COMS 4771 Machine Learning (2022 Fall)
Problem Set #3

Xinhao Li - xl2778@columbia.edu

2022/11/16

Problem 1

(i)

The marginal distribution of a subset of a collection of random variables is the probably distribution of the variables contained in the subset. It gives the probabilities of various values of the variables in the subset without reference to the values of the other variables(wiki). Also the pdf of multivariate Gaussian distribution is given by (wiki). Therefore the marginal distribution of x_1 is:

$$\begin{aligned} f(x_1) &= \int f(x_1, x_2) dx_2 \\ &= \int \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^{d_1}|\boldsymbol{\Sigma}|}} dx_2 \\ &= \frac{\exp\left(-\frac{1}{2}(x_1-\mu_1)^{\mathrm{T}}{\Sigma_{11}}^{-1}(x_1-\mu_1)\right)}{\sqrt{(2\pi)^{d_1}|\Sigma_{11}|}} \end{aligned} \tag{1}$$

(ii)

The joint distribution on x is:

$$\begin{aligned} f(\boldsymbol{x}) &= f(x_1, x_2) \\ &= \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^{d}|\boldsymbol{\Sigma}|}} \end{aligned} \tag{2}$$

where $d = d_1 + d_2$. Let's calculate $(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ this part first.

$$\begin{aligned}
(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) &= [(x_1 - \mu_1)^T, (x_2 - \mu_2)^T] \begin{bmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \\
&= \left[(x_1 - \mu_1)^T, (x_2 - \mu_2)^T\right] \begin{bmatrix} \Sigma^{11}(x_1 - \mu_1) + \Sigma^{12}(x_2 - \mu_2) \\ \Sigma^{21}(x_1 - \mu_1) + \Sigma^{22}(x_2 - \mu_2) \end{bmatrix} \\
&= (x_1 - \mu_1)^T \Big[\Sigma^{11}(x_1 - \mu_1) + \Sigma^{12}(x_2 - \mu_2)\Big] \\
&\quad + (x_2 - \mu_2)^T \Big[\Sigma^{21}(x_1 - \mu_1) + \Sigma^{22}(x_2 - \mu_2)\Big] \\
&= (x_1 - \mu_1)^T \Sigma^{11}(x_1 - \mu_1) + (x_1 - \mu_1)^T \Sigma^{12}(x_2 - \mu_2) \\
&\quad + (x_2 - \mu_2)^T \Sigma^{21}(x_1 - \mu_1) + (x_2 - \mu_2)^T \Sigma^{22}(x_2 - \mu_2) \qquad (3)
\end{aligned}$$

where

$$(x_1 - \mu_1)^T \Sigma^{12}(x_2 - \mu_2) = (x_2 - \mu_2)^T \Sigma^{21}(x_1 - \mu_1) \qquad (4)$$

Therefore, we have:

$$\begin{aligned}
(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) &= (x_1 - \mu_1)^T \Sigma^{11}(x_1 - \mu_1) + 2(x_1 - \mu_1)^T \Sigma^{12}(x_2 - \mu_2) \\
&\quad + (x_2 - \mu_2)^T \Sigma^{22}(x_2 - \mu_) \qquad (5)
\end{aligned}$$

Based on the facts of Σ_{11}, Σ_{12}, and Σ_{22} in the question, we can have:

$$\begin{aligned}
(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) &= (x_1 - \mu_1)^T \Big[\Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12})^{-1} \Sigma_{12}^T \Sigma_{11}^{-1}\Big](x_1 - \mu_1) \\
&\quad - 2(x_1 - \mu_1)^T \Big[\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12})^{-1}\Big](x_2 - \mu_2) \\
&\quad + (x_2 - \mu_2)^T \Big[(\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12})^{-1}\Big](x_2 - \mu_2) \\
&= (x_1 - \mu_1)^T \Sigma_{11}^{-1}(x_1 - \mu_1) \\
&\quad + (x_1 - \mu_1)^T \Big[\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12})^{-1} \Sigma_{12}^T \Sigma_{11}^{-1}\Big](x_1 - \mu_1) \\
&\quad - 2(x_1 - \mu_1)^T \Big[\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12})^{-1}\Big](x_2 - \mu_2) \\
&\quad + (x_2 - \mu_2)^T \Big[(\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12})^{-1}\Big](x_2 - \mu_2) \qquad (6)
\end{aligned}$$

Let $m = \Sigma_{12}^T \Sigma_{11}^{-1}(x_1 - \mu_1)$, then $m^T = (x_1 - \mu_1)^T \Sigma_{11}^{-1} \Sigma_{12}$. $A = \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}$. $n = (x_2 - \mu_2)$, then $n^T = (x_2 - \mu_2)^T$. Therefore, we have:

$$\begin{aligned}
(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) &= (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1) \\
&\quad + m^T A^{-1} m - 2 m^T A^{-1} n + n^T A^{-1} n \\
&= (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1) \\
&\quad + m^T A^{-1} m - m^T A^{-1} n - m^T A^{-1} n + n^T A^{-1} n \\
&= (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1) \\
&\quad + m^T A^{-1} (m - n) - (m - n)^T A^{-1} n \\
&= (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1) \\
&\quad + m^T A^{-1} (m - n) - n^T A^{-1} (m - n) \\
&= (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1) \\
&\quad + (m - n)^T A^{-1} (m - n) \\
&= (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1) \\
&\quad + (n - m)^T A^{-1} (n - m) \qquad (7)
\end{aligned}$$

Substitute m, A, and n back into equation 7, and we have:

$$\begin{aligned}
(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) &= (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1) + (m - n)^T A^{-1} (m - n) \\
&= (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1) \\
&\quad + ((x_2 - \mu_2) - \Sigma_{12}^T \Sigma_{11}^{-1} (x_1 - \mu_1))^T A^{-1} ((x_2 - \mu_2) - \Sigma_{12}^T \Sigma_{11}^{-1} (x_1 - \mu_1)) \\
&= (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1) \\
&\quad + (x_2 - (\mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1} (x_1 - \mu_1)))^T A^{-1} (x_2 - (\mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1} (x_1 - \mu_1))) \\
&= (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1) + (x_2 - b)^T A^{-1} (x_2 - b) \qquad (8)
\end{aligned}$$

where $b = \mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1} (x_1 - \mu_1)$ and $A = \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}$.
Finally, we have:

$$\begin{aligned}
f(\boldsymbol{x}) &= \frac{\exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \\
&= \frac{\exp\left[-\frac{1}{2} \left((x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1) + (x_2 - b)^T A^{-1} (x_2 - b)\right)\right]}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \qquad (9)
\end{aligned}$$

where $b = \mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1} (x_1 - \mu_1)$, $A = \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}$, and $d = d_1 + d_2$.

(iii)

Let $g(\boldsymbol{x}) = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$. For x_1 and x_1 and x_2, we have:

$$g(x_1) = (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1) \qquad (10)$$

$$g(x_2) = (x_2 - b)^T A^{-1}(x_2 - b) \tag{11}$$

$$g(x_1, x_2) = g(x_1) + g(x_2) \tag{12}$$

So,

$$\begin{aligned}
f(\boldsymbol{x}) &= \frac{\exp\left[-\frac{1}{2}\left((x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1) + (x_2-b)^T A^{-1}(x_2-b)\right)\right]}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \\
&= \frac{\exp\left[-\frac{1}{2}\left((x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1) + (x_2-b)^T A^{-1}(x_2-b)\right)\right]}{\sqrt{(2\pi)^d|\Sigma_{11}||A|}} \\
&= \frac{\exp\left[-\frac{1}{2}\left(g(x_1) + g(x_2)\right)\right]}{\sqrt{(2\pi)^d|\Sigma_{11}||A|}} \\
&= \frac{\exp\left[-\frac{1}{2}g(x_1)\right]\exp\left[-\frac{1}{2}g(x_2)\right]}{\sqrt{(2\pi)^d|\Sigma_{11}||A|}} \\
&= \frac{\exp\left[-\frac{1}{2}g(x_1)\right]}{\sqrt{(2\pi)^{d_1}|\Sigma_{11}|}}\ \frac{\exp\left[-\frac{1}{2}g(x_2)\right]}{\sqrt{(2\pi)^{d_2}|A|}} \\
&= \frac{\exp\left[-\frac{1}{2}(x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)\right]}{\sqrt{(2\pi)^{d_1}|\Sigma_{11}|}}\ \frac{\exp\left[-\frac{1}{2}(x_2-b)^T A^{-1}(x_2-b)\right]}{\sqrt{(2\pi)^{d_2}|A|}} \\
&= N(x_1;\mu_1,\Sigma_{11})N(x_2;b,A)
\end{aligned} \tag{13}$$

where $b = \mu_2 + \Sigma_{12}^T\Sigma_{11}^{-1}(x_1 - \mu_1)$ and $A = \Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}$.

(iv)

The conditional distribution of x_2 given x_1 is:

$$\begin{aligned}
f(x_2|x_1) = \frac{f(x_1,x_2)}{f(x_1)} &= \frac{\frac{\exp\left[-\frac{1}{2}(x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)\right]}{\sqrt{(2\pi)^{d_1}|\Sigma_{11}|}}\ \frac{\exp\left[-\frac{1}{2}(x_2-b)^T A^{-1}(x_2-b)\right]}{\sqrt{(2\pi)^{d_2}|A|}}}{\frac{\exp\left(-\frac{1}{2}(x_1-\mu_1)^T{\Sigma_{11}}^{-1}(x_1-\mu_1)\right)}{\sqrt{(2\pi)^d|\Sigma_{11}|}}} \\
&= \frac{\exp\left[-\frac{1}{2}(x_2-b)^T A^{-1}(x_2-b)\right]}{\sqrt{(2\pi)^{d_2}|A|}}
\end{aligned} \tag{14}$$

where $b = \mu_2 + \Sigma_{12}^T\Sigma_{11}^{-1}(x_1 - \mu_1)$ and $A = \Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}$.

(v)

For $\mu_n = 0$ and $\Sigma_{n\times n} = I$, the plots are shown in figure 1. We can see that these functions are not smooth. As shown in figure 2, when $\mu_n = 0$ and $\Sigma_{n\times n}$ is set to all ones matrix, the

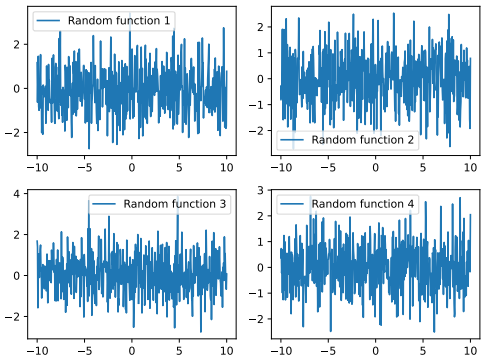

Figure 1: $\mu_n = 0$ and $\Sigma_{n\times n} = I$

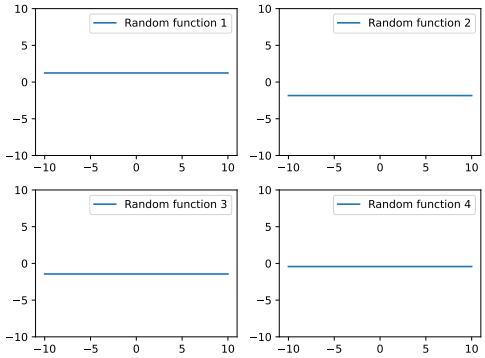

Figure 2: $\mu_n = 0$ and $\Sigma_{n\times n}$: all ones matrix.

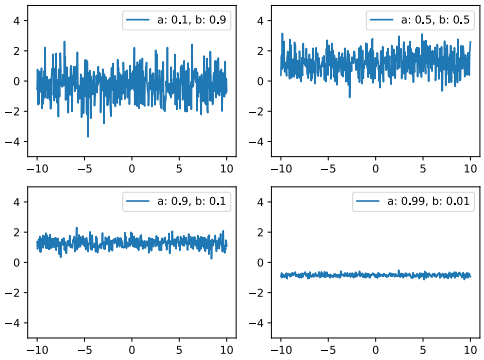

Figure 3: cov = a * np.ones((500,500)) + b * np.eye(500)

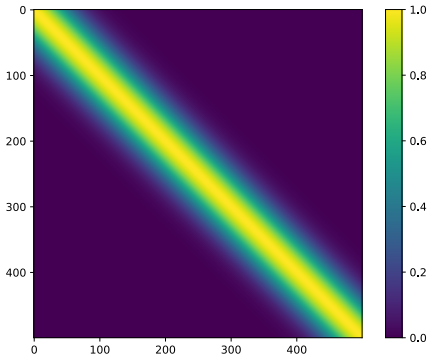

Figure 4: Demo of covariance matrix defined by kernel function k.

functions behave like a horizontal straight line. Figure 3 shows how the smoothness of functions change with Σ. The equation used here to compute Σ is: Σ = a * np.ones((500,500)) + b * np.eye(500). When the covariances of randomly selected samples are large, this means they are highly correlated, which will make the function more smooth. The clear trend can be seen when the value of a increased from 0.1 to 0.99. The functions become more and more smooth.

(vi)

Here we have $\mu_n = 0$ and $\Sigma_{n\times n} = \exp(-\frac{(x_i-x_j)^2}{5}))$. As shown in figure 4, if the samples are near to each other, then the covariance between them is quite large. Therefore, the functions are smooth (figure 5).

(vii)

If one is interested in random periodic functions, the setting of μ and Σ should be: mu should have the periodicity to make sure that the function fluctuates periodically. To make sure the function repeats similarly for every period, Σ can be set to all ones matrix. The results are shown in figure 6.

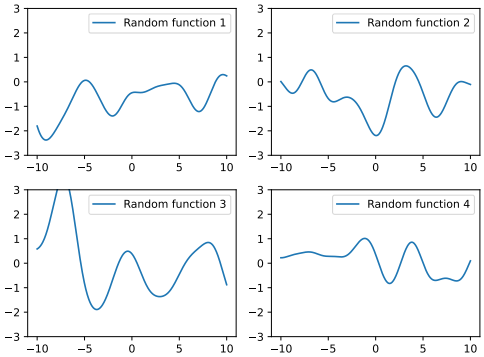

Figure 5: q-vi

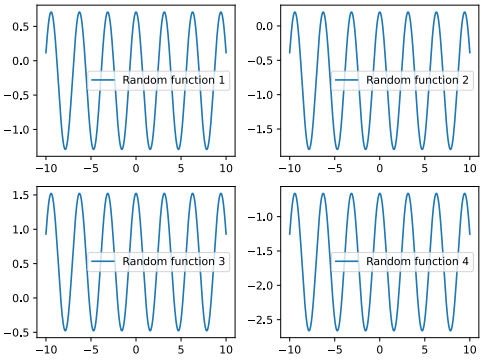

Figure 6: q-vii

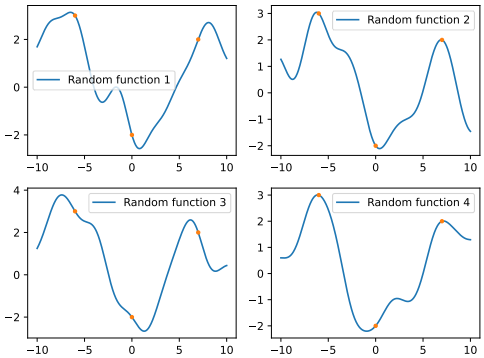

Figure 7: q-ix

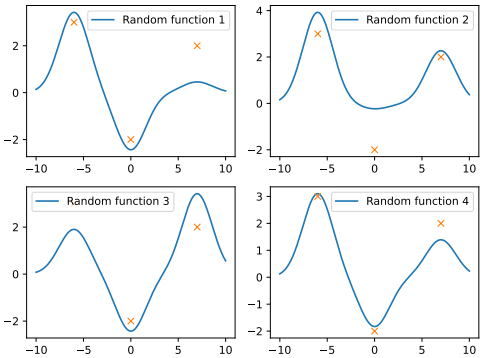

Figure 8: q-x

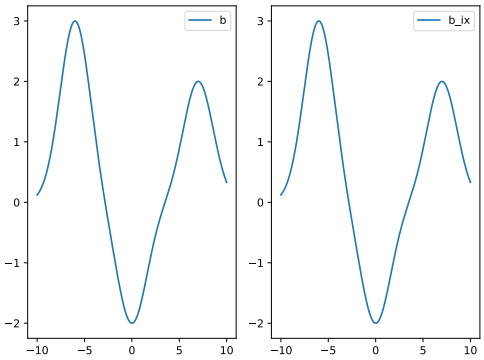

Figure 9: q-xii

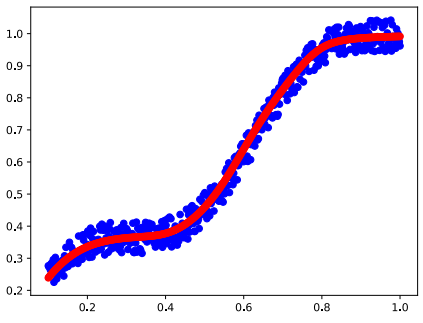

Figure 10: Polynomial function

(viii)

From part (iv), we know the conditional distribution of x_2 given x_1 is: $N(x_2; b, A)$. Therefore, the posterior $Y|\overline{Y}$ is:

$$N(x_2; b, A)$$

where $b = \mu_n + K(\overline{X}, X)^T K(\overline{X}, \overline{X})^{-1}(\bar{Y} - \mu_m)$ and $A = K(X, X) - K(\overline{X}, X)^T K(\overline{X}, \overline{X})^{-1} K(\overline{X}, X)$.

(ix)

As shown in figure 7, we can see the training data points are always on the random functions and the random functions are smooth.

(x)

As shown in figure 8, the training data points are always not on the random functions and the random functions are smooth.

(xi)

As mentioned in (viii), the mean is:

$$b = \mu_n + K(\overline{X}, X)^T K(\overline{X}, \overline{X})^{-1}(\bar{Y} - \mu_m) \tag{15}$$

Here, we have $\mu_n = \mu_m = 0$, so:

$$b = K(\overline{X}, X)^T K(\overline{X}, \overline{X})^{-1}\bar{Y} \tag{16}$$

(xii)

The mean "functions" for parts (ix) and (x) are shown in figure 9 and they are identical.

Problem 2

(i)

$\sigma(x)$ is defined as: $\sigma(x) = \frac{1}{1+e^{-x}}$. The partial derivative of $\sigma(x)$ w.r.t. x is:

$$\begin{aligned}\frac{\partial \sigma(x)}{\partial x} = \frac{\partial (1+e^{-x})^{-1}}{\partial x} &= (-1)(1+e^{-x})^{-2}(-1)(e^{-x}) \\ &= \frac{e^{-x}}{(1+e^{-x})^2} \\ &= \frac{1}{1+e^{-x}}\frac{e^{-x}}{1+e^{-x}} \\ &= \sigma(x)\Big[\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\Big] \\ &= \sigma(x)(1-\sigma(x)) \end{aligned} \tag{17}$$

(ii)

$$E(W,b) = \frac{1}{2n}\sum_{i=1}^{n} ||\sigma(W^T x_i + b) - y_i||^2$$

$$\frac{\partial E}{\partial W} = \frac{\partial E}{\partial \sigma}\frac{\partial \sigma}{\partial A}\frac{\partial A}{\partial W} = \frac{1}{n}\sum_{i=1}^{n}\Big[\big(\sigma(W^T x_i + b) - y_i\big)\sigma(W^T x_i + b)\big(1-\sigma(W^T x_i + b)\big)\Big]x_i \tag{18}$$

$$\frac{\partial E}{\partial b} = \frac{\partial E}{\partial \sigma}\frac{\partial \sigma}{\partial A}\frac{\partial A}{\partial b} = \frac{1}{n}\sum_{i=1}^{n}\Big[\big(\sigma(W^T x_i + b) - y_i\big)\sigma(W^T x_i + b)\big(1-\sigma(W^T x_i + b)\big)\Big] \tag{19}$$

(iii)

(a)

$$\frac{\partial N(x)}{\partial x} = \frac{\partial \sigma_1}{\partial f1}\cdot\frac{\partial f1}{\partial \sigma_2}\cdot\frac{\partial \sigma_2}{\partial f2}\cdot\frac{\partial f2}{\partial \sigma_3}\cdot\ldots\cdot\frac{\partial \sigma_n}{\partial f_n}\cdot\frac{\partial f_n}{\partial \sigma_x} \tag{20}$$

Each item (i) in the equation above takes i $\cdot$ O(1) time to calculate. So in total, we have:

$$\sum_{i=1}^{n} i \cdot O(1) = O(n^2). \tag{21}$$

(b)

We can compute $\frac{\partial f_n}{\partial x}$ first and then compute $\frac{\partial \sigma_n}{\partial f_n}$. Then we will have:

$$\sum_{i=1}^{n} 1 \cdot O(1) = O(n). \tag{22}$$

(iv)

The flexible framework for learning neural networks has been implemented in neural_network_gd_adam.py. The standard gradient descent will be used if no optimizer is defined. tanh activation function is used here and the minibatch has also been implemented to accelerate learning. For example:

```
# Number of epochs
epochs = 1500
# Learning rate
lr = 0.01
# batch size
batch_size = 64

# No optimizer is defined, use standard gradient descent
model1.set_config(epochs=epochs, learning_rate=lr, optimizer=None)

# Define Adam optimizer
adam = Adam(lr)
model2.set_config(epochs=epochs, learning_rate=lr, optimizer=adam)
```

The number of layers can be dynamically changed by having these defined:

```
model1 = Network()
model1.add(Layer(input_feature_dimention, activation='tanh'))
model1.add(Layer(number_layers1, activation='tanh'))
model1.add(Layer(number_layers2, activation='tanh'))
...
...
model1.add(Layer(output_feature_dimention, activation='tanh'))
```

Some simple functions have been used first to test my framework. The first example is, XOR. The input is [[0 0], [0 1], [1 0], [1 1]] and the group truth is [[0], [1], [1], [0]]. Using Adam optimizer with 3 layers (input-65 neurons-output), it only takes 0.14s to get the results [[0.00113234], [0.99719923], [0.99837704], [0.00233957]], which is correct. However, using standard gradient descent with 3 layers (input-65 neurons-output), even though it is faster, but the accuracy is not good ([[0.11079135], [0.42985902], [0.88478167], [0.5512474]]), which means either more neurons or more layers are needed. When 4 layers are used with 256 neurons for the middle two layers, it takes 1.9s to reach the similar accuracy as using Adam optimizer. So we can see using Adam optimizer can definitely reduce the time needed

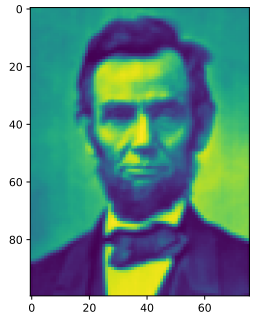

Figure 11: Image 1

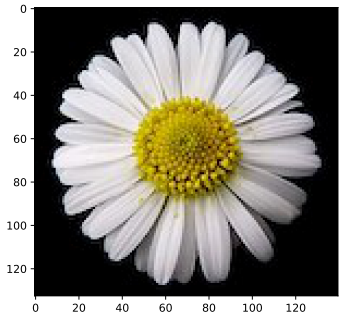

Figure 12: Image 2

to converge.
Another example is a polynomial function. As shown in the figure 10, we can see the neutral network can accurately predict the trend.
Now we are going to use the framework to learn the mapping function.

(v)

Image 1 and 2 are shown in figure 11 and 12.

Image 1

Figure 13, 14, 15, 16, 17, and 18 show the predicted images of image 1. The caption of figures simply means: number of layers - number of neurons 1 - number of neurons 2 - ... - number of epochs - minibatch size.

1. So by comparing figure 16 with figure 17, we can see that if the number of epochs is not large enough, simply increasing number of neurons in the layer will dramatically increase the performance of the network.

2. By comparing figure 14 and 15, we can see that if the number of epochs is large enough (1500 here), simply increasing number of neurons in the layer will not increase the performance of network too much.

3. Then by comparing figure 13 with figure 16, we can see that if the number of epochs is the same, simply increasing number of layers in the framework will dramatically increase the performance of the network.

4. Finally, by comparing figure 13 with figure 14 and figure 16 with figure 18, we can see that if the number of layers and number of neurons are the same, increasing number of epochs will also dramatically increase the performance of the network.

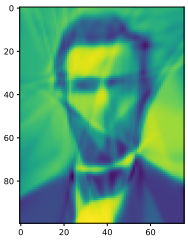

Figure 13: 4-128-128-500-64

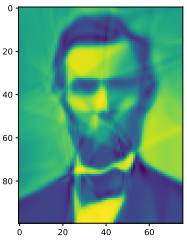

Figure 14: 4-128-128-1500-64

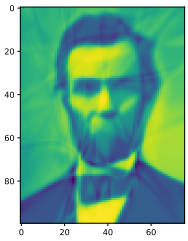

Figure 15: 4-128-256-1500-64

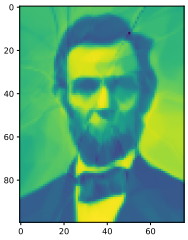

Figure 16: 5-128-128-128-500-64

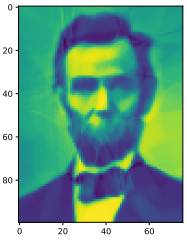

Figure 17: 5-128-128-256-500-64

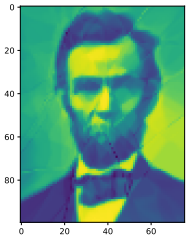

Figure 18: 5-128-128-128-1000-64

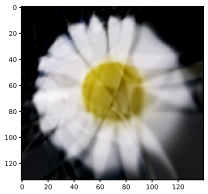

Figure 19: 4-128-128-1000-64

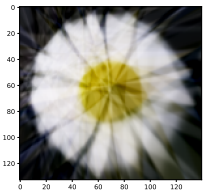

Figure 20: 4-256-256-300-64

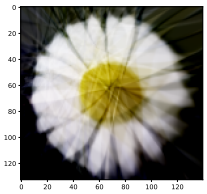

Figure 21: 4-256-256-500-64

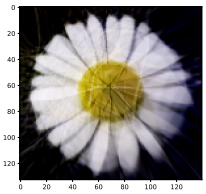

Figure 22: 4-256-256-1500-64

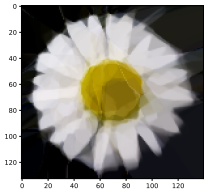

Figure 23: 5-128-128-128-300-64

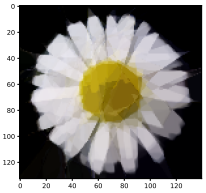

Figure 24: 5-128-128-128-1000-64

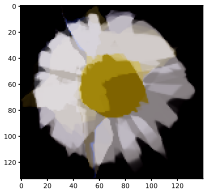

Figure 25: 5-256-256-256-300-64

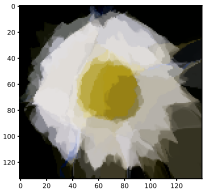

Figure 26: 6-256-256-256-256-300-64

Image 2

Basically, all the previous findings hold true. Besides that, I also found that since image 2 contains more information than image 1, so number of epochs has the most important impact on the performance of network.

Conclusion

In conclusion, figure 18 and 22 give the best results.

Appendix

Code

- neural_network_gd_adam.py
- learn_pattern.py
- learn_simple_functions.py

References

- Definition of marginal distribution.
- PDF of multivariate Gaussian distribution.
- General idea of the neural network framework and Adam optimizer.

Useful discussions with

- Yueyue Ma
- Yiwei Wang