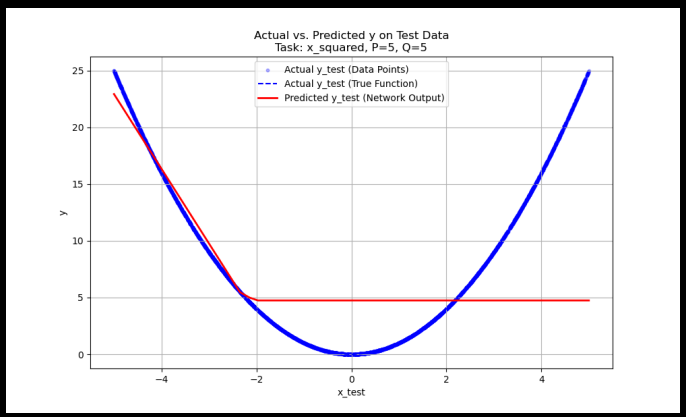


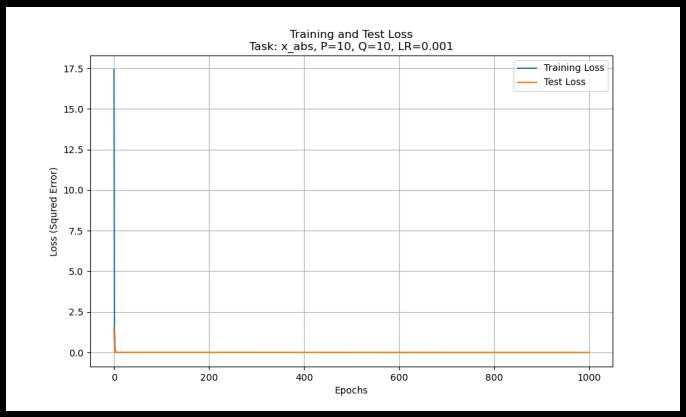


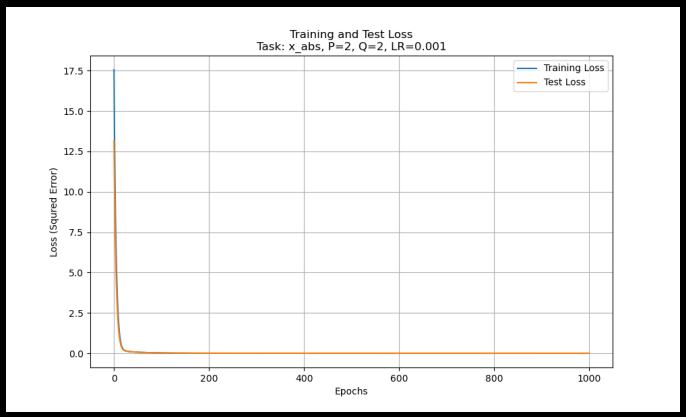

Actual vs. Predicted y on Test Data
Task: x_abs, P=5, Q=5
Actual y_test (Data Points)
Actual y_test (True Function)
Predicted y_test (Network Output)
x_test
y

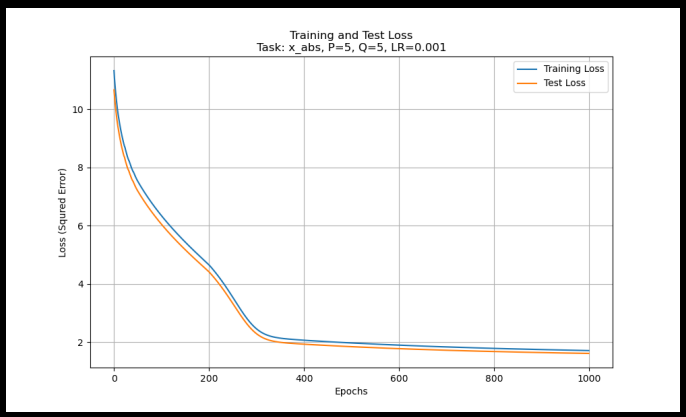

Actual vs. Predicted y on Test Data
Task: x_squared, P=10, Q=10
Actual y_test (Data Points)
Actual y_test (True Function)
Predicted y_test (Network Output)
x_test
y

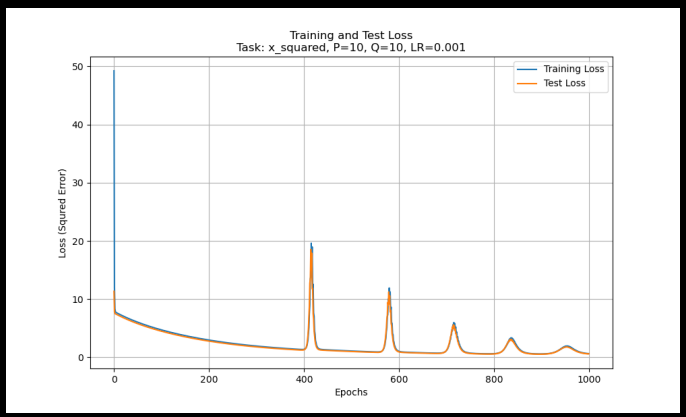

Actual vs. Predicted y on Test Data
Task: x_squared, P=2, Q=2
Actual y_test (Data Points)
Actual y_test (True Function)
Predicted y_test (Network Output)
x_test
y

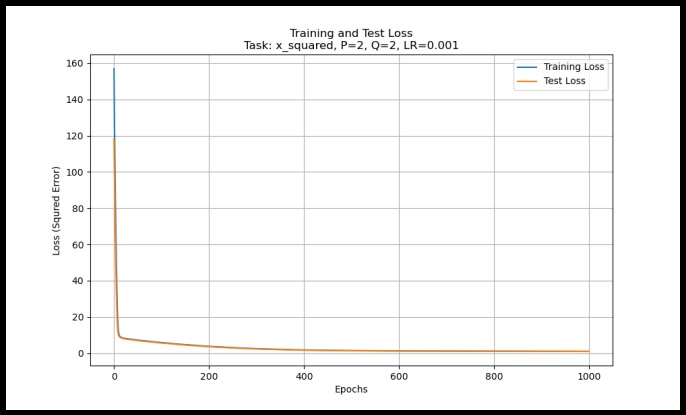

Actual vs. Predicted y on Test Data
Task: x_squared, P=5, Q=5
Actual y_test (Data Points)
Actual y_test (True Function)
Predicted y_test (Network Output)
x_test
y

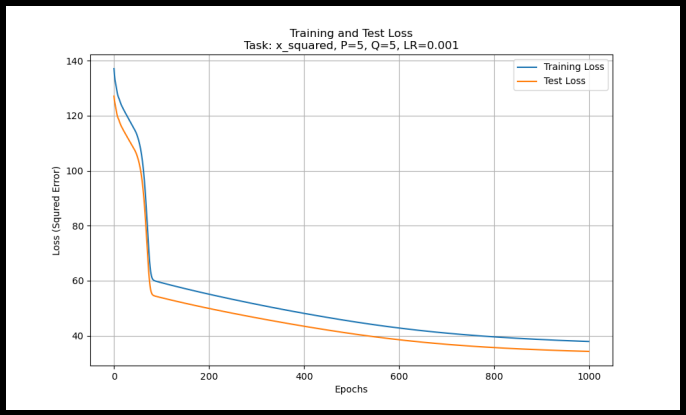

Training and Test Loss
Task: x_abs, P=10, Q=10, LR=0.001
Training Loss
Test Loss
Loss (Squred Error)
Epochs


Training and Test Loss
Task: x_abs, P=2, Q=2, LR=0.001
Training Loss
Test Loss
Loss (Squred Error)
Epochs


Training and Test Loss
Task: x_abs, P=5, Q=5, LR=0.001
Training Loss
Test Loss
Loss (Squred Error)
Epochs


Training and Test Loss
Task: x_squared, P=10, Q=10, LR=0.001
Training Loss
Test Loss
Loss (Squred Error)
Epochs


Training and Test Loss
Task: x_squared, P=2, Q=2, LR=0.001
Training Loss
Test Loss
Loss (Squred Error)
Epochs
0
20
40
60
80
100
120
140
160
0
200
400
600
800
1000


Training and Test Loss
Task: x_squared, P=5, Q=5, LR=0.001
Training Loss
Test Loss
Loss (Squred Error)
Epochs
40
60
80
100
120
140
0
200
400
600
800
1000