A Appendix-I: Preliminaries for Gaussian Process Regression

The Gaussian Process Regression involves two steps. Firstly, we need to introduce the prior by specifying a prior mean function and a prior covariance function. After that, we can use these functions to calculate a posterior mean function. The posterior mean function is exactly the estimator \( \hat{y}(x) \) that we want. We introduce these two steps one by one in the following.

Consider the first step. The prior of the Gaussian process based on \( T_f \) is specified by two components. The first component is the mean function taking the features of an instance as an output, denoted by \( m(\cdot) \), and the second component is the covariance function taking two features as an input, denoted by \( k(\cdot, \cdot) \). In the second component, for any two features \( x_i \) and \( x_j \) where \( i \in [1, n] \) and \( j \in [1, n] \), \( k(x_i, x_j) \) outputs a real value denoting the correlation between \( x_i \) and \( x_j \). Formally, the distribution is represented in the form of \( GP(m(x), k(\cdot, \cdot)) \).

Following previous studies [9], we set the mean function \( m(\cdot) \) to 0.5. We adopt the Radial Basis Function (RBF) [9] as a covariance function \( k(\cdot, \cdot) \) since it has a nice theoretical property to be used in our theoretical analysis.

We define an \( n \times n \) matrix denoted by \( K \) where the entry at the \( i \)-th row and at the \( j \)-th column in \( K \) is \( k(x_i, x_j) \) for \( i \in [1, n] \) and \( j \in [1, n] \). This matrix will be used in the second step of the model.

Consider the second step. We define the posterior mean function of the Gaussian process as follows. According to [9], since \( \hat{y}(x) \) follows \( GP(m(x), k(\cdot, \cdot)) \) and the RBF function is used as \( k(\cdot, \cdot) \), we can express \( \hat{y}(x) \) as follows.

\[
\hat{y}(x) = k(x)^T (K + \sigma^2 I)^{-1} f.
\]

where \( k(x) = [k(x, x_1), \ldots, k(x, x_n)] \) and \( I \) denotes the \( n \times n \) identity matrix. We say that \( k(x, x_i) \) is an instance-based kernel function where \( x \in X \) and \( i \in [1, n] \) since it involves an instance with its feature \( x_i \).

Note that \( \hat{y}(x) \) can be written as a weighted linear combination of instance-based kernel functions. Specifically, it can be written as \( k(x)^T a \) where \( a \) is an \( n \)-dimensional vector and

\[
a = (K + \sigma^2 I)^{-1} f.
\]

We define \( F \) to be the function class containing all possible functions \( \hat{y}(\cdot) \) in the above form of \( k(x)^T a \) such that the \( a \) vector associated with each function has its \( L_2 \)-norm value at most a given value \( A \) where \( A \) is a positive real number given by users. \( A \) can be regarded as a parameter describing the complexity of the function class. If \( A \) is larger, then the complexity of this class is higher.

B Appendix-II: Proof of Theorem 4.1

Proof. Before we give this proof, we first give the following lemma which will be used in the proof.

**Lemma B.1.** Given a confidence parameter \( \delta \in (0, 1) \), there exist three constants \( C_1, C_2 \) and \( C_3 \) which are independent of \( n \) such that with probability at least \( 1 - \delta \),

\[
E_{x,f}[(\hat{y}(x) - f)^2] \leq \Delta \;
\]

where \( \Delta = \frac{C_1 + C_2 \ln n + C_3 \ln \frac{1}{\delta}}{n} \).

**Proof.** Given \( x \in X \), the hypothesis \( h(x) \) used in this paper is \( \|h(x)\|_{\ell^2} \geq 1/2 \), where \( \hat{y}(\cdot) \) is the regression function for estimating the conditional probability. We write it as \( h(\cdot) \) if the context is clear. We define the hypothesis space \( \mathcal{H} \) to be \( \{h(\cdot) : \|h(\cdot)\|_{\ell^2} \geq 1/2 \} \) for each \( \hat{y}(\cdot) \in F \). Let \( d \) be the VC dimension of \( \mathcal{H} \).

Given a function \( \hat{y} \in F \), \( x \in X \) and \( f \in [0, 1] \), we define the square loss of \( \hat{y} \), denoted by \( g_{\hat{y}}(x, f) \), to be

\[
g_{\hat{y}}(x, f) = (\hat{y}(x) - f)^2
\]

Let \( \mathcal{G} = \{g_{\hat{y}}(\cdot, \cdot) : \hat{y} \in F\} \). For simplicity, we write \( g_{\hat{y}}(\cdot, \cdot) \) as \( g(\cdot, \cdot) \) if \( \hat{y} \) is clear in the context.

In order to prove this lemma, we used the following existing lemma (Lemma 20.8 in [14]).

**Lemma B.2.** ([14]) Suppose that we are given a set \( Z \) of elements and we observed \( n \) elements in \( Z \), namely \( z_1, z_2, \ldots, z_n \). Consider a class \( \mathcal{L} \) of real-valued functions defined on \( Z \), and suppose that for each \( l \in \mathcal{L} \) and each \( z \in Z \), \( |l(z)| \leq K_1 \) where \( K_1 \) is a real number greater than 0. Given \( \epsilon \in (0, 1) \), we denote \( M(\mathcal{L}, \epsilon) \) to be the covering number of the \( \epsilon \)-cover of class \( \mathcal{L} \) [14]. Let \( P(Z) \) be a probability distribution on \( Z \) for which \( E[l(z)] \geq 0 \) and \( E[l(z)^2] \leq K_2 \cdot E[l(z)] \) for each \( l \in \mathcal{L} \) where \( K_2 \) is another real number at least 1. Then, for \( \epsilon > 0 \), \( 0 < \alpha < \frac{1}{2} \) and \( n \geq \max \{4(K_1 + K_2)/(\alpha^2 \epsilon), K_2/(\alpha^2 \epsilon)\} \),

\[
Pr(\exists l \in \mathcal{L}, \frac{E[l(z)] - \frac{1}{2} \sum_{i=1}^{n} l(z_i)}{E[l(z)] + \epsilon} \geq \alpha) \\
\leq 2M(\mathcal{L}, \frac{\alpha \epsilon}{8}) \exp \left(-\frac{3\alpha^2\epsilon}{8K_1 + 324K_2}\right) + \\
4M(\mathcal{L}, \frac{\alpha \epsilon}{8\sqrt{\pi}}) \exp \left(-\frac{\alpha^2 \epsilon}{8}\right)
\]

Consider a function \( g \in \mathcal{G} \). Note that for any \( x \in X \) and \( f \in [0, 1] \), \( |g(x, f)| \leq 1 \) and \( E[g(x, f)^2] \leq E[g(x, f)] \). Let \( X_F = \{(x, f) : x \in X, f \in [0, 1]\} \).

We use Lemma B.2 by setting the parameters in this lemma as follows. We set \( Z \) to \( X_F \). Each observation \( z_i \) is set to \( (x_i, f_i) \) where \( i \in [1, n] \). Besides, we set \( \mathcal{L} = \mathcal{G} \), \( l = g, \alpha = \frac{1}{2}, K_1 = 1 \) and \( K_2 = 1 \). By Lemma B.2, we have

\[
Pr(\exists g \in \mathcal{G}, E[g(x, f)] - \frac{2}{n} \sum_{i=1}^{n} g(x_i, f_i) \geq \epsilon) \\
\leq 6M(\mathcal{G}, \frac{\epsilon}{10}) \exp \left(-\frac{3\epsilon n}{1328}\right)
\]
where $E$ note that $(g)$ From (B.3), we know that since from Equation (A.2), we have $1 + \frac{\ln}{\ln 6}$, $C_2 = \frac{1328d}{3}$ and $C_3 = \frac{1328}{3}$. Next, we want to find the upper bound of $\sum_{i=1}^{n} g(x_i, f_i)$. From (B.3), we know that

$$\sum_{i=1}^{n} g(x_i, f_i) = \sum_{i=1}^{n} (\hat{g}(x_i) - f_i)^2$$

Note that $\hat{g}(x_i) = k(x_i)^T a$ for $i \in [1, n]$. Besides, it is easy to verify that $K$ can be expressed as $(k(x_1), \ldots, k(x_i), \ldots, k(x_n))^T$. Let $\tilde{y} = \{\hat{g}(x_i)\}_{i=1}^{n}$. We can deduce that $\tilde{y} = Ka$. Thus, it is easy to show that

$$\sum_{i=1}^{n} (\hat{g}(x_i) - f_i)^2 = (Ka - f) \cdot (Ka - f)$$

From (B.6) and (B.7), we have

$$\sum_{i=1}^{n} g(x_i, f_i) = (Ka - f) \cdot (Ka - f)$$

From Equation (A.2), we have

$$a = (K + \sigma^2 I)^{-1} f$$

$$(K + \sigma^2 I) a = f$$

$$Ka - f = -\sigma^2 a$$

From (B.8) and (B.9), we derive that

$$\sum_{i=1}^{n} g(x_i, f_i) = \sigma^4 a \cdot a$$

Since $\|a\| = \sqrt{a \cdot a}$ and $\|a\| \leq A$, we have

$$\sum_{i=1}^{n} g(x_i, f_i) \leq \sigma^4 A^2$$

Therefore, by combining (B.10) and (B.6), we have

$$E[g(x, f)] \leq C_1 + C_2 \ln n + C_3 \ln \frac{1}{\delta}$$

where $C_1 = \frac{1328d}{3} \ln \frac{\epsilon}{\delta} + 2\sigma^4 A^2$, $C_2 = \frac{1328d}{3}$ and $C_3 = \frac{1328}{3}$. Since $g(x, f) = (\hat{g}(x) - f)^2$, we have $E[(\hat{g}(x) - f)^2] \leq C_1 + C_2 \ln n + C_3 \ln \frac{1}{\delta}$. Let $C_1 = C_1 + C_2 \ln n + C_3 \ln \frac{1}{\delta}$.

We have $E[(\hat{g}(x) - f)^2] \leq \Delta$.

We have just given Lemma B.1. We are ready to give the proof of Theorem 4.1.

In this proof, for convenience, $E_{x \sim P(x)}[\cdot]$ is represented by $E[\cdot]$, and $P_{x \sim P(x)}(\cdot)$ is represented by $Pr(\cdot)$. We know that

$$E(h) = Pr_{x,y}(y \neq h(x)) - Pr_{x,y}(y \neq \hat{h}(x))$$

$$= E_{x}[Pr_{y|x}(y \neq h(x))] - E_{x}[Pr_{y|x}(y \neq \hat{h}(x))].$$

(B.12) $E_{x}[Pr_{y|x}(y \neq h(x)) - Pr_{y|x}(y \neq \hat{h}(x))]$

Consider a certain feature $x$. We want to show that $Pr_{y|x}(y \neq h(x)) - Pr_{y|x}(y \neq \hat{h}(x))$ can be expressed as $|\eta(x) - 1| \cdot |h(x) - \hat{h}(x)|$. Note that $Pr_{y|x}(y \neq h(x))$ is equal to either $\eta(x)$ or $1 - \eta(x)$. Similarly, $Pr_{y|x}(y \neq \hat{h}(x))$ is equal to either $\eta(x)$ or $1 - \eta(x)$. Consider two cases. Case 1: $h(x) = \hat{h}(x)$. In this case, $|h(x) - \hat{h}(x)| = 0$. Since there is no error of hypothesis $h$ (compared with $h^*$), we derive that $Pr_{y|x}(y \neq h(x)) - Pr_{y|x}(y \neq \hat{h}(x)) = 0$. It is easy to see that $Pr_{y|x}(y \neq h(x)) - Pr_{y|x}(y \neq \hat{h}(x))$ can be expressed as $|\eta(x) - 1| \cdot |h(x) - \hat{h}(x)|$. Case 2: $h(x) \neq \hat{h}(x)$. In this case, since $h^*(\cdot)$ is optimal, we know that $Pr_{y|x}(y \neq \hat{h}(x)) = \min\{\eta(x), 1 - \eta(x)\}$ (because $h^*(\cdot)$ introduces the smallest error). Since $h(x) \neq \hat{h}(x)$, we derive that $Pr_{y|x}(y \neq \hat{h}(x)) = \max\{\eta(x), 1 - \eta(x)\}$. Thus, we have $Pr_{y|x}(y \neq h(x)) - Pr_{y|x}(y \neq \hat{h}(x)) = [2\eta(x) - 1] \cdot |h(x) - \hat{h}(x)|$. Note that $|h(x) - \hat{h}(x)| = 1$. Thus, $Pr_{y|x}(y \neq h(x)) - Pr_{y|x}(y \neq \hat{h}(x))$ can be expressed as $|2\eta(x) - 1| \cdot |h(x) - \hat{h}(x)|$. Therefore, from (B.12), we conclude that

$$E(h) = E[|2\eta(x) - 1| \cdot |h(x) - \hat{h}(x)|].$$

We know that when $h(x) \neq \hat{h}(x)$, we have $|\eta(x) - \frac{1}{2}| \leq |\eta(x) - \hat{\eta}(x)|$. This is because if $\eta(x) \leq \frac{1}{2}$, then we know that $\hat{\eta}(x) > \frac{1}{2}$ and thus we derive that $|\eta(x) - \frac{1}{2}| \leq |\eta(x) - \hat{\eta}(x)|$. Besides, if $\eta(x) > \frac{1}{2}$, then we have a similar conclusion.

Since $|\eta(x) - \frac{1}{2}| \leq |\eta(x) - \hat{\eta}(x)|$, we derive that $|2\eta(x) - 1| \leq 2|\eta(x) - \hat{\eta}(x)|$. Besides, from (B.13), we have

$$E(h) \leq 2|E[|\eta(x) - \hat{\eta}(x)| \cdot |h(x) - \hat{h}(x)|]|$$

(B.14) $2 \cdot E[|\eta(x) - \hat{\eta}(x)| \cdot I_{h(x) \neq \hat{h}(x)}]|$

According to Hölder Inequality, we have $E[(\eta(x) - \hat{\eta}(x)) \cdot I_{h(x) \neq \hat{h}(x)}] \leq \sqrt{\mathbb{E}[|\eta(x) - \hat{\eta}(x)|^2]} \cdot \sqrt{\mathbb{E}[I_{h(x) \neq \hat{h}(x)}]^2]}$. Since $I_{h(x) \neq \hat{h}(x)}^2 = 1_{h(x) \neq \hat{h}(x)}$, we have $E[|\eta(x) - \hat{\eta}(x)| \cdot I_{h(x) \neq \hat{h}(x)}] \leq \sqrt{\mathbb{E}[|\eta(x) - \hat{\eta}(x)|^2]} \cdot \sqrt{\mathbb{E}[I_{h(x) \neq \hat{h}(x)}]^2]}$. Since $E[I_{h(x) \neq \hat{h}(x)}] = Pr(h(x) \neq \hat{h}(x))$, we have $E[|\eta(x) - \hat{\eta}(x)| \cdot I_{h(x) \neq \hat{h}(x)}] \leq \sqrt{\mathbb{E}[|\eta(x) - \hat{\eta}(x)|^2]} \cdot \sqrt{Pr(h(x) \neq \hat{h}(x))}$. From (B.14), we derive the following.

$$E(h) \leq 2 \sqrt{\mathbb{E}[|\eta(x) - \hat{\eta}(x)|^2]} \cdot \sqrt{Pr(h(x) \neq \hat{h}(x))}$$
After we find the upper bound of the right-hand side of the above inequality, we can complete the proof. In the following, we will show that with probability at least 1 − \(\delta\),
\[
\mathbb{E}[|\eta(x) - \hat{\eta}(x)|^2] \leq \Delta
\]
and \(Pr(h(x) \neq h^*(x)) \leq c \cdot \Delta^{3/2}\).

With these results, we derive that with probability at least 1 − \(\delta\), \(E(h) \leq 2 \cdot \sqrt{\Delta} \cdot \sqrt{c \cdot \Delta^{3/2}} = 2 \cdot \sqrt{\Delta} \cdot \Delta^{3/4}\). Thus, after substituting Equation (B.11) into the above inequality, we have
\[
\mathbb{E}\left[(\hat{\eta}(x) - \eta(x))^2 + \sigma^2\right] = \mathbb{E}[|\eta(x) - \hat{\eta}(x)|^2 + (\eta(x) - f(x))^2]
\]
\[
= \mathbb{E}[|\eta(x) - \hat{\eta}(x)|^2 + (\eta(x) - f(x))^2]
\]
\[
- 2(\eta(x) - \hat{\eta}(x))(\eta(x) - \eta(x))
\]
\[
= \mathbb{E}[|\eta(x) - \hat{\eta}(x)|^2 + (\eta(x) - f(x))^2]
\]
\[
- 2(\eta(x) - \hat{\eta}(x))(\eta(x) - \eta(x))
\]
\[
= \mathbb{E}[|\eta(x) - \hat{\eta}(x) - (\eta(x) - f(x))|^2]
\]
\[
= \mathbb{E}[(\hat{\eta}(x) - \eta(x))^2]
\]
we know that
\[
\mathbb{E}[|\eta(x) - \hat{\eta}(x)|^2] \leq \mathbb{E}[(\hat{\eta}(x) - f(x))^2]
\]
as \(\sigma^2 \geq 0\).

From Lemma B.1, we derive that with probability at least 1 − \(\delta\),
\[
E[(\eta(x) - \hat{\eta}(x))^2] \leq \Delta
\]
(B.15) \(\mathbb{E}[(\eta(x) - \hat{\eta}(x))^2] \leq \Delta\)

Next, we will show that with probability at least 1 − \(\delta\), \(Pr(h(x) \neq h^*(x)) \leq c \Delta^{3/2}\).

Since \(\mathbb{E}[|\eta(x) - \hat{\eta}(x)|^2] < \mathbb{E}[(\eta(x) - \hat{\eta}(x))^2]\), from (B.15), we derive that with probability at least 1 − \(\delta\),
\[
\mathbb{E}[|\hat{\eta}(x) - \eta(x)|^2] \leq \sqrt{\Delta}
\]
(B.16) \(\mathbb{E}[|\hat{\eta}(x) - \eta(x)|^2] \leq \sqrt{\Delta}\)

Since
\[
Pr(h(x) \neq h^*(x))
\]
\[
= \mathbb{E}[I_{h(x) \neq h^*(x)}]
\]
\[
= \mathbb{E}[I_{h(x) \neq h^*(x), |\hat{\eta}(x) - \eta(x)|^2} \leq \sqrt{\Delta}]
\]
\[
+ \mathbb{E}[I_{h(x) \neq h^*(x), |\hat{\eta}(x) - \eta(x)|^2} > \sqrt{\Delta}]
\]
and according to Inequality (B.16), the second term above equals 0 (i.e., \(\mathbb{E}[|\hat{\eta}(x) - \eta(x)|^2] \leq \sqrt{\Delta}\)) with probability at least 1 − \(\delta\), we claim that with probability at least 1 − \(\delta\),
\[
Pr(h(x) \neq h^*(x)) \leq c \cdot \Delta^{3/2}
\]
(B.17) \(Pr(h(x) \neq h^*(x)) \leq c \cdot \Delta^{3/2}\)

As we discussed before, \(h(x) \neq h^*(x)\) implies that \(|\eta(x) - \frac{1}{2}| \leq |\hat{\eta}(x) - \eta(x)|\) for any \(x \in \mathcal{X}\). Thus, we have
\[
\mathbb{E}[|\hat{\eta}(x) - \eta(x)|^2] \leq \mathbb{E}[|\eta(x) - \eta(x)|]
\]
when \(h(x) \neq h^*(x)\). From (B.17), we derive that \(h(x) \neq h^*(x)\) and \(\mathbb{E}[|\hat{\eta}(x) - \eta(x)|^2] \leq \sqrt{\Delta} \) implies \(\mathbb{E}[|\eta(x) - \frac{1}{2}|] < \sqrt{\Delta}\) with probability at least 1 − \(\delta\). Therefore, with probability at least 1 − \(\delta\),
\[
Pr(h(x) \neq h^*(x), |\hat{\eta}(x) - \eta(x)|^2) \leq \sqrt{\Delta}
\]
\[
\leq Pr(\mathbb{E}[|\eta(x) - \frac{1}{2}|] < \sqrt{\Delta})
\]
\[
\leq c \cdot \Delta^{3/2}
\]
(B.18) \(Pr(h(x) \neq h^*(x)) \leq c \cdot \Delta^{3/2}\)

Finally, we will show that event \(E_1\) : “\(\mathbb{E}[|\eta(x) - \hat{\eta}(x)|^2] \leq \Delta\)” occurs if and only if event \(E_2\) : “\(Pr(h(x) \neq h^*(x)) \leq c \cdot \Delta^{3/2}\)” occurs. With this result, we conclude that with probability at least 1 − \(\delta\), these two events occur simultaneously. Thus, we complete the proof. Note that when we show “\(Pr(h(x) \neq h^*(x)) \leq c \cdot \Delta^{3/2}\)”, we make use of “\(\mathbb{E}[|\eta(x) - \hat{\eta}(x)|^2] \leq \Delta\)”. Thus, if \(E_1\) is true, then \(E_2\) is true. Otherwise, then \(E_2\) is not true.

Thus, we complete the proof. \(\square\)