Introduction to Machine Learning

Project Phase 2

Ilia Hashemi Rad
MohammadPouya Toroghi

99102456 400109479

Theory Questions

Question 1

Solution: $p_{\mathbf{Z},\mathbf{Y}}(\mathbf{z},\mathbf{y})$:

$$p_{\mathbf{Z},\mathbf{Y}}(\mathbf{z},\mathbf{y}) = p_{\mathbf{Z}}(\mathbf{z})p_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z}) = \frac{1}{(2\pi)^{\frac{L}{2}}|\Sigma_{\mathbf{z}}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{z}-\mu_{\mathbf{z}})^T\Sigma_{\mathbf{z}}^{-1}(\mathbf{z}-\mu_{\mathbf{z}})} \times \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_{\mathbf{Y}|\mathbf{Z}}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}-(\mathbf{W}\mathbf{z}+\mathbf{b}))^T\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}(\mathbf{y}-(\mathbf{W}\mathbf{z}+\mathbf{b}))}$$

$$= \frac{1}{(2\pi)^{\frac{L+D}{2}}|\Sigma_{\mathbf{z}}\Sigma_{\mathbf{Y}|\mathbf{Z}}|^{\frac{1}{2}}} e^{\underbrace{-\frac{1}{2}((\mathbf{z}-\mu_{\mathbf{z}})^T\Sigma_{\mathbf{z}}^{-1}(\mathbf{z}-\mu_{\mathbf{z}}) + (\mathbf{y}-(\mathbf{W}\mathbf{z}+\mathbf{b}))^T\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}(\mathbf{y}-(\mathbf{W}\mathbf{z}+\mathbf{b})))}_{\triangleq A}}$$

$$\Rightarrow A = \mathbf{z}^T(\Sigma_{\mathbf{z}}^{-1} + \mathbf{W}^T\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}\mathbf{W})\mathbf{z} - \mathbf{z}^T(\Sigma_{\mathbf{z}}^{-1} + \mathbf{W}^T\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}\mathbf{W})\mu_{\mathbf{z}} - \mathbf{z}^T(\mathbf{W}^T\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1})\mathbf{y} + \mathbf{z}^T(\mathbf{W}^T\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1})(\mathbf{W}\mu_{\mathbf{z}}+\mathbf{b})$$

$$-\mu_{\mathbf{z}}^T(\Sigma_{\mathbf{z}}^{-1} + \mathbf{W}^T\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}\mathbf{W})\mathbf{z} + \mu_{\mathbf{z}}^T(\Sigma_{\mathbf{z}}^{-1} + \mathbf{W}^T\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}\mathbf{W})\mu_{\mathbf{z}} + \mu_{\mathbf{z}}^T(\mathbf{W}^T\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1})\mathbf{y} - \mu_{\mathbf{z}}^T(\mathbf{W}^T\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1})(\mathbf{W}\mu_{\mathbf{z}}+\mathbf{b})$$

$$-\mathbf{y}^T(\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}\mathbf{W})\mathbf{z} + \mathbf{y}^T(\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}\mathbf{W})\mu_{\mathbf{z}} + \mathbf{y}^T(\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1})\mathbf{y} - \mathbf{y}^T(\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1})(\mathbf{W}\mu_{\mathbf{z}}+\mathbf{b})$$

$$+(\mu_{\mathbf{z}}^T\mathbf{W}^T+\mathbf{b}^T)(\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}\mathbf{W})\mathbf{z} - (\mu_{\mathbf{z}}^T\mathbf{W}^T+\mathbf{b}^T)(\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}\mathbf{W})\mu_{\mathbf{z}} - (\mu_{\mathbf{z}}^T\mathbf{W}^T+\mathbf{b}^T)(\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1})\mathbf{y} + (\mu_{\mathbf{z}}^T\mathbf{W}^T+\mathbf{b}^T)(\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1})(\mathbf{W}\mu_{\mathbf{z}}+\mathbf{b})$$

Now, we notice that A can be expressed in the form below:

$$A = (\begin{bmatrix}\mathbf{z}\\ \mathbf{y}\end{bmatrix} - \underbrace{\begin{bmatrix}\mu_{\mathbf{z}}\\ \mathbf{W}\mu_{\mathbf{z}}+\mathbf{b}\end{bmatrix}}_{\mu_{\mathbf{Z},\mathbf{Y}}})^T \underbrace{\begin{bmatrix}\Sigma_{\mathbf{z}}^{-1} + \mathbf{W}^T\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}\mathbf{W} & -\mathbf{W}^T\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}\\ -\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}\mathbf{W} & \Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}\end{bmatrix}}_{\Sigma_{\mathbf{Z},\mathbf{Y}}^{-1}} (\begin{bmatrix}\mathbf{z}\\ \mathbf{y}\end{bmatrix} - \begin{bmatrix}\mu_{\mathbf{z}}\\ \mathbf{W}\mu_{\mathbf{z}}+\mathbf{b}\end{bmatrix})$$

Now from linear algebra, we knew that for some block matrix R in the form below, for which D^{-1} and $(A-BD^{-1}C)^{-1}$ exist, R^{-1} can be derived as:

$$R = \begin{bmatrix}A & B\\ C & D\end{bmatrix} \Rightarrow R^{-1} = \begin{bmatrix}(A-BD^{-1}C)^{-1} & -(A-BD^{-1}C)^{-1}BD^{-1}\\ -D^{-1}C(A-BD^{-1}C)^{-1} & D^{-1}+D^{-1}C(A-BD^{-1}C)^{-1}BD^{-1}\end{bmatrix}$$

Now, assuming $R = \Sigma_{\mathbf{Z},\mathbf{Y}}^{-1}$, we get: $D = \Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}$, and $A - BD^{-1}C = \Sigma_{\mathbf{z}}^{-1}$, which are in fact inversible.

$$\Rightarrow R^{-1} = (\Sigma_{\mathbf{Z},\mathbf{Y}}^{-1})^{-1} = \Sigma_{\mathbf{Z},\mathbf{Y}} = \begin{bmatrix}\Sigma_{\mathbf{z}} & \Sigma_{\mathbf{z}}\mathbf{W}^T\\ \mathbf{W}\Sigma_{\mathbf{z}} & \Sigma_{\mathbf{Y}|\mathbf{Z}} + \mathbf{W}\Sigma_{\mathbf{z}}\mathbf{W}^T\end{bmatrix}$$

Then again, from linear algebra, we knew that for some other block matrix R' in the form below, for which A'^{-1} exists, $|R'|$ can be calculated by:

$$R' = \begin{bmatrix}A' & B'\\ C' & D'\end{bmatrix} \Rightarrow |R'| = |A'||D' - C'A'^{-1}B'|$$

Now, assuming $R' = \Sigma_{\mathbf{Z},\mathbf{Y}}$, we get: $A' = \Sigma_{\mathbf{Z}}$, which is in fact inversible.

$$|R'| = |\Sigma_{\mathbf{Z},\mathbf{Y}}| = |\Sigma_{\mathbf{Z}}||\Sigma_{\mathbf{Y}|\mathbf{Z}}| = |\Sigma_{\mathbf{Z}}\Sigma_{\mathbf{Y}|\mathbf{Z}}|$$

$$\Rightarrow p_{\mathbf{Z},\mathbf{Y}}(\mathbf{z},\mathbf{y}) = \frac{1}{(2\pi)^{\frac{L+D}{2}}|\Sigma_{\mathbf{z}}\Sigma_{\mathbf{Y}|\mathbf{Z}}|^{\frac{1}{2}}} e^{-\frac{1}{2}A} = \frac{1}{(2\pi)^{\frac{L+D}{2}}|\Sigma_{\mathbf{Z},\mathbf{Y}}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\begin{bmatrix}\mathbf{z}\\ \mathbf{y}\end{bmatrix} - \mu_{\mathbf{Z},\mathbf{Y}})^T\Sigma_{\mathbf{Z},\mathbf{Y}}^{-1}(\begin{bmatrix}\mathbf{z}\\ \mathbf{y}\end{bmatrix} - \mu_{\mathbf{Z},\mathbf{Y}})}$$

$$\Rightarrow p_{\mathbf{Z},\mathbf{Y}}(\mathbf{z},\mathbf{y}) = \mathcal{N}(\mu_{\mathbf{Z},\mathbf{Y}}, \Sigma_{\mathbf{Z},\mathbf{Y}})$$

Solution: $p_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y})$:

From **Theoretical Homework 1**, we knew that, if $(Z^T, Y^T)^T = X \sim \mathcal{N}(\mu_{\mathbf{X}}, \Sigma_{\mathbf{X}})$, for which:

$$\mu_{\mathbf{X}} = \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix},\ \Sigma_{\mathbf{X}} = \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix},\ \Delta_{\mathbf{X}} = \Sigma_{\mathbf{X}}^{-1} = \begin{bmatrix}\Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22}\end{bmatrix}$$

Then, $p_{\mathbf{Y}}(\mathbf{y})$ can be calculated as:

$$p_{\mathbf{Y}}(\mathbf{y}) = \mathcal{N}(\mu_2, \Sigma_{22})$$

Now, for our problem, $\mu_X = \mu_{Z,Y}$, $\Sigma_X = \Sigma_{Z,Y}$ and $\Delta_X = \Sigma_{Z,Y}^{-1}$, which were calculated earlier.

$$\Rightarrow p_{\mathbf{Y}}(\mathbf{y}) = \mathcal{N}(\mathbf{W}\mu_{\mathbf{z}} + \mathbf{b}, \Sigma_{\mathbf{Y}|\mathbf{Z}} + \mathbf{W}\Sigma_{\mathbf{z}}\mathbf{W}^T)$$

Now, having $p_{\mathbf{Z}}(\mathbf{z})$, $p_{\mathbf{Y}}(\mathbf{y})$, and $p_{\mathbf{Z},\mathbf{Y}}(\mathbf{z},\mathbf{y})$, we will use the formula derived in **Theoretical Homework 1** to calculate $p_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y})$:

$$\Rightarrow p_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y}) = \mathcal{N}(\mu_{\mathbf{Z}|\mathbf{Y}}, \Sigma_{\mathbf{Z}|\mathbf{Y}})$$

Where: $\Sigma_{\mathbf{Z}|\mathbf{Y}} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Sigma_{\mathbf{z}} - \Sigma_{\mathbf{z}}\mathbf{W}^T(\Sigma_{\mathbf{Y}|\mathbf{Z}} + \mathbf{W}\Sigma_{\mathbf{z}}\mathbf{W}^T)^{-1}\mathbf{W}\Sigma_{\mathbf{z}},\ \Sigma_{\mathbf{Z}|\mathbf{Y}}^{-1} = \Delta_{11} = \Sigma_{\mathbf{z}}^{-1} + \mathbf{W}^T\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}\mathbf{W}$

And: $\mu_{\mathbf{Z}|\mathbf{Y}} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y} - \mu_2) = \mu_{\mathbf{z}} - \Sigma_{\mathbf{z}}\mathbf{W}^T(\Sigma_{\mathbf{Y}|\mathbf{Z}} + \mathbf{W}\Sigma_{\mathbf{z}}\mathbf{W}^T)^{-1}(\mathbf{y} - (\mathbf{W}\mu_{\mathbf{z}} + \mathbf{b}))$

$$\Rightarrow \mu_{\mathbf{Z}|\mathbf{Y}} = \Sigma_{\mathbf{Z}|\mathbf{Y}}(\mathbf{W}^T\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}(\mathbf{y} - \mathbf{b}) + \Sigma_{\mathbf{z}}^{-1}\mu_{\mathbf{z}})$$

Question 2

Solution:

Assume we define some variable $\mathbf{t} = \begin{bmatrix}t_1 \\ t_2 \\ \vdots \\ t_K\end{bmatrix}$, where $p_{\mathbf{t}}(t_k = 1) = \pi_k$. We may write:

$$p_{\mathbf{Z}|\mathbf{t}}(\mathbf{z}|t_k = 1) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

According to the results of the previous question:

$$p_{\mathbf{Z}|\mathbf{Y},\mathbf{t}}(\mathbf{z}|\mathbf{y}, t_k = 1) = \mathcal{N}(\mathbf{z}|\mathbf{y}; \boldsymbol{\mu}_{k;\mathbf{Z}|\mathbf{Y}}, \boldsymbol{\Sigma}_{k;\mathbf{Z}|\mathbf{Y}})$$

Where:

$$\boldsymbol{\Sigma}_{k;\mathbf{Z}|\mathbf{Y}} = \boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k\mathbf{W}^T(\Sigma_{\mathbf{Y}|\mathbf{Z}} + \mathbf{W}\boldsymbol{\Sigma}_k\mathbf{W}^T)^{-1}\mathbf{W}\boldsymbol{\Sigma}_k, \quad \boldsymbol{\mu}_{k;\mathbf{Z},\mathbf{Y}} = \boldsymbol{\Sigma}_{k;\mathbf{Z}|\mathbf{Y}}(\mathbf{W}^T\Sigma_{\mathbf{Y}|\mathbf{Z}}^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k)$$

$$\Rightarrow p_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y}) = \sum_{k=1}^{K} p_{\mathbf{Z}|\mathbf{Y},\mathbf{t}}(\mathbf{z}|\mathbf{y}, t_k = 1)p_{\mathbf{t}}(t_k = 1) = \sum_{k=1}^{K} \pi_k\mathcal{N}(\mathbf{z}|\mathbf{y}; \boldsymbol{\mu}_{k;\mathbf{Z}|\mathbf{Y}}, \boldsymbol{\Sigma}_{k;\mathbf{Z}|\mathbf{Y}})$$

We can see that the posterior $p_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y})$ is also **GMM**, with the same π_k's, but different μ and Σ's, which all their values were derived above.

Question 3

Solution:

The Gaussian Mixture Model(**GMM**) is often preferred for modeling the prior distribution, especially when the likelihood distribution is normal, due to its flexibility in capturing complex patterns in the data. Here are a few reasons for that matter which we could come up with:

1. As was mentioned in the previous question, the posterior will still be **GMM**, which makes our analysis much easier.

2. As was mentioned in the previous phase, the provided flexibility of **GMM** is much higher than any one distribution, as **GMM** allows for the representation of multimodal distributions, which can be useful when the underlying data may come from multiple distinct sources or clusters.
3. **GMM**s naturally incorporate uncertainty by assigning probabilities to each component of the mixture. This is particularly useful when dealing with data that exhibits inherent uncertainty or when the data contains outliers. By assigning lower weights to components with low probabilities, **GMM**s can effectively handle these situations.
4. In some cases, using a **GMM** as a prior can improve the performance of statistical models. The added flexibility of **GMM**s allows them to better approximate the underlying distribution of the data, resulting in better model fit and more accurate predictions.
5. **GMM**s are often used in unsupervised learning tasks, such as clustering, where the goal is to discover underlying patterns or groups in the data. In these scenarios, the use of **GMM**s as prior distributions can help guide the clustering process and improve the accuracy of the resulting clusters.
It's important to note that the choice of prior distribution depends on the specific problem and the characteristics of the data. While a **GMM** is a popular choice, there may be situations where a different prior distribution, such as a single normal distribution, is more appropriate or sufficient to capture the data's characteristics.

Simulation Questions

Question 4

As was said in the **.ipynb** file, we have already calculated the best σ^2, for each m, using **cross validation**. This was done by splitting the train dataset into train, and test datasets, and used cross validation to elicit the best value for σ, for each m, which could have been done because the noise is independent of the other parameters. So, we will not be trying different σs in this question. Also, keep in mind that to improve the runtime of our code, we ran the code for parts of the train dataset, to derive the parameters, but after calculating the optimal parameters we will be run the code for the whole test dataset.

Results: $K = 3$, $m = 8$

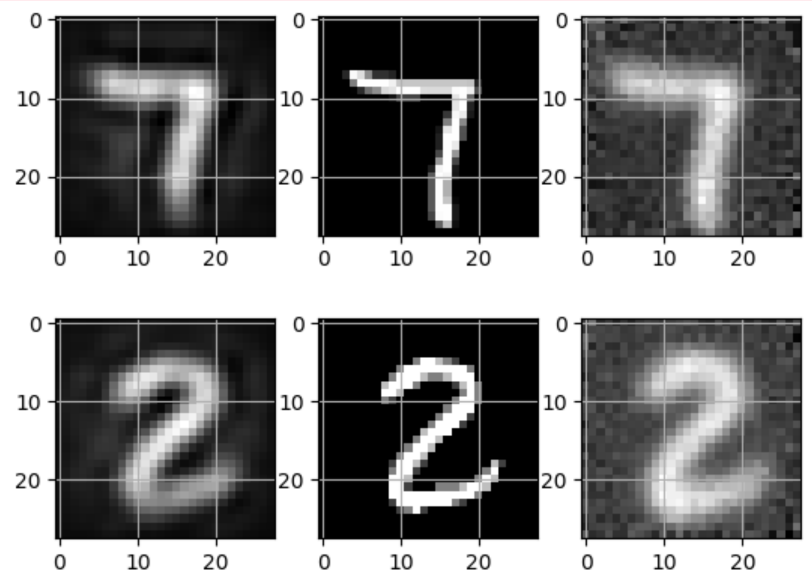

Figure 1: Denoised Images, Original Images, Corrupted Images

Calculated MSE = 33.6363901108913
Keep in mind that in calculation of MSE, we used mean of pixels that were in multiple patches, but for plotting we came to realize that by only summing patches of the picture, the resulting picture will have better quality!

Results: $K = 7$, $m = 8$

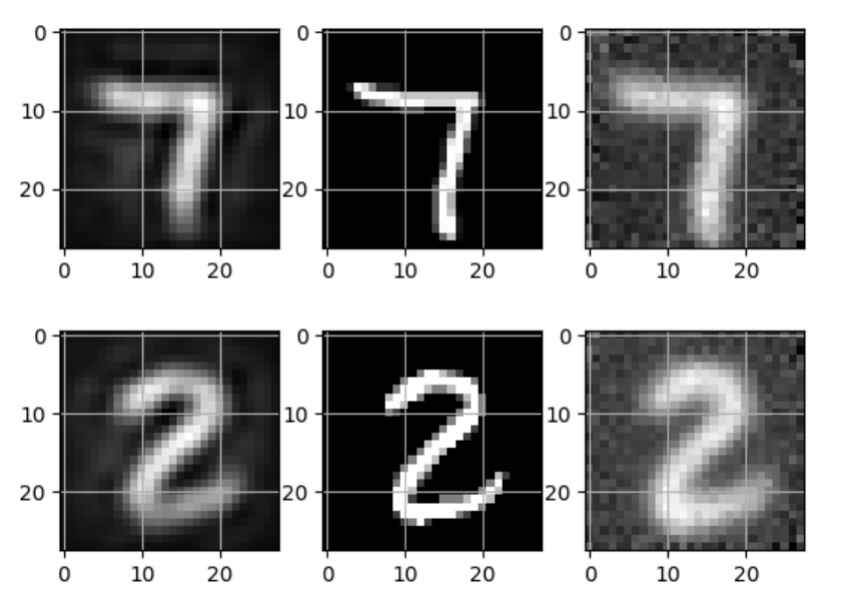

Figure 2: Denoised Images, Original Images, Corrupted Images

Calculated MSE = 32.96005407902957
Keep in mind that in calculation of MSE, we used mean of pixels that were in multiple patches, but for plotting we came to realize that by only summing patches of the picture, the resulting picture will have better quality!

Results: $K = 10$, $m = 8$

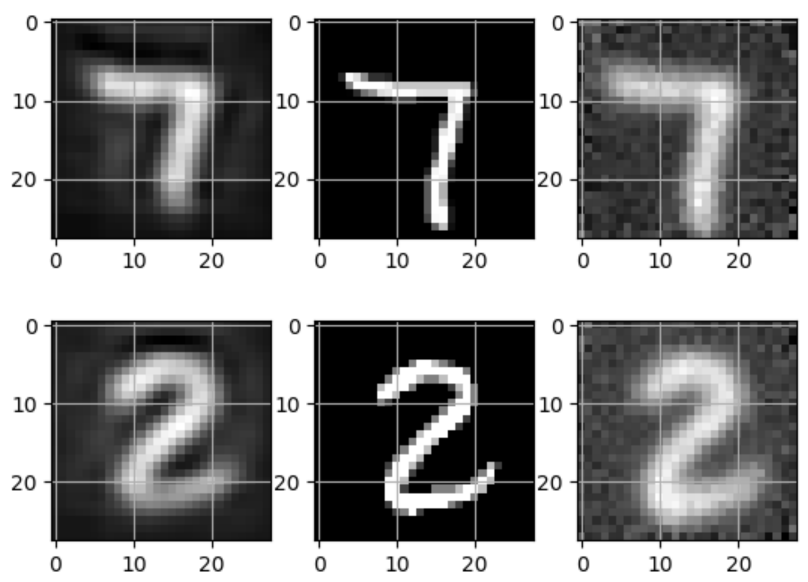

Figure 3: Denoised Images, Original Images, Corrupted Images

Calculated MSE = 32.34941658423573
Keep in mind that in calculation of MSE, we used mean of pixels that were in multiple patches, but for plotting we came to realize that by only summing patches of the picture, the resulting picture will have better quality!

Results: $K = 11$, $m = 8$

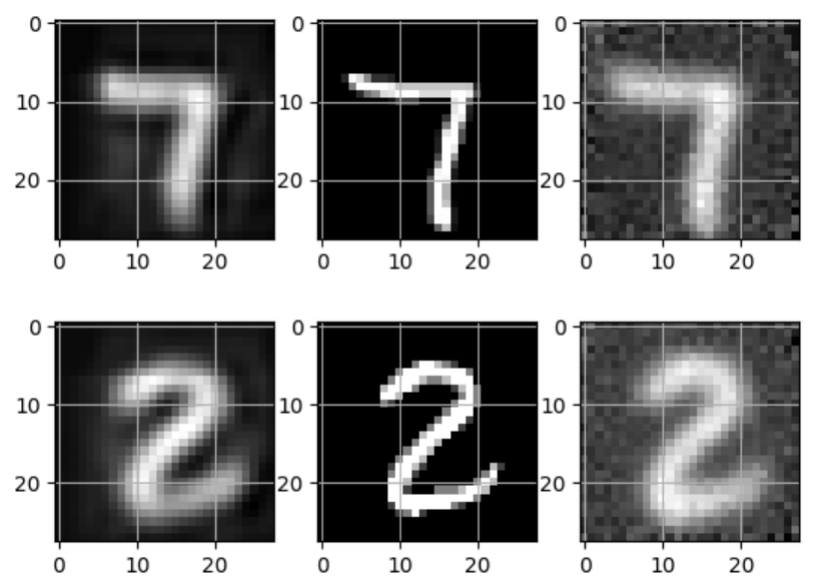

Figure 4: Denoised Images, Original Images, Corrupted Images

Calculated MSE = 16.896197222511454
Keep in mind that in calculation of MSE, we used mean of pixels that were in multiple patches, but for plotting we came to realize that by only summing patches of the picture, the resulting picture will have better quality!

Results: $K = 64$, $m = 8$

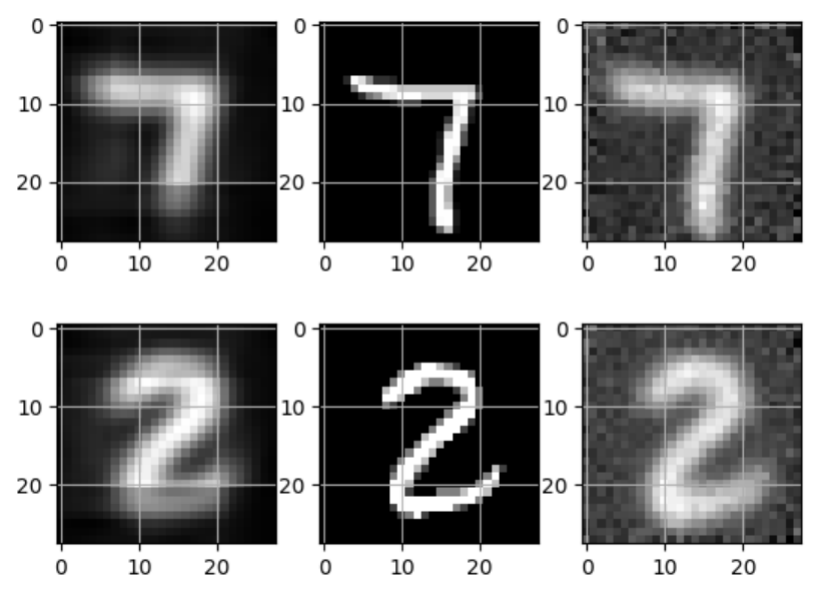

Figure 5: Denoised Images, Original Images, Corrupted Images

Calculated MSE = 12.009813619325282
Keep in mind that in calculation of MSE, we used mean of pixels that were in multiple patches, but for plotting we came to realize that by only summing patches of the picture, the resulting picture will have better quality!

Comparison:

We can see that between the relatively small values if K, $K = 11$ has the least MSE by a large gap(!), which was expected considering the fact that we ultimately want to cluster the **MNIST** data into 10 clusters, and considering that a lot of our patches are empty in the original images, so we choose $K = 11$, in our computations here on out. Now, the reason that we did not choose $K = 64$ as our optimal K value, is that despite having the best calculated MSE, its resulting pictures are somewhat blurry, and have less quality than the chosen value of $K = 11$' results, so it seems that it has led to overfitting.

Results: $K = 11$, $m = 4$

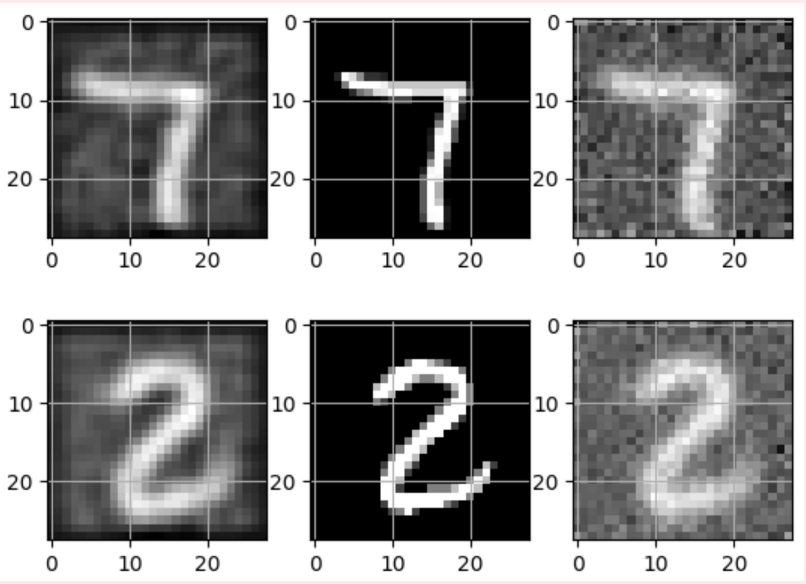

Figure 6: Denoised Images, Original Images, Corrupted Images

Calculated MSE = 65.905693591212

Keep in mind that in calculation of MSE, we used mean of pixels that were in multiple patches, but for plotting we came to realize that by only summing patches of the picture, the resulting picture will have better quality!

Results: $K = 11$, $m = 12$

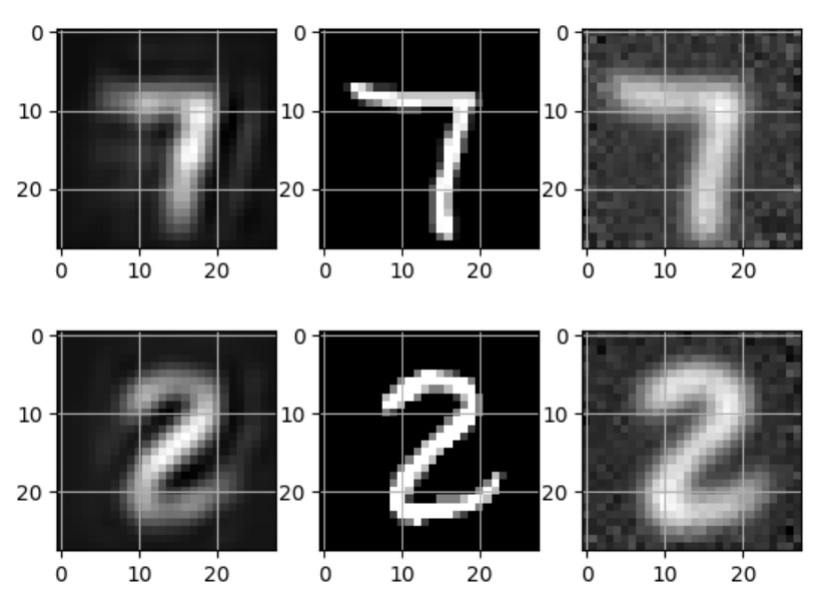

Figure 7: Denoised Images, Original Images, Corrupted Images

Calculated MSE = 13.633746355685131
Keep in mind that in calculation of MSE, we used mean of pixels that were in multiple patches, but for plotting we came to realize that by only summing patches of the picture, the resulting picture will have better quality!

Results: $K = 11$, $m = 16$

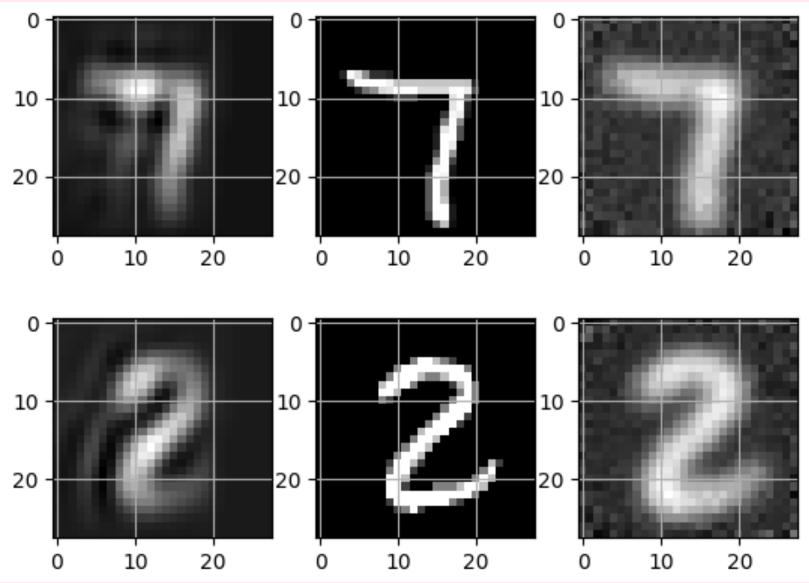

Figure 8: Denoised Images, Original Images, Corrupted Images

Calculated MSE = 9.998195738754685
Keep in mind that in calculation of MSE, we used mean of pixels that were in multiple patches, but for plotting we came to realize that by only summing patches of the picture, the resulting picture will have better quality!

Results: $K = 11$, $m = 20$

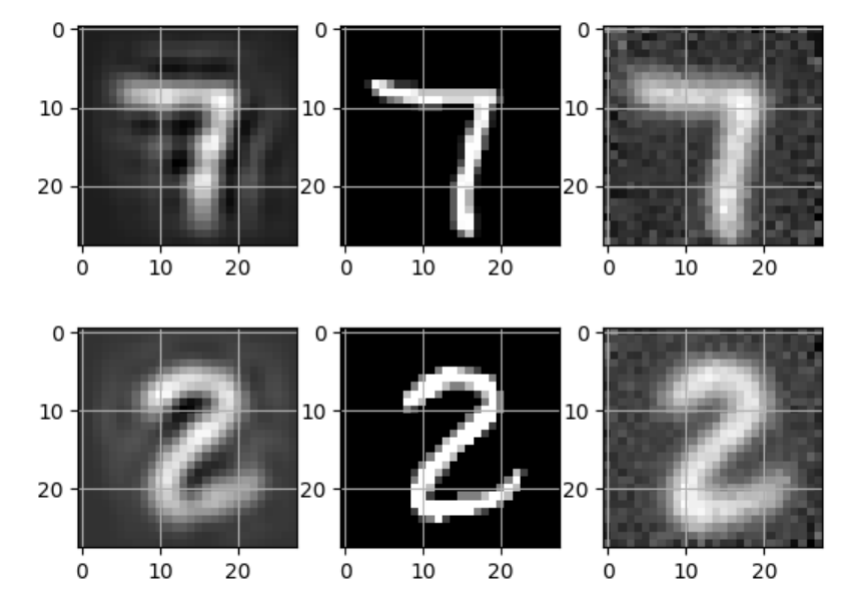

Figure 9: Denoised Images, Original Images, Corrupted Images

Calculated MSE = 8.52209528581841
Keep in mind that in calculation of MSE, we used mean of pixels that were in multiple patches, but for plotting we came to realize that by only summing patches of the picture, the resulting picture will have better quality!

Results: $K = 11$, $m = 28$

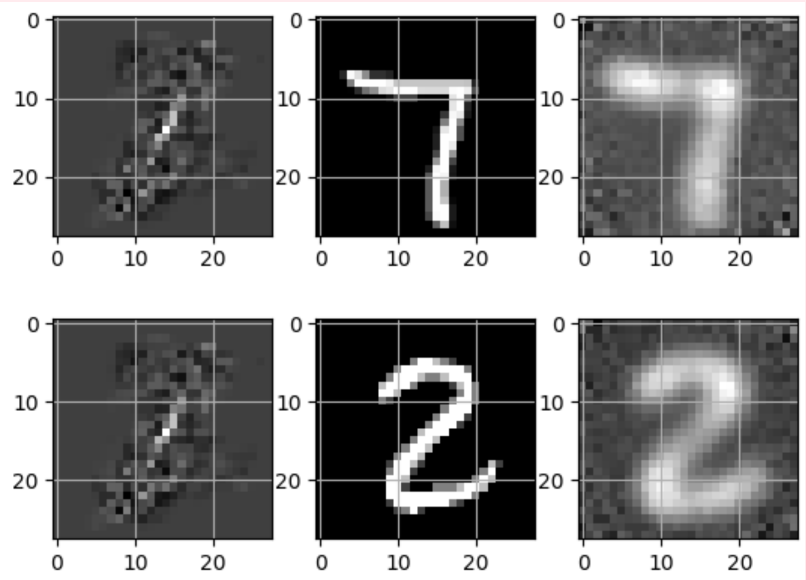

Figure 10: Denoised Images, Original Images, Corrupted Images

Calculated MSE = 366.8611873958767
Keep in mind that in calculation of MSE, we used mean of pixels that were in multiple patches, but for plotting we came to realize that by only summing patches of the picture, the resulting picture will have better quality!

Comparison:

We can see that even though $m = 20$ appears to have the least MSE, the quality of $m = 8$, is the best. Although one notable aspect of these recent plots is that, as expected, $m = 28$ has the worst quality, because it is essentially not using the patching idea and is actually implementing a simple MLP network instead of convolution.

As a result of all mentioned in this question, we have arrived at our ideal hyperparameters, which are, $K = 11$, $m = 8$ and $\sigma^2 = 100$. These are the values for which we will be running the code for the test dataset.

Final Results:

Using the previously derived parameters, we run the code for the whole test dataset, and these are a few of the results:

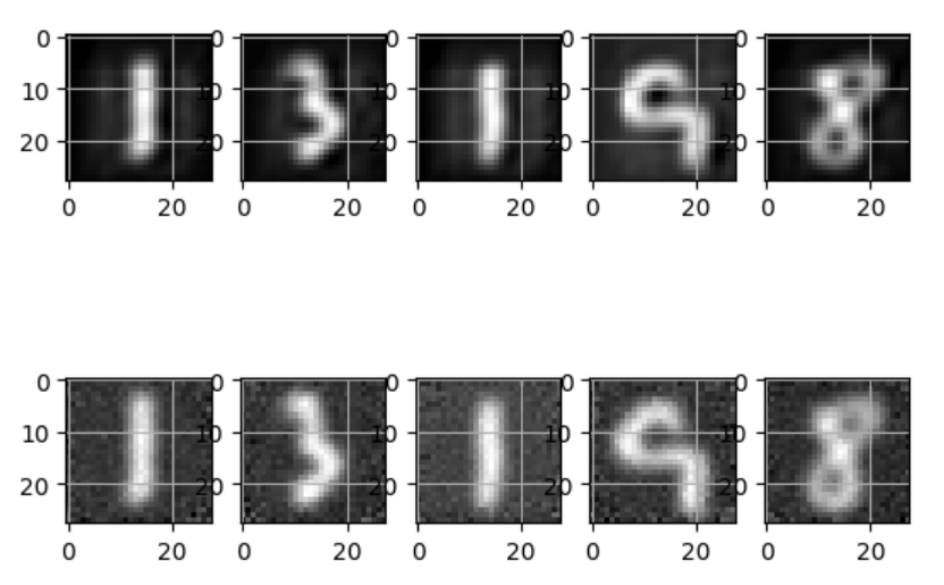

Figure 11: The upper ones are Denoised Images, and the lower ones are Corrupted Images

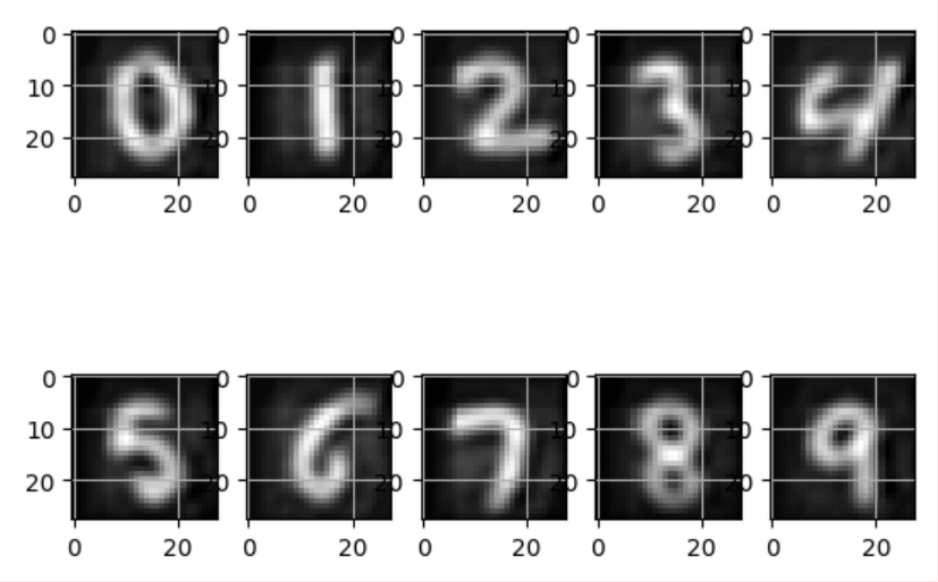

Figure 12: All are Denoised Images

We can see the acceptable quality of the output images.
Although these are the outputs of the code we had a bit of a problem downloading them into our devices and the quality of the images decreased as we did, regardless of how we did the transition. The best possible quality of the saved images we could reach can be seen in the **Results** folder.