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 $\hat{\eta}(\mathbf{x})$ 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 input, 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 $\mathbf{x}_i$ and $\mathbf{x}_j$ where $i \in [1,n]$ and $j \in [1,n]$, $k(\mathbf{x}_i,\mathbf{x}_j)$ outputs a real value denoting the *correlation* between $\mathbf{x}_i$ and $\mathbf{x}_j$. Formally, the distribution is represented in the form of $\mathcal{GP}(m(\cdot),k(\cdot,\cdot))$.

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 $k(\mathbf{x}_i,\mathbf{x}_j)$ 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 $\hat{\eta}(\mathbf{x})$ follows $\mathcal{GP}(m(\mathbf{x}),k(\cdot,\cdot))$ and the RBF function is used as $k(\cdot,\cdot)$, we can express $\hat{\eta}(\mathbf{x})$ as follows.

$$\hat{\eta}(\mathbf{x}) = \mathbf{k}(\mathbf{x})^T(K+\sigma^2\mathbf{I})^{-1}\mathbf{f}. \tag{A.1}$$

where $\mathbf{k}(\mathbf{x}) = \{k(\mathbf{x},\mathbf{x}_i)\}_{i=1}^n$, $\mathbf{f} = \{f_i\}_{i=1}^n$ and $\mathbf{I}$ denotes the $n\times n$ *identity matrix*. We say that $k(\mathbf{x},\mathbf{x}_i)$ is an *instance-based kernel function* where $\mathbf{x}\in\mathcal{X}$ and $i\in[1,n]$ since it involves an instance with its feature $\mathbf{x}_i$.

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

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

We define $\mathcal{F}$ to be the *function class* containing all possible functions $\hat{\eta}(\cdot)$ in the above form of $k(\mathbf{x})^T\mathbf{a}$ such that the $\mathbf{a}$ 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 $1-\delta$, $\mathbb{E}_{\mathbf{x},f}[(\hat{\eta}(\mathbf{x})-f)^2] \le \Delta$ where $\Delta = \frac{C_1+C_2\ln n+C_3\ln\frac{1}{\delta}}{n}$.*

Proof. Given $\mathbf{x}\in\mathcal{X}$, the hypothesis $h(\mathbf{x})$ used in this paper is $\mathbb{I}_{\hat{\eta}(\mathbf{x})\ge 1/2}$, where $\hat{\eta}(\cdot)\in\mathcal{F}$ 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 $\{h(\cdot) : h(\cdot) = \mathbb{I}_{\hat{\eta}(\cdot)\ge 1/2}$ for each $\hat{\eta}(\cdot)\in\mathcal{F}\}$. Let d be the VC dimension of $\mathcal{H}$.

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

$$g_{\hat{\eta}}(\mathbf{x},f) = (\hat{\eta}(\mathbf{x})-f)^2 \tag{B.3}$$

Let $\mathcal{G} = \{g_{\hat{\eta}}(\cdot,\cdot) : \hat{\eta}\in\mathcal{F}\}$. For simplicity, we write $g_{\hat{\eta}}(\cdot,\cdot)$ as $g(\cdot,\cdot)$ if $\hat{\eta}$ 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 $z_1,z_2,...,z_n$. Consider a class $\mathcal{L}$ of real-valued functions defined on set Z, and suppose that for each $l\in\mathcal{L}$ and each $z\in Z$, $|l(z)\le K_1|$ where K_1 is a real number greater than 0. Given $\epsilon\in(0,1)$, we denote $\mathcal{M}(\mathcal{L},\epsilon)$ to be the covering number of the ϵ-cover of class $\mathcal{L}$ [14]. Let $P(Z)$ be a probability distribution on Z for which* $\mathbb{E}[l(z)]\ge 0$ *and* $\mathbb{E}[l(z)^2]\le K_2\cdot\mathbb{E}[l(z)]$ *for each $l\in\mathcal{L}$ where K_2 is another real number at least 1. Then, for $\epsilon>0$, $0<\alpha\le\frac{1}{2}$ and $n\ge\max\{4(K_1+K_2)/(\alpha^2\epsilon), K_1^2/(\alpha^2\epsilon)\}$,*

$$\begin{aligned}
&Pr(\exists l\in\mathcal{L}, \tfrac{\mathbb{E}[l(z)]-\frac{1}{n}\sum_{i=1}^n l(z_i)}{\mathbb{E}[l(z)]+\epsilon}\ge\alpha) \\
\le\ &2\mathcal{M}(\mathcal{L},\tfrac{\alpha\epsilon}{8})\exp(-\tfrac{3\alpha^2\epsilon n}{8K_1+324K_2})+ \\
&4\mathcal{M}(\mathcal{L},\tfrac{\alpha\epsilon}{8K_1})\exp(-\tfrac{\alpha^2\epsilon n}{4K_1^2})
\end{aligned}$$

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

We use Lemma B.2 by setting the parameters in this lemma as follows. We set Z to XF. Each observation z_i is set to $(\mathbf{x}_i,f_i)$ 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

$$\begin{aligned}
&Pr(\exists g\in\mathcal{G}, \mathbb{E}[g(\mathbf{x},f)]-\tfrac{2}{n}\sum_{i=1}^n g(\mathbf{x}_i,f_i)\ge\epsilon) \\
\le\ &6\mathcal{M}(\mathcal{G},\tfrac{\epsilon}{16})\exp(-\tfrac{3\epsilon n}{1328})
\end{aligned} \tag{B.4}$$

We set $\delta = 6\mathcal{M}(\mathcal{G}, \frac{\epsilon}{16})\exp(-\frac{3\epsilon n}{1328})$. Note that $\mathcal{M}(\mathcal{G}, \frac{\epsilon}{16}) \le (\frac{en}{d})^d$ where e is the natural logarithmic base [14]. Thus, we derive that

$$\epsilon \le \frac{1328}{3n}(d \cdot \ln\frac{en}{d} + \ln\frac{6}{\delta}) \tag{B.5}$$

From (B.4) and (B.5), we derive that with probability at least $1-\delta$, there exists a function $g \in \mathcal{G}$ such that

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

where $C_1 = \frac{1328}{3}(d\ln\frac{e}{d} + \ln 6)$, $C_2 = \frac{1328d}{3}$ and $C_3 = \frac{1328}{3}$.

Next, we want to find the upper bound of $\sum_{i=1}^n g(\mathbf{x}_i, f_i)$.

From (B.3), we know that

$$\sum_{i=1}^n g(\mathbf{x}_i, f_i) = \sum_{i=1}^n (\hat{\eta}(\mathbf{x}_i) - f_i)^2 \tag{B.6}$$

Note that $\hat{\eta}(\mathbf{x}_i) = \mathbf{k}(\mathbf{x}_i)^T\mathbf{a}$ for $i \in [1, n]$. Besides, it is easy to verify that K can be expressed as $(\mathbf{k}(\mathbf{x}_1), ..., \mathbf{k}(\mathbf{x}_i), ..., \mathbf{k}(\mathbf{x}_n))^T$. Let $\overrightarrow{\eta} = \{\hat{\eta}(\mathbf{x}_i)\}_{i=1}^n$. We can deduce that $\overrightarrow{\eta} = K\mathbf{a}$. Thus, it is easy to show that

$$\sum_{i=1}^n (\hat{\eta}(\mathbf{x}_i) - f_i)^2 = (K\mathbf{a} - \mathbf{f}) \cdot (K\mathbf{a} - \mathbf{f}) \tag{B.7}$$

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

$$\sum_{i=1}^n g(\mathbf{x}_i, f_i) = (K\mathbf{a} - \mathbf{f}) \cdot (K\mathbf{a} - \mathbf{f}) \tag{B.8}$$

From Equation (A.2), we have

$$\begin{aligned} \mathbf{a} &= (K + \sigma^2\mathbf{I})^{-1}\mathbf{f} \\ (K + \sigma^2\mathbf{I})\mathbf{a} &= \mathbf{f} \\ K\mathbf{a} - \mathbf{f} &= -\sigma^2\mathbf{a} \\ (K\mathbf{a} - \mathbf{f}) \cdot (K\mathbf{a} - \mathbf{f}) &= \sigma^4\mathbf{a} \cdot \mathbf{a} \quad \text{(B.9)} \end{aligned}$$

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

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

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

$$\sum_{i=1}^n g(\mathbf{x}_i, f_i) \le \sigma^4 A^2 \tag{B.10}$$

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

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

where $C_1 = \frac{1328d}{3}\ln\frac{e}{d} + 2\sigma^4 A^2$, $C_2 = \frac{1328d}{3}$ and $C_3 = \frac{1328}{3}$.

Since $g(\mathbf{x}, f) = (\hat{\eta}(\mathbf{x}) - f)^2$, we have $\mathbb{E}[(\hat{\eta}(\mathbf{x}) - f)^2] \le \frac{C_1 + C_2 \cdot \ln n + C_3 \cdot \ln\frac{1}{\delta}}{n}$. Let

$$\Delta = \frac{C_1 + C_2 \cdot \ln n + C_3 \cdot \ln\frac{1}{\delta}}{n}. \tag{B.11}$$

We have $\mathbb{E}[(\hat{\eta}(\mathbf{x}) - f)^2] \le \Delta$.

$\square$

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

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

$$\begin{aligned} E(h) &= Pr_{\mathbf{x},y}(y \ne h(\mathbf{x})) - Pr_{\mathbf{x},y}(y \ne h^*(\mathbf{x})) \\ &= \mathbb{E}_{\mathbf{x}}[Pr_{y|\mathbf{x}}(y \ne h(\mathbf{x}))] - \mathbb{E}_{\mathbf{x}}[Pr_{y|\mathbf{x}}(y \ne h^*(\mathbf{x}))]. \\ &= \mathbb{E}_{\mathbf{x}}[Pr_{y|\mathbf{x}}(y \ne h(\mathbf{x})) - Pr_{y|\mathbf{x}}(y \ne h^*(\mathbf{x}))] \quad \text{(B.12)} \end{aligned}$$

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

$$E(h) = \mathbb{E}[|2\eta(\mathbf{x}) - 1| \cdot |h(\mathbf{x}) - h^*(\mathbf{x})|]. \tag{B.13}$$

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

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

$$\begin{aligned} E(h) &\le \mathbb{E}[2|\eta(\mathbf{x}) - \hat{\eta}(\mathbf{x})| \cdot |h(\mathbf{x}) - h^*(\mathbf{x})|] \\ &= 2 \cdot \mathbb{E}[|\eta(\mathbf{x}) - \hat{\eta}(\mathbf{x})| \cdot \mathbb{I}_{h(\mathbf{x}) \ne h^*(\mathbf{x})}] \quad \text{(B.14)} \end{aligned}$$

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

$$E(h) \le 2\sqrt{\mathbb{E}[(\eta(\mathbf{x}) - \hat{\eta}(\mathbf{x}))^2]}\sqrt{Pr(h(\mathbf{x}) \ne h^*(\mathbf{x}))}$$

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 $1-\delta$, $\mathbb{E}[(\eta(\mathbf{x})-\hat{\eta}(\mathbf{x}))^2] \leq \Delta$ and $Pr(h(\mathbf{x}) \neq h^*(\mathbf{x})) \leq c \cdot \triangle^{\frac{\gamma}{2}}$. With these results, we derive that with probability at least $1-\delta$, $E(h) \leq 2 \cdot \sqrt{\Delta} \cdot \sqrt{c \cdot \Delta^{\frac{\gamma}{2}}} = 2 \cdot \sqrt{c} \cdot \Delta^{\frac{2+\gamma}{4}}$. Thus, after substituting Equation (B.11) into the above inequality, we have $E(h) \leq 2 \cdot \sqrt{c} \cdot (\frac{C_1+C_2 \ln n + C_3 \ln \frac{1}{\delta}}{n})^{\frac{2+\gamma}{4}}$, where $C_1 = \frac{1328}{3}(d \ln \frac{e}{d} + \ln 6)$, $C_2 = \frac{1328 d}{3}$ and $C_3 = \frac{1328}{3}$.

The remaining part of this proof is to show the correctness of the upper bound of the right-hand side of the inequality. Firstly, we will show that with probability at least $1-\delta$, $\mathbb{E}[(\eta(\mathbf{x})-\hat{\eta}(\mathbf{x}))^2] \leq \Delta$.

Since

$$\begin{aligned}
& \mathbb{E}[(\eta(\mathbf{x})-\hat{\eta}(\mathbf{x}))^2] + \sigma^2 \\
= \;& \mathbb{E}[(\eta(\mathbf{x})-\hat{\eta}(\mathbf{x}))^2 + (\eta(\mathbf{x})-f)^2] \\
= \;& \mathbb{E}[(\eta(\mathbf{x})-\hat{\eta}(\mathbf{x}))^2 + (\eta(\mathbf{x})-f)^2 \\
& -2(\eta(\mathbf{x})-\hat{\eta}(\mathbf{x}))(\eta(\mathbf{x})-\eta(\mathbf{x}))] \\
= \;& \mathbb{E}[(\eta(\mathbf{x})-\hat{\eta}(\mathbf{x}))^2 + (\eta(\mathbf{x})-f)^2 \\
& -2(\eta(\mathbf{x})-\hat{\eta}(\mathbf{x}))(\eta(\mathbf{x})-f)] \\
= \;& \mathbb{E}[((\eta(\mathbf{x})-\hat{\eta}(\mathbf{x}))-(\eta(\mathbf{x})-f))^2] \\
= \;& \mathbb{E}[(\hat{\eta}(\mathbf{x})-f)^2]
\end{aligned}$$

we know that

$$\mathbb{E}[(\eta(\mathbf{x})-\hat{\eta}(\mathbf{x}))^2] \quad \leq \quad \mathbb{E}[(\hat{\eta}(\mathbf{x})-f)^2]$$

as $\sigma^2 \geq 0$.

From Lemma B.1, we derive that with probability at least $1-\delta$,

$$\text{(B.15)} \qquad \mathbb{E}[(\eta(\mathbf{x})-\hat{\eta}(\mathbf{x}))^2] \leq \Delta$$

Next, we will show that with probability at least $1-\delta$, $Pr(h(\mathbf{x}) \neq h^*(\mathbf{x})) \leq c\triangle^{\frac{\gamma}{2}}$.

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

$$\text{(B.16)} \qquad \mathbb{E}[|\hat{\eta}(\mathbf{x})-\eta(\mathbf{x})|] \leq \sqrt{\Delta}$$

Since

$$\begin{aligned}
& Pr(h(\mathbf{x}) \neq h^*(\mathbf{x})) \\
= \;& \mathbb{E}[\mathbb{I}_{h(\mathbf{x}) \neq h^*(\mathbf{x})}] \\
= \;& \mathbb{E}[\mathbb{I}_{h(\mathbf{x}) \neq h^*(\mathbf{x}), \mathbb{E}[|\hat{\eta}(\mathbf{x})-\eta(\mathbf{x})|] \leq \sqrt{\Delta}}] \\
& + \mathbb{E}[\mathbb{I}_{h(\mathbf{x}) \neq h^*(\mathbf{x}), \mathbb{E}[|\hat{\eta}(\mathbf{x})-\eta(\mathbf{x})|] > \sqrt{\Delta}}]
\end{aligned}$$

and according to Inequality (B.16), the second term above equals 0 (i.e., $\mathbb{E}[|\hat{\eta}(\mathbf{x})-\eta(\mathbf{x})|] \leq \sqrt{\Delta}$) with probability at least $1-\delta$, we claim that with probability at least $1-\delta$,

$$\begin{aligned}
& Pr(h(\mathbf{x}) \neq h^*(\mathbf{x})) \\
\text{(B.17)} = \;& Pr(h(\mathbf{x}) \neq h^*(\mathbf{x}), \mathbb{E}[|\hat{\eta}(\mathbf{x})-\eta(\mathbf{x})|] \leq \sqrt{\Delta})
\end{aligned}$$

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

$$\begin{aligned}
& Pr(h(\mathbf{x}) \neq h^*(\mathbf{x}), \mathbb{E}[|\hat{\eta}(\mathbf{x})-\eta(\mathbf{x})|] \leq \sqrt{\Delta}) \\
\leq \;& Pr(\mathbb{E}[|\eta(\mathbf{x})-\frac{1}{2}|] < \sqrt{\Delta}) \\
\leq \;& c \cdot \Delta^{\frac{\gamma}{2}} \qquad \text{(By Definition 1)}
\end{aligned}$$

Thus, with probability at least $1-\delta$,

$$\text{(B.18)} \qquad Pr(h(\mathbf{x}) \neq h^*(\mathbf{x})) \leq c \cdot \Delta^{\frac{\gamma}{2}}$$

Finally, we will show that event E_1 : "$\mathbb{E}[(\eta(\mathbf{x})-\hat{\eta}(\mathbf{x}))^2] \leq \Delta$" occurs if and only if event E_2 : "$Pr(h(\mathbf{x}) \neq h^*(\mathbf{x})) \leq c \cdot \Delta^{\frac{\gamma}{2}}$" 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 "$Pr(h(\mathbf{x}) \neq h^*(\mathbf{x})) \leq c \cdot \Delta^{\frac{\gamma}{2}}$", we make use of "$\mathbb{E}[(\eta(\mathbf{x})-\hat{\eta}(\mathbf{x}))^2] \leq \Delta$". Thus, if E_1 is true, then E_2 is true. Otherwise, then E_2 is not true.

Thus, we complete the proof. $\square$