Assignment 2
COMP8600

May 15, 2022

Name: Xuecheng Zhang (**u6284513**),
Yichao Jiang (**u7236045**)

03/05/2022

1 Question 1

Answer to Question 1.1

$$R_D[f] = E_{(X,Y)\sim D}[l(f(x), y)] = \iint P(x, y)l[f(x), y]dxdy \tag{1}$$

$$R_{\tilde{D}}[f] = E_{(X,Y)\sim \tilde{D}}[l(f(x), y)] = \iint \tilde{P}(x, y)l[f(x), y]dxdy \tag{2}$$

$$R_D[f] - R_{\tilde{D}}[f] = \iint (P(x, y) - \tilde{P}(x, y))l[f(x), y]dxdy \tag{3}$$

$$\tag{4}$$

According to the text, the loss function l is bounded between 0 and 1 on its domain.

$$0 < l[f(x), y] < 1 \tag{5}$$

$$\therefore |R_p[f] - R_{\tilde{y}}[f]| \leq \iint |P(x, y) - \tilde{P}(x, y)|dxdy = D_{TV}(p, \tilde{p}) \tag{6}$$

Answer to Question 1.2

Link the question to 1.1 and use the triangle inequality.

$$\left|R_D[f] - \hat{R}_{\widetilde{S}}[f]\right| \Leftrightarrow \left|R_D[f] - R_{\tilde{D}}[f] + R_{\tilde{D}}[f] - \hat{R}_{\widetilde{S}}[f]\right| \tag{7}$$

$$\leqslant |R_D[f] - R_{\tilde{D}}[f]| + |R_{\tilde{D}}[f] - \hat{R}_{\widetilde{S}}[f]| \tag{8}$$

$$\leqslant \mathrm{D}_{TV}\left(p, \tilde{p}\right) + \left|R_{\widetilde{D}}[f] - \hat{R}_{\widetilde{S}}[f]\right| \tag{9}$$

For the second term,we can use Hoeffding's inequality:

$$\Pr\left\{\forall f \in \mathcal{H}, \left|R_{\widetilde{D}}[f] - \hat{R}_{\widetilde{S}}[f]\right| < \varepsilon\right\} \Leftrightarrow 1 - \Pr\left\{\exists f \in \mathcal{H}, \left|R_{\widetilde{D}}[f] - \hat{R}_{\widetilde{S}}[f]\right| \geq \varepsilon\right\} \tag{10}$$

$$= 1 - \Pr\left\{\bigcup_{f\in\mathcal{H}} \left|R_{\widetilde{D}}[f] - \hat{R}_{\widetilde{S}}[f]\right| \geq \varepsilon\right\} \tag{11}$$

$$\geq 1 - \Pr\left\{\sum_{f\in\mathcal{H}} \left|R_{\widetilde{D}}[f] - \hat{R}_{\widetilde{S}}[f]\right| \geq \varepsilon\right\} \tag{12}$$

$$\geq 1 - 2|\mathcal{H}|exp(-2N\epsilon^2) = 1 - \delta \tag{13}$$

$$\text{Set}\, \varepsilon = \sqrt{\frac{\ln(2|\mathcal{H}|) + \ln(\frac{1}{\delta})}{2N}} \xrightarrow[\frac{2}{|\mathcal{H}|} \leq \frac{1}{\delta}]{\text{Looser bound}} \sqrt{\frac{\ln|\mathcal{H}| + \ln\left(\frac{1}{\delta}\right)}{N}} \tag{14}$$

Then we could tell

$$\Pr\left\{\forall f \in \mathcal{H}, \left|R_{\widetilde{D}}[f] - \hat{R}_{\widetilde{S}}[f]\right| \leq \varepsilon\right\} \geq 1 - \delta \tag{15}$$

$$\Pr\left\{\forall f \in \mathcal{H} : \left|R_D[f] - \hat{R}_{\widetilde{S}}[f]\right| \leq \mathrm{D}_{TV}(p, \tilde{p}) + \sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{N}}\right\} \geq 1 - \delta.\ \text{Since}(\frac{2}{|\mathcal{H}|} \geq \frac{1}{\delta}) \tag{16}$$

Answer to Question 1.3

It is clearly the generalisation bound will be larger due to the noisy data since compare to the :

$$\Pr\left\{\forall f \in \mathcal{H} : \left|R_D[f] - \hat{R}_{\widetilde{S}}[f]\right| \leq \sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{N}}\right\} \geq 1 - \delta. \tag{17}$$

We always have $\mathrm{D}_{TV}(p, \tilde{p})$ on the right side after adding the noisy data,and the

$$\iint |P(x,y) - \tilde{P}(x,y)| dxdy = D_{TV}(p, \tilde{p}) \tag{18}$$

is always greater than zero. Since the training error is non-trivial and the empirical risk is relatively large.$|\Delta_{gen} \leq \sqrt{\frac{\frac{\ln|\mathcal{H}|}{\delta}}{N}}|$.The denominator is constant if the $\mathcal{H}$ is constant.So the rate decrease with the number of data points will be $\sqrt{\frac{1}{N}}$.

Answer to Question 1.4

$$\begin{aligned}
\Pr(\hat{Y} = y \mid X) &= \frac{\Pr(\hat{Y} = y, X)}{\Pr(X)} && (19)\\
&= \frac{\int_Y \Pr(\hat{Y} = y, X, Y)}{\Pr(X)} && (20)\\
&= \frac{\int_Y \Pr(\hat{Y} = y, X|Y)\Pr(Y)}{\Pr(X)} && (21)\\
&\overset{LN3}{=} \frac{\int_Y \Pr(\hat{Y} = y \mid Y)\Pr(X \mid Y)\Pr(Y)}{\Pr(X)} && (22)\\
&= \int_Y \Pr(\hat{Y} = y \mid Y)\Pr(Y \mid X) && (23)\\
&= \underbrace{(1 - \sigma_y) \cdot \Pr(Y = y \mid X)}_{Y=y\text{,and not slip}} + \underbrace{\sigma_{-y} \cdot (1 - \Pr(Y = y \mid X))}_{Y=-y\text{,filp}}. && (24)
\end{aligned}$$

Answer to Question 1.5

Start by bounding use the triangle inequality:

$$\left|R_D[f] - \hat{R}_{\tilde{S}}[f]\right| \leqslant |R_D[f] - R_{\tilde{D}}[f]| + |R_{\tilde{D}}[f] - \hat{R}_s[f]| \tag{25}$$

We have solved the second term, so scale for the first term, let us first prove the below equation, when $p(\tilde{x}) = p(x)$:

$$\begin{aligned}
p(\tilde{y} \mid \tilde{x}) = \frac{p(\tilde{x} \mid \tilde{y})p(\tilde{y})}{p(\tilde{x})} &= \frac{p(x \mid \tilde{y})p(\tilde{y})}{p(x)} && (26)\\
&= \frac{p(x, \tilde{y})}{p(x)} && (27)\\
&= p(\tilde{y} \mid x) && (28)
\end{aligned}$$

Scale for the first term:

$$\begin{aligned}
|R_D[f] - R_{\tilde{D}}[f]| \leqslant \iint |p(x,y) - \tilde{p}(x,y)| dxdy \Leftrightarrow &\int |p(x,y) - p(\tilde{x}, \tilde{y})| dxdy && (29)\\
= &\int |p(x)(p(y \mid x) - p(\tilde{y} \mid x)) \mid dxdy && (30)
\end{aligned}$$

Due to 1.4 we have:

$$p(\tilde{y} \mid x) \overset{\sigma_y=\sigma_{-y}}{=} (1-\sigma)p(y \mid x) + \sigma(1 - p(y \mid x)) \tag{31}$$

so we can expand the equation:

$$\begin{aligned}
&\iint |p(x)(p(y \mid x) - p(\tilde{y} \mid x))|dxdy && (32)\\
&= \iint |p(x)(p(y \mid x) + 2\sigma p(y \mid x) - \sigma - p(y \mid x))|dxdy && (33)\\
&= 2\sigma \underbrace{\iint |p(x,y) - p(x)|dxdy}_{\text{it is smaller than 1}} && (34)\\
&\leqslant 2\sigma && (35)
\end{aligned}$$

Combine with question 1.2. we could tell:

$$\Pr\left\{\forall f \in \mathcal{H} : |R_D[f] - \hat{R}_{\tilde{S}}[f]| \leq 2\sigma + \sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{N}}\right\} \geq 1 - \delta.\ (\frac{2}{|\mathcal{H}|} \geq \frac{1}{\delta}) \tag{36}$$

2 Question 2

Answer to Question 2.1

See code.

Answer to Question 2.2

See code.

Answer to Question 2.3

For the Gaussian Process, the function value at a given point $\mathbf{x}$ can be considered as a normal distribution with mean $\mu(\mathbf{x})$ and variance $\sigma^2(\mathbf{x})$, i.e. $f(\mathbf{x}) \backsim \mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$. Thus,

$$\mathbb{E}[I(x)] = \mathbb{E}_{f(\mathbf{x})\backsim\mathcal{N}(\mu(\mathbf{x}),\sigma^2(\mathbf{x}))}[I(\mathbf{x})] \tag{37}$$

We can perform the reparameterisation trick like VAE so that $f(\mathbf{x}) = \mu(\mathbf{x}) + \sigma(\mathbf{x})\epsilon$ where $\epsilon \backsim \mathcal{N}(0,1)$. Thus, we can rewrite (37) as belows:

$$\mathbb{E}[I(x)] = \mathbb{E}_{\epsilon\backsim\mathcal{N}(0,1)}[I(\mathbf{x})] \tag{38}$$

The expected improvement (EI) is given by integral:

$$\mathbb{E}[I(x)] = \int_{-\infty}^{+\infty} I(\mathbf{x})\phi(\epsilon)d\epsilon = \int_{-\infty}^{+\infty} \max\{0, f(\mathbf{x}) - f(\mathbf{x}^+) - \xi\}\phi(\epsilon)d\epsilon$$

where $\phi(x)$ is the probability distribution function.

When $f(\mathbf{x}) - f(\mathbf{x}^+) - \xi = 0$,

$$\begin{aligned}
f(\mathbf{x}) &= f(\mathbf{x}^+) + \xi\\
\mu(\mathbf{x}) + \sigma(\mathbf{x})\epsilon &= f(\mathbf{x}^+) + \xi
\end{aligned}$$

1. If $\sigma(\mathbf{x}) = 0 \implies f(\mathbf{x}) = \mu(\mathbf{x})$ and $\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi = 0\ \forall \epsilon$, i.e. $f(\mathbf{x}) - f(\mathbf{x}^+) - \xi = 0\ \forall\epsilon$. Thus

$$\mathbb{E}[I(x)] = \int_{-\infty}^{+\infty} 0 d\epsilon = 0 \tag{39}$$

2. If $\sigma(\mathbf{x}) > 0$

$$\epsilon = \frac{f(\mathbf{x}^+) + \xi - \mu(\mathbf{x})}{\sigma(\mathbf{x})} = \epsilon_0 \tag{40}$$

We set $\epsilon_0 = \frac{f(\mathbf{x}^+)+\xi-\mu(\mathbf{x})}{\sigma(\mathbf{x})}$, followed by (39)

$$\begin{aligned}
\mathbb{E}[I(x)] &= \int_{-\infty}^{\epsilon_0} 0 \cdot \phi(\epsilon) d\epsilon + \int_{\epsilon_0}^{+\infty} \big(f(\mathbf{x}) - f(\mathbf{x}^+) - \xi\big)\, \phi(\epsilon) d\epsilon \\
&= \int_{\epsilon_0}^{+\infty} \big(\mu(\mathbf{x}) + \sigma(\mathbf{x})\epsilon - f(\mathbf{x}^+) - \xi\big)\, \phi(\epsilon) d\epsilon \\
&= \big(\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi\big) \int_{\epsilon_0}^{+\infty} \phi(\epsilon) d\epsilon + \int_{\epsilon_0}^{+\infty} \sigma(\mathbf{x})\epsilon\phi(\epsilon) d\epsilon \\
&= L1 + L2
\end{aligned} \tag{41}$$

Since $\int_{\epsilon_0}^{+\infty} \phi(\epsilon) d\epsilon = 1 - \Phi(\epsilon_0) = \Phi(-\epsilon_0) = \Phi(Z)$ where Z=$\frac{\mu(\mathbf{x})-f(\mathbf{x}^+)-\xi}{\sigma(\mathbf{x})}$, $\Phi(x)$ is the normal cumulative function, we have $L1 = (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)\, \Phi(Z)$.

Moreover, we can simplify L2 as below:

$$\begin{aligned}
L2 &= \frac{\sigma(\mathbf{x})}{\sqrt{2\pi}} \int_{\epsilon_0}^{+\infty} \epsilon \cdot \exp(-1/2\epsilon^2) d\epsilon \\
&= -\frac{\sigma(\mathbf{x})}{\sqrt{2\pi}} \int_{\epsilon_0}^{+\infty} \left(\frac{d}{d\epsilon}\exp(-1/2\epsilon^2)\right) d\epsilon \\
&= -\frac{\sigma(\mathbf{x})}{\sqrt{2\pi}} \exp(-1/2\epsilon^2)\Big|_{\epsilon_0}^{+\infty} \\
&= \frac{\sigma(\mathbf{x})}{\sqrt{2\pi}} \exp(-1/2\epsilon_0^2) \\
&= \frac{\sigma(\mathbf{x})}{\sqrt{2\pi}} \exp(-1/2(-\epsilon_0)^2) \\
&= \sigma(\mathbf{x})\phi(-\epsilon_0) \\
&= \sigma(\mathbf{x})\phi(Z)
\end{aligned} \tag{42}$$

Thus, $L1 + L2 = (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)\, \Phi(Z) + \sigma(\mathbf{x})\phi(Z)$.

$$\mathbb{E}[I(x)] = \begin{cases} (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)\, \Phi(Z) + \sigma(\mathbf{x})\phi(Z) & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases} \tag{43}$$

Answer to Question 2.4

See code.

Answer to Question 2.5

```
Algorithm 1: Scheduler $\xi$
input  :  n_iter, i, $\xi^{(0)}$, $\xi^{(i)}$
/* where i presents number of previous iterations                    */
initialise $X_{sample}$, $Y_{sample}$, decay_point= 0.7*n_iter, $\xi_{max} = 1.5$, $\xi_{min} = 5 \times 10^{-3}$
/* In the exploration phase                                           */
if $i <$ decay_point - 1 then
    /* increment $\xi$ until at $\xi_{max}$ point                     */
    $\xi^{(i+1)} \leftarrow \xi^{(i)} + (\xi_{max} - \xi^{(0)})/$(decay_point-1)
else
    /* In the exploitation phase                                      */
    /* $\alpha$ is a value where sets the last value to $5 \times 10^{-3}$    */
    $\alpha = \ln(\xi_{min}/\xi_{max})$/-(n_iter-2)
    /* When $i \geq$ decay_point, we decay $\xi$ with $\xi = \xi_{max} * e^{-\alpha i}$   */
    $\xi^{(i+1)} \leftarrow \xi_{max}$ * exp($\alpha *$ decay_point $- i$)
output: $\xi^{(i+1)}$
```

The ξ scheduler is shown on the above algorithm. At first, ξ is initialised as a small value ξ_0 so that $f(x^+)$ is not influenced by ξ, which may produce a better result. In the subsequent iterations (1-7 iterations), we increase the value of ξ from ξ_0 to ξ_{max} so that the sampling point can fall into uncertainty areas which is to prevent stopping at the local minima. We forced the acquisition function to explore more points within the bound. We evaluate ξ_{max} under different values and we select 1.5 as ξ_{max}. After exploring several points, we decay ξ with exponential function $f(x) = e^{-kx}$ in 7-10 iterations so that the acquisition function obtains higher exploitation. The exponential function e^{-kx} can dramatically decrease to a small value and enable the acquisition function to have enough time to exploit. We compel the function to exploit the knowledge about likely high reward regions in the function's domain. For the final step, the ξ value is around 5×10^{-3} which is relatively small and the function can convergence within 10 steps.

And see code

Answer to Question 2.6

1) see code

2) Without trained hyperparameters, every data point in the function is not connected smoothly. Instead, the hyperparameters (σ_f^2 and length_scale) trained through Gaussian process estimate the covariance between two different data points so that the function becomes smoother and the fit model is closer to the true distribution.