The Australian National University Semester 1, 2022
School of Computing Assignment 2
Group member name(s): Ziyang Chen, Han Zhang
Group member UID(s): u6908560, u7235649

COMP4670/8600: Statistical Machine Learning

Contribution statement:
Ziyang Chen (u6908560) mainly contributed on Section 1.
Han Zhang (u7235649) mainly contributed on Section 2.

Answer to Question 1.1

Because X is continuous, therefore, we can get following equation in distribution $\mathcal{D}$ according to Markov's Inequality and the given equation (1.1) in question sheet.

$$\mathcal{R}_{\mathcal{D}}[f] = \mathbb{E}_{\mathcal{D}}[\ell(f(X), Y)] = \iint \ell(f(x), y)p(x, y)\mathrm{d}x\mathrm{d}y \tag{1}$$

where $q(x, y)$ is the density function of $\mathcal{D}$.

Combining the given equation $0 \leq \ell(y, y') \leq 1$, we can get

$$\iint \ell(f(x), y)p(x, y)\mathrm{d}x\mathrm{d}y \leq \iint p(x, y)\mathrm{d}x\mathrm{d}y \tag{2}$$

Therefore, we can get following equation according to equation 1 and equation 2.

$$\mathcal{R}_{\mathcal{D}}[f] \leq \iint p(x, y)\mathrm{d}x\mathrm{d}y \tag{3}$$

Similarly, we can get the following equation in distribution $\widetilde{\mathcal{D}}$

$$\mathcal{R}_{\widetilde{\mathcal{D}}}[f] \leq \iint \widetilde{p}(x, y)\mathrm{d}x\mathrm{d}y \tag{4}$$

where $\widetilde{p}(x, y)$ is the density function of $\widetilde{\mathcal{D}}$.

Combining equation 3 and equation 4, we can get

$$\left|\mathcal{R}_{\mathcal{D}}[f] - \mathcal{R}_{\widetilde{\mathcal{D}}}[f]\right| \leq \iint \left|p(x, y) - \widetilde{p}(x, y)\right| \mathrm{d}x\mathrm{d}y \tag{5}$$

Therefore, according to the given equation (1.6) about the definition of $D_{TV}(p, \widetilde{p})$, we can get

$$\left|\mathcal{R}_{\mathcal{D}}[f] - \mathcal{R}_{\widetilde{\mathcal{D}}}[f]\right| \leq D_{TV}(p, \widetilde{p}) \tag{6}$$

Answer to Question 1.2

Using the triangle inequality, we can get:

$$\left|\mathcal{R}_{\mathcal{D}}[f] - \widehat{\mathcal{R}}_{\widetilde{\mathcal{S}}}[f]\right| = \left|\mathcal{R}_{\mathcal{D}}[f] - \mathcal{R}_{\widetilde{\mathcal{D}}}[f] + \mathcal{R}_{\widetilde{\mathcal{D}}}[f] - \widehat{\mathcal{R}}_{\widetilde{\mathcal{S}}}[f]\right| \leq \left|\mathcal{R}_{\mathcal{D}}[f] - \mathcal{R}_{\widetilde{\mathcal{D}}}[f]\right| + \left|\mathcal{R}_{\widetilde{\mathcal{D}}}[f] - \widehat{\mathcal{R}}_{\widetilde{\mathcal{S}}}[f]\right| \tag{7}$$

According to the above proof in Question 1.1, we can have $\left|\mathcal{R}_{\mathcal{D}}[f] - \mathcal{R}_{\widetilde{\mathcal{D}}}[f]\right| \leq D_{TV}(p, \widetilde{p})$, therefore, combining equation 7, we can get

$$\left|\mathcal{R}_{\mathcal{D}}[f] - \widehat{\mathcal{R}}_{\widetilde{\mathcal{S}}}[f]\right| \leq D_{TV}(p, \widetilde{p}) + \left|\mathcal{R}_{\widetilde{\mathcal{D}}}[f] - \widehat{\mathcal{R}}_{\widetilde{\mathcal{S}}}[f]\right| \tag{8}$$

According to the equation (A, 1): $\Pr\{|\hat{\mu} - \mu_Z| \geq \epsilon\} \leq 2\exp(-2N\epsilon^2)$ in Proposition 1, μ_Z is the true mean and $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} Z_i$ is the sample mean. Therefore, μ_Z and $\hat{\mu}$ correspond to $\mathcal{R}_{\widetilde{\mathcal{D}}}[f]$ and $\widehat{\mathcal{R}}_{\widetilde{\mathcal{S}}}[f]$ in equation 8.

Therefore, equation (A, 1) can be written

$$\Pr\{\left|\widehat{\mathcal{R}}_{\widetilde{\mathcal{S}}}[f] - \mathcal{R}_{\widetilde{\mathcal{D}}}[f]\right| \geq \epsilon\} \leq 2\exp(-2N\epsilon^2) \tag{9}$$

It also can be written

$$\Pr\{\left|\mathcal{R}_{\widetilde{\mathcal{D}}}[f] - \widehat{\mathcal{R}}_{\widetilde{\mathcal{S}}}[f]\right| \leq \epsilon\} \geq 1 - 2\exp(-2N\epsilon^2) \tag{10}$$

Let $\delta = 2\exp(-2N\epsilon^2)$, we can get $\epsilon = \sqrt{\frac{\log(\frac{2}{\delta})}{2N}}$. Through loosing bound, we can get $\epsilon \leq \sqrt{\frac{\log(\frac{1}{\delta})}{N}}$. Therefore, the equation 10 can be changed into

$$\Pr\{\left|\mathcal{R}_{\widetilde{\mathcal{D}}}[f] - \widehat{\mathcal{R}}_{\widetilde{\mathcal{S}}}[f]\right| \leq \sqrt{\frac{\log(\frac{1}{\delta})}{N}}\} \geq 1 - \delta \tag{11}$$

According to the model complexity and uniform convergence, we can get, with probability $1 - \delta$, $\forall f \in \mathcal{H} : \triangle_{gen}(f) =\leq \sqrt{\frac{\ln(|\mathcal{H}|)+\ln(\frac{1}{\delta})}{N}}$.

So, we can have

$$\Pr\{\left|\mathcal{R}_{\widetilde{\mathcal{D}}}[f] - \widehat{\mathcal{R}}_{\widetilde{\mathcal{S}}}[f]\right| \leq \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(\frac{1}{\delta})}{N}}\} \geq 1 - \delta \tag{12}$$

Therefore, adding $D_{TV}(p, \widetilde{p})$ into two sides of the equation within curly bracket, we can get

$$\Pr\{D_{TV}(p, \widetilde{p}) + \left|\mathcal{R}_{\widetilde{\mathcal{D}}}[f] - \widehat{\mathcal{R}}_{\widetilde{\mathcal{S}}}[f]\right| \leq D_{TV}(p, \widetilde{p}) + \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(\frac{1}{\delta})}{N}}\} \geq 1 - \delta \tag{13}$$

Combining equation 8 and equation 13, we can get

$$\Pr\{\left|\mathcal{R}_{\mathcal{D}}[f] - \widehat{\mathcal{R}}_{\widetilde{\mathcal{S}}}[f]\right| \leq D_{TV}(p, \widetilde{p}) + \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(\frac{1}{\delta})}{N}}\} \geq 1 - \delta \tag{14}$$

Answer to Question 1.3

Comparing with the equation 14 and the given equation without noisy data, we can see the noisy data will increase the generalisation bound and the difference of them will be $D_{TV}(p, \widetilde{p})$, which is the risk distance between true data and noisy data. The rate will increase when it decreases with the number of data points.

Answer to Question 1.4

For a noisy point $y \in \mathcal{Y}$, it may be got from 2 situations.

1. it was $-y$ before, when the label noise happened, it changed into y.
2. it was y before and there is no change when the label noise happened;

For the situation 1, we can get the probability is

$$\Pr(\widetilde{Y} = y|Y = -y) \cdot \Pr(Y = -y|X) \tag{15}$$

For the situation 2, we can get the probability is

$$\Pr(\widetilde{Y} = y|Y = y) \cdot \Pr(Y = y|X) \tag{16}$$

Therefore,

$$\begin{aligned}\Pr(\widetilde{Y} = y|X) &= \Pr(\widetilde{Y} = y|Y = -y) \cdot \Pr(Y = -y|X) + \Pr(\widetilde{Y} = y|Y = y) \cdot \Pr(Y = y|X) \\ &= \Pr(\widetilde{Y} = y|Y = -y) \cdot (1 - \Pr(Y = y|X))(1 - \Pr(\widetilde{Y} = -y|Y = y)) \cdot \Pr(Y = y|X)\end{aligned} \tag{17}$$

Because the label flipping probability $\Pr(\widetilde{Y} = -y|Y = y) = \sigma_y$, when $-y$ changes into y, the label flipping probability in this situation is $\Pr(\widetilde{Y} = y|Y = -y) = \sigma_{-y}$.

Therefore, equation 17 can be written as

$$\Pr(\widetilde{Y} = y|X) = \sigma_{-y} \cdot (1 - \Pr(Y = y|X)) + (1 - \sigma_y) \cdot \Pr(Y = y|X) \tag{18}$$

Answer to Question 1.5

According to the Question 1.2, we already have equation 14. Combining the equation (1.6) in question sheet, we can get

$$\Pr\{\left|\mathcal{R}_{\mathcal{D}}[f] - \widehat{\mathcal{R}}_{\widetilde{\mathcal{S}}}[f]\right| \leq \iint |p(x,y) - \widetilde{p}(x,y)| \, \mathrm{d}x\mathrm{d}y + \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(\frac{1}{\delta})}{N}}\} \geq 1 - \delta \tag{19}$$

In this question, we can get

$$p(x,y) = \Pr(X = x, Y = y) = \Pr(X)\Pr(Y = y|X = x) = \Pr(X)\Pr(Y = y|X) \tag{20}$$

$$\widetilde{p}(x,y) = \Pr(\widetilde{X} = x, \widetilde{Y} = y) = \Pr(\widetilde{X})\Pr(\widetilde{Y} = y|\widetilde{X} = x) = \Pr(\widetilde{X})\Pr(\widetilde{Y} = y|\widetilde{X}) \tag{21}$$

According to the **LN2** property, we can know for all $x \in \mathcal{X}$

$$\Pr(\widetilde{X} = x) = \Pr(X = x) \quad \textit{similarly} \quad \Pr(\widetilde{X}) = \Pr(X) \tag{22}$$

which can imply

$$\Pr(\widetilde{X} = x|\widetilde{Y} = y) = \Pr(X = x|\widetilde{Y} = y) \quad \textit{similarly} \quad \Pr(\widetilde{X}|\widetilde{Y} = y) = \Pr(X|\widetilde{Y} = y) \tag{23}$$

Using the Bayes theorem and combining equation 22 and equation 23, we can get

$$\Pr(\widetilde{Y} = y|\widetilde{X}) = \frac{\Pr(\widetilde{Y} = y) \cdot \Pr(\widetilde{X}|\widetilde{Y} = y)}{\Pr(\widetilde{X})} = \frac{\Pr(\widetilde{Y} = y) \cdot \Pr(X|\widetilde{Y} = y)}{\Pr(X)} = \Pr(\widetilde{Y} = y|X) \tag{24}$$

Therefore, combining equation 20, equation 21, equation 23, equation 24 and the equation 18 in

question 1.4, and $\sigma_y = \sigma_{-y} = \sigma$ in the symmetric label noise, we can get

$$\begin{aligned} p(x,y) - \widetilde{p}(x,y) &= \Pr(X) \cdot \Pr(Y = y|X) - \Pr(\widetilde{X}) \cdot \Pr(\widetilde{Y} = y|\widetilde{X}) \\ &= \Pr(X) \cdot \Pr(Y = y|X) - \Pr(X) \cdot \Pr(\widetilde{Y} = y|X) \\ &= \Pr(X) \cdot \Pr(Y = y|X) - \Pr(X) \cdot [\sigma_{-y} \cdot (1 - \Pr(Y = y|X)) + (1 - \sigma_y) \cdot \Pr(Y = y|X)] \\ &= \Pr(X) \cdot \Pr(Y = y|X) - \Pr(X) \cdot [\sigma \cdot (1 - \Pr(Y = y|X)) + (1 - \sigma) \cdot \Pr(Y = y|X)] \\ &= \Pr(X) \cdot (2\sigma \Pr(Y = y|X) - \sigma) \\ &= 2\sigma \Pr(X) \Pr(Y = y|X) - \sigma \Pr(X) \end{aligned} \tag{25}$$

Because $0 \leq \Pr \leq 1$ and $0 \leq \Pr(Y = y|X) \leq 1$. Thus, we can get

$$p(x,y) - \widetilde{p}(x,y) = 2\sigma \Pr(X) \Pr(Y = y|X) - \sigma \Pr(X) \leq 2\sigma \tag{26}$$

Therefore, according to $0 \leq \sigma \leq 1$, we can get

$$\iint |p(x,y) - \widetilde{p}(x,y)| \, \mathrm{d}x\mathrm{d}y \leq \iint 2\sigma \mathrm{d}x\mathrm{d}y = 2\sigma \tag{27}$$

Combining the equation 19, we can get

$$\Pr\{\left|\mathcal{R}_{\mathcal{D}}[f] - \widehat{\mathcal{R}}_{\widetilde{\mathcal{S}}}[f]\right| \leq 2\sigma + \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(\frac{1}{\delta})}{N}}\} \geq 1 - \delta \tag{28}$$

Answer to Question 2.1

The *Matern32* and *Matern52* functions in *kernels.py* have been implemented and submitted.

Answer to Question 2.2

The *optimisation()*, *fit()*, *predict()* and *log_marginal_likelihood()* in *gp.py* have been implemented and submitted.

Answer to Question 2.3

According to the question sheet, we can get

$$f(\mathbf{x}) = \mu(\mathbf{x}) + \sigma(\mathbf{x})Z \tag{29}$$

where $\mu(\mathbf{x})$ is the mean of the $f(\mathbf{x})$ and $\sigma(\mathbf{x})$ is the standard deviation of the $f(\mathbf{x})$.
Seeing the equation (2.11), we can get

$$EI(\mathbf{x}) = \mathbb{E}(I(\mathbf{x})) = \int_{-\infty}^{+\infty} I(\mathbf{x}) \Pr(Z) \mathrm{dZ} = \int_{-\infty}^{+\infty} I(\mathbf{x}) \phi(Z) \mathrm{dZ} \tag{30}$$

where $\phi(\cdot)$ is the normal probability distribution functions, which is

$$\phi(Z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-Z^2}{2}} \tag{31}$$

According to the equation (2.8), let $I(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{x}^+) - \xi = 0$. Combining to equation 29, we can get

$$Z_0 = \frac{f(\mathbf{x}^+) + \xi - \mu(\mathbf{x})}{\sigma(\mathbf{x})} \tag{32}$$

Continuing in equation 30, we can get

$$\begin{aligned}\mathbb{E}(I(\mathbf{x})) &= \int_{-\infty}^{+\infty} I(\mathbf{x})\phi(Z)\mathrm{dZ} \\ &= \int_{-\infty}^{Z_0} I(\mathbf{x})\phi(Z)\mathrm{dZ} + \int_{Z_0}^{+\infty} I(\mathbf{x})\phi(Z)\mathrm{dZ}\end{aligned} \tag{33}$$

Because $I(\mathbf{x}) = 0$, $\int_{-\infty}^{Z_0} I(\mathbf{x})\phi(Z)\mathrm{dZ} = 0$. Therefore, we can write equation 33 as

$$\begin{aligned}\mathbb{E}(I(\mathbf{x})) &= \int_{Z_0}^{+\infty} I(\mathbf{x})\phi(Z)\mathrm{dZ} \\ &= \int_{Z_0}^{+\infty} (\mu(\mathbf{x}) + \sigma(\mathbf{x})Z - f(\mathbf{x}^+) - \xi)\phi(Z)\mathrm{dZ} \\ &= \int_{Z_0}^{+\infty} (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)\phi(Z)\mathrm{dZ} + \int_{Z_0}^{+\infty} (\sigma(\mathbf{x})Z)\phi(Z)\mathrm{dZ} \\ &= (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)\int_{Z_0}^{+\infty} \phi(Z)\mathrm{dZ} + \sigma(\mathbf{x})\int_{Z_0}^{+\infty} Z\phi(Z)\mathrm{dZ} \\ &= (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)(1 - \Phi(Z_0)) + \sigma(\mathbf{x})\int_{Z_0}^{+\infty} Z\frac{1}{\sqrt{2\pi}}e^{\frac{-Z^2}{2}}\mathrm{dZ} \\ &= (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)(1 - \Phi(Z_0)) + \frac{\sigma(\mathbf{x})}{\sqrt{2\pi}}\int_{Z_0}^{+\infty} Ze^{\frac{-Z^2}{2}}\mathrm{dZ} \\ &= (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)(1 - \Phi(Z_0)) + \frac{\sigma(\mathbf{x})}{\sqrt{2\pi}}[e^{\frac{-Z^2}{2}}]_{Z_0}^{+\infty} \\ &= (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)(1 - \Phi(Z_0)) + \sigma(\mathbf{x})\phi(Z_0)\end{aligned} \tag{34}$$

where $\Phi(\cdot)$ is the normal cumulative distribution functions.

According to the equation (B.6) in question sheet, we can get $1 - \Phi(-Z) = \Phi(Z)$. Because $\Phi(\cdot)$ is the normal cumulative distribution functions, we can also get $\Phi(Z) = \Phi(-Z)$. Therefore, we can get the following equation from equation 34

$$\mathbb{E}(I(\mathbf{x})) = (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)\Phi(Z_0) + \sigma(\mathbf{x})\phi(Z_0) \tag{35}$$

Combining all the information and equation (2.10) in question sheet, we can get

$$\mathbb{E}(I(\mathbf{x})) = \begin{cases} (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)\Phi(Z) + \sigma(\mathbf{x})\phi(Z) & if \;\; \sigma(\mathbf{x}) > 0 \\ 0 & if \;\; \sigma(\mathbf{x}) = 0 \end{cases} \tag{36}$$

where

$$Z = \begin{cases} \frac{\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi}{\sigma(\mathbf{x})} & if \;\; \sigma(\mathbf{x}) > 0 \\ 0 & if \;\; \sigma(\mathbf{x}) = 0 \end{cases} \tag{37}$$

Answer to Question 2.4

The Probability of Improvement (PI) and Expected Improvement (EI) acquisitions functions in *acquisitions.py* have been implemented and submitted.

Answer to Question 2.5

Answer to Question 2.6

It is usually difficult to find the optimal hyperparameter combination among complex hyperparameters when the dimensions are high. The hyperparameter optimization approach can make this process easy.