A Proof of Theorem 3.1

In this appendix, we give the proof of Theorem 3.1. Before we give the proofs, we first give some notations used in the proofs in Section A.1. Then, we give the proof of Theorem 3.1 in Section A.2.

A.1 Notations For the ease of presentation, we define some new notations which will be used in the proofs.

In the training process, whenever we determine whether we need to obtain the fractional score of $\mathbf{x}_i$ with probability p_i, if we obtain its fractional score, Q_i is set to 1 and we can obtain its fractional score f_i. Otherwise, Q_i is set to 0 and f_i is an undefined value. Thus, each instance can be represented by $(\mathbf{x}_i, f_i, p_i, Q_i)$.

Next, we give error measurements, called the *excess square loss*, which are based on *both* an estimated function $\hat{\eta}$ and the "best" estimated function.

Given an estimator $\hat{\eta} \in \mathcal{F}$, a feature $\mathbf{x}$ and a fractional score f, we define the *excess square loss* of $\hat{\eta}$ to be

$$g_{\hat{\eta}}(\mathbf{x}, f) = \ell_{\hat{\eta}}(\mathbf{x}, f) - \ell_{\hat{\eta}^*}(\mathbf{x}, f) \tag{A.1}$$

Next, we define a set $\mathcal{G}$ which contains functions denoting the excess square loss. Specifically, we define $\mathcal{G}$ to be $\{g_{\hat{\eta}} : \hat{\eta} \in \mathcal{F}\}$. For simplicity, we write $g_{\hat{\eta}}$ as g if $\hat{\eta}$ is clear in the context.

Given a function $g_{\hat{\eta}} \in \mathcal{G}$, we define $\mathbb{E}[g_{\hat{\eta}}]$ and $\hat{\mathbb{E}}_t[g_{\hat{\eta}}]$ as follows.

$$\mathbb{E}[g_{\hat{\eta}}] = \mathbb{E}_{\mathbf{x},f}[g_{\hat{\eta}}(x, f)] \tag{A.2}$$

and

$$\hat{\mathbb{E}}_t[g_{\hat{\eta}}] = \frac{1}{t}\sum_{i=1}^{t} \frac{Q_i}{p_i}(\ell_{\hat{\eta}}(\mathbf{x}_i, f_i) - \ell_{\hat{\eta}^*}(\mathbf{x}_i, f_i)), \tag{A.3}$$

Note that according to Equation (A.1) and Equation (A.2), $\mathbb{E}[g_{\hat{\eta}}] = \mathbb{E}_{\mathbf{x},f}[\ell_{\hat{\eta}}(\mathbf{x}, f) - \ell_{\hat{\eta}^*}(\mathbf{x}, f)]$. $\mathbb{E}[g_{\hat{\eta}}]$ corresponds to the *expected excess square loss* of a function $g_{\hat{\eta}} \in \mathcal{G}$.

We also define *ϵ-cover* and *covering number* as follows. Given an error parameter $\epsilon \in (0, 1)$ and a set $\widetilde{\mathcal{G}}$ of functions, namely $g_1, ..., g_N$, where each function takes a feature in $\mathcal{X}$ as an input and a value in $[0, 1]$ as an output for $i \in [1, N]$, $\widetilde{\mathcal{G}}$ is said to be an *ϵ-cover* of $\mathcal{G}$ if for each $g \in \mathcal{G}$, there exists a function $g_i \in \widetilde{\mathcal{G}}$ such that $\mathbb{E}[(g_j(\mathbf{x}) - g(\mathbf{x}))^2] \leq \epsilon$. Given an error parameter $\epsilon \in (0, 1)$, we define the *ϵ-covering number* of $\mathcal{G}$, denoted by $\mathcal{M}(\epsilon, \mathcal{G})$, to be the minimum size of the ϵ-cover of $\mathcal{G}$ among all possible ϵ-covers of $\mathcal{G}$. Assume that $\mathcal{G}$ follows the *uniform Glivenko-Canelli* (UGC) property. Then, we know that $\mathcal{M}(\epsilon, \mathcal{G}) < \infty$ for any $\epsilon > 0$. We define $\mathcal{M}(\epsilon, \mathcal{F}) = \mathcal{M}(\epsilon, \mathcal{G})$. It is known that the complexity of the ϵ-cover of a function class is $O(\frac{1}{\epsilon})$ [15]. That is, $\mathcal{M}(\epsilon, \mathcal{G}) = O(\frac{1}{\epsilon})$.

A.2 Major Idea of the Proof of Theorem 3.1 In the previous section, we defined some notations and some concepts. In this section, we give the major idea of the proof of Theorem 3.1, the major result of this paper.

Major Idea of Theorem 3.1: The following shows the major idea of showing the proof of Theorem 3.1. Note that in Algorithm 1, for the t-th iteration where $t \in [1, n]$, we need to estimate the optimal value η^*. The estimated value is denoted by $\hat{\eta}_t$. After we have this estimated value, we create a hypothesis h_t and set it to $\mathbb{I}_{\hat{\eta}_t(\cdot)\geq 0.5}$. There are different error measurements to evaluate the *error* introduced by this estimated value.

- The first error measurement is the expected excessive square loss of $\hat{\eta}_t$, denoted by $\mathbb{E}[g_{\hat{\eta}}]$ which is equal to the difference between the square loss of $\hat{\eta}_t$ and the square loss of $\hat{\eta}^*$. Formally, it is equal to $\mathbb{E}_{\mathbf{x},f}[\ell_{\hat{\eta}_t}(\mathbf{x}, f) - \ell_{\eta^*}(\mathbf{x}, f)]$.
- The second error measurement is the expected difference between $\hat{\eta}_t$ and its fractional score f. Formally, it is equal to $\mathbb{E}_{\mathbf{x},f}[|\hat{\eta}_t(\mathbf{x}) - f|]$.

After we obtain the above two error measurements, we can derive $E(h_t)$.

In the following, we want to show the upper bound of the first error measurement first in the following Lemma A.1. Then, we show that the upper bound of the second error measurement in the following Lemma A.2. Finally, we show the upper bound of $E(h_f)$ in the following Lemma A.3.

LEMMA A.1. (FIRST ERROR MEASUREMENT) *Let $\hat{\eta}_t \in \mathcal{F}$, which is returned at the t-th round in Algorithm 1. For a confidence parameter $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\mathbb{E}[g_{\hat{\eta}_t}] \leq 128 \cdot \frac{\ln \mathcal{M}(\epsilon/32, \mathcal{G}) + \ln \frac{2}{\delta}}{p_{min}^2 t} + \frac{2\sigma^4 A^2}{p_{min} t}$$

LEMMA A.2. (SECOND ERROR MEASUREMENT) *Let $\hat{\eta}_t \in \mathcal{F}$, which is returned at the t-th round in Algorithm 1. For a confidence parameter $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\mathbb{E}_{\mathbf{x},f}[|\hat{\eta}_t(\mathbf{x}) - f|] \leq \sqrt{128 \cdot \frac{\ln \mathcal{M}(\epsilon/32, \mathcal{G}) + \ln \frac{2}{\delta}}{p_{min}^2 t} + \frac{2\sigma^4 A^2}{p_{min} t}}$$

LEMMA A.3. (ERROR BOUND FOR CLASSIFICATION) *Let $h_t = \mathbb{I}_{\hat{\eta}_t \geq 0.5}$ be the classifier returned at the t-th round. For a confidence parameter $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$E(h_t) \leq c \cdot (128 \cdot \frac{\ln \mathcal{M}(\epsilon/32, \mathcal{G}) + \ln \frac{2}{\delta}}{p_{min}^2 t} + \frac{2\sigma^4 A^2}{p_{min} t})^{\frac{2+\gamma}{4}} \tag{A.4}$$

In the following, we give the proof of Lemma A.1, the proof of Lemma A.2 and the proof of Lemma A.3 in Section B, Section B.2 and Section B.3, respectively.

B Proof of Lemma A.1

Lemma A.1 is the lemma showing the upper bound of the first error measurement, $\mathbb{E}_{\mathbf{x},f}[\ell_{\hat{\eta}_t}(\mathbf{x},f) - \ell_{\eta^*}(\mathbf{x},f)]$. Note that $g_{\hat{\eta}}(\mathbf{x},f) = \ell_{\hat{\eta}}(\mathbf{x},f) - \ell_{\hat{\eta}^*}(\mathbf{x},f)$. Lemma A.1 is a lemma studying the upper bound of $\mathbb{E}_{\mathbf{x},f}[g_{\hat{\eta}}(\mathbf{x},f)]$. In order to show this, we first compute the upper bounds of the following probabilities.

- $Pr(\frac{\mathbb{E}[g]-\hat{\mathbb{E}}_t[g]}{\sqrt{\mathbb{E}[g]}} \geq \psi)$ for any real number $\psi \in (0,1)$
- $Pr(\frac{\mathbb{E}[g]-\hat{\mathbb{E}}_t[g]}{\mathbb{E}[g]+\epsilon} \geq \beta)$ for any real number $\beta \in (0,1)$

After knowing the upper bounds of these two probabilities, we can show the correctness of Lemma A.1. The first upper bound can be found in Lemma B.1, while the second upper bound can be found in Lemma B.2.

LEMMA B.1. *Given a function $g \in \mathcal{G}$, for any real number $\psi \in (0,1)$,*

$$Pr(\frac{\mathbb{E}[g]-\hat{\mathbb{E}}_t[g]}{\sqrt{\mathbb{E}[g]}} \geq \psi) \leq \exp(-\frac{p_{min}^2\psi^2 t}{8})$$

LEMMA B.2. *Given a function $g \in \mathcal{G}$ and an error parameter $\epsilon \in (0,1)$, for any real number $\beta \in (0,1)$,*

$$Pr(\frac{\mathbb{E}[g]-\hat{\mathbb{E}}_t[g]}{\mathbb{E}[g]+\epsilon} \geq \beta) \leq \exp(-\frac{p_{min}^2\beta^2\epsilon t}{8(1-\beta)})$$

Next, we show the proof of Lemma B.1 in Section B.1 and the proof of Lemma B.2 in Section B.1.1. Finally, we give the proof of Lemma A.1 in Section B.1.2.

B.1 Proof of Lemma B.1

Now, we show the proof of Lemma B.1.

Proof. (I) **Construct a martingale sequence.** In the following, we construct a martingale sequence of $t+1$ random variables, namely $Z_0, Z_1, ..., Z_t$.

Given a function $g \in \mathcal{G}$, we define $t+1$ random variables variables $Z_0, Z_1, Z_2, ..., Z_t$. Let $Z_0 = 0$. For any $i \in [1,t]$, we define

$$Z_i = i \cdot p_{min}(\mathbb{E}[g] - \hat{\mathbb{E}}_i[g]) \tag{B.5}$$

According to the definition of $\mathbb{E}[g]$ and $\hat{\mathbb{E}}_i[g]$, Z_i can also be written as follows.

$$\begin{aligned} Z_i &= p_{min}\textstyle\sum_{j=1}^{i}(\mathbb{E}[\ell_{\hat{\eta}} - \ell_{\hat{\eta}^*}] - \\ &\quad \tfrac{Q_j}{p_j}((\hat{\eta}(\mathbf{x}_j) - f_j)^2 - (\hat{\eta}^*(\mathbf{x}_j) - f_j)^2)) \end{aligned}$$

For any $i \in [1,t]$,

$$\begin{aligned} &\mathbb{E}[Z_i|Z_{i-1},...,Z_0] \\ =\ &\mathbb{E}[Z_{i-1} + p_{min}(\mathbb{E}[\ell_{\hat{\eta}} - \ell_{\hat{\eta}^*}] - \\ &\tfrac{Q_i}{p_i}((\hat{\eta}(\mathbf{x}_i) - f_i)^2 - (\hat{\eta}^*(\mathbf{x}_i) - f_i)^2))|Z_{i-1},...,Z_0] \\ =\ &Z_{i-1} + p_{min}\cdot\mathbb{E}_{\mathbf{x}_i,f_i}[\mathbb{E}[\ell_{\hat{\eta}} - \ell_{\hat{\eta}^*}] - \\ &((\hat{\eta}(\mathbf{x}_i) - f_i)^2 - (\hat{\eta}^*(\mathbf{x}_i) - f_i)^2)]|Z_{i-1},...,Z_0] \\ =\ &Z_{i-1} \end{aligned}$$

Therefore, the sequence $\{Z_0, Z_1, Z_2, ...Z_t\}$ is martingale according to its definition.

(II) **Applying a Hoeffding-type bound for the martingale**

In order to apply a Hoeffding-type bound for the martingale, we should calculate two terms beforehand. They are $\mathbb{E}[Z_t]$ and the upper bound on $\text{Var}(Z_i|Z_{i-1},...,Z_0)$.

First, we have

$$\mathbb{E}[Z_t] = t \cdot p_{min}\mathbb{E}[\mathbb{E}[g] - \hat{\mathbb{E}}_t[g]] = 0. \tag{B.6}$$

Second,

$$\begin{aligned} &\text{Var}(Z_i|Z_{i-1},...,Z_0) \\ =\ &\mathbb{E}_{\mathbf{x}_i,Q_i,f_i}[(Z_i - \mathbb{E}[Z_i|Z_{i-1},...,Z_0])^2|Z_{i-1},...,Z_0] \\ =\ &\mathbb{E}_{\mathbf{x}_i,Q_i,f_i}[(Z_i - Z_{i-1})^2|Z_{i-1},...,Z_0] \\ &-(\hat{\eta}^*(\mathbf{x}_i) - f_i)^2))^2|\{\mathbf{x}_j, f_j, Q_j\}_{j=1}^{i-1}] \\ =\ &p_{min}\cdot((\mathbb{E}[g])^2 - 2(\mathbb{E}[g])^2 + \tfrac{1}{p_i}\mathbb{E}[g^2]) \\ =\ &p_{min}\cdot(\tfrac{1}{p_i}\mathbb{E}[g^2] - \mathbb{E}[g]) \\ \leq\ &\mathbb{E}[g^2] \end{aligned} \tag{B.7}$$

Since the function class $\mathcal{F}$ is a class of functions represented as the weighted linear combination of instance-based kernel functions, we know that $\mathcal{F}$ is *closure convex* [1] and $|f_i - \hat{\eta}_t(\mathbf{x}_i)| \leq 1$ for any $(\mathbf{x}_i, f_i)$. Therefore, according to Lemma 20.9 in [2], we know that

$$\mathbb{E}[g^2] \leq 4\mathbb{E}[g]. \tag{B.8}$$

Combining Inequality (B.7) and Inequality (B.8), we have

$$\text{Var}(Z_i|Z_{i-1},...,Z_0) \leq 4\mathbb{E}[g] \tag{B.9}$$

According to Azuma's inequality, for any $\lambda > 0$,

$$Pr(Z_t - \mathbb{E}[Z_t] > \lambda) \leq \exp(-\frac{\lambda^2}{2t\cdot\text{Var}[Z_t|Z_{t-1},...,Z_0]})$$

By substituting Equation (B.5) and Equation (B.6) into the left hand side of the above inequality, we have

$$\begin{aligned} &Pr(t\cdot p_{min}(\mathbb{E}[g] - \hat{\mathbb{E}}_t[g]) - 0 > \lambda) \\ \leq\ &\exp(-\frac{\lambda^2}{2t\cdot\text{Var}[Z_t|Z_{t-1},...,Z_0]}) \end{aligned} \tag{B.10}$$

Furthermore, after substituting Inequality (B.9) into the right hand side of Inequality (B.10), we obtain

$$Pr(t\cdot p_{min}(\mathbb{E}[g] - \hat{\mathbb{E}}_t[g]) > \lambda) \leq \exp(-\frac{\lambda^2}{8t\cdot\mathbb{E}[g]})$$

Let $\psi = \lambda/(t\cdot p_{min}\sqrt{\mathbb{E}[g]})$. Then, we have

$$Pr(\frac{\mathbb{E}[g]-\hat{\mathbb{E}}_t[g]}{\sqrt{\mathbb{E}[g]}} > \psi) \leq \exp(-\frac{p_{min}^2\psi^2 t}{8})$$

So the proof is done.

[1] *closure convex* is defined in Definition 20.1 in [2].

B.1.1 Proof of Lemma B.2 We show the proof of Lemma B.2 now.

Proof. According to Lemma B.1, given $g \in \mathcal{G}$, suppose that $\mathbb{E}[g] - \hat{\mathbb{E}}_t[g] \geq \psi\sqrt{\mathbb{E}[g]}$. Then, for any $\psi' > 0$ we have two cases:

(I) If $\mathbb{E}[g] \geq (1 + 1/\psi')^2\psi^2$, then $\mathbb{E}[g] \geq \hat{\mathbb{E}}_t[g] + (1 + 1/\psi')\psi^2$.

(II) If $\mathbb{E}[g] < (1 + 1/\psi')^2\psi^2$, then $\mathbb{E}[g] \geq \hat{\mathbb{E}}_t[g] + \frac{\psi'}{1+\psi'}\mathbb{E}[g]$, and so $\mathbb{E}[g] \geq (1 + \psi')\hat{\mathbb{E}}_t[g]$.

In either case, $\mathbb{E}[g] \geq (1 + \psi')\hat{\mathbb{E}}_t[g] + (1 + 1/\psi')\psi^2$. Hence, given a fixed function $g \in \mathcal{G}$,

$$\begin{aligned} & Pr(\mathbb{E}[g] \geq (1 + \psi')\hat{\mathbb{E}}_t[g] + (1 + 1/\psi')\psi^2) \\ \leq \quad & \exp\left(-\tfrac{p_{min}^2\psi^2 t}{8}\right) \end{aligned}$$

Choosing $\psi' = \beta/(1 - \beta)$ and $\psi^2 = \epsilon\beta^2/(1 - \beta)$, we have

$$Pr(\mathbb{E}[g] \geq \tfrac{1}{1-\beta}\hat{\mathbb{E}}_t[g] + \epsilon\tfrac{\beta}{1-\beta}) \leq \exp\left(-\tfrac{p_{min}^2\epsilon\beta^2 t}{8(1-\beta)}\right)$$

which implies the result.

B.1.2 Proof of Lemma A.1 After we know the correctness of Lemma B.1 and Lemma B.2, we are ready to show the correctness of Lemma A.1.

Proof. We divide our proof of Lemma A.1 into two steps. Firstly, we resort to the Symmetrization technique to extend the bound on the right hand side of Lemma B.2 from a fixed function to a set of functions by using the idea of covering numbers $\mathcal{M}(\epsilon, \mathcal{G})$.

(I) **Symmetrization**

Denote by $\hat{\mathbb{E}}'_t[g]$ the empirical measure in terms of another t instance-label pairs. That is,

$$\hat{\mathbb{E}}'_t[g] = \tfrac{1}{t}\sum_{i=1}^{t}\tfrac{Q_{t+i}}{p_{t+i}}(\ell_{\hat{\eta}}(\mathbf{x}_{t+i}, f_{t+i}) - \ell_{\hat{\eta}^*}(\mathbf{x}_{t+i}, f_{t+i}))$$

Conventionally, we called these t instance-label pairs $\{\mathbf{x}_{t+i}, f_{t+i}, Q_{t+i}, p_{t+i}\}_{i=1}^t$ as "ghost samples", which indicates that they are mentioned for the sake of theoretical analysis but are not required in the real learning process.

Let $\beta = 0.5$ in Lemma B.2. Then, given a fixed function $g \in \mathcal{G}$,

$$Pr(\mathbb{E}[g] - 2\hat{\mathbb{E}}_t[g] \geq \epsilon) \leq \exp\left(-\tfrac{p_{min}^2\epsilon t}{16}\right) \tag{B.11}$$

For simplicity, let

$$p = Pr(\exists g \in \mathcal{G} \text{ s.t. } \mathbb{E}[g] - 2\hat{\mathbb{E}}_t[g] \geq \epsilon). \tag{B.12}$$

It is easy to verify that

$$p = Pr(\sup_{g\in\mathcal{G}}\ \mathbb{E}[g] - 2\hat{\mathbb{E}}_t[g] \geq \epsilon)$$

Let g_s be the function achieving the supremum (note that it depends on $\{(\mathbf{x}_i, f_i, Q_i, p_i)\}_{i=1}^t$). We have

$$\begin{aligned} & \mathbb{I}_{\mathbb{E}[g_s]-2\hat{\mathbb{E}}_t[g_s]\geq\epsilon}\mathbb{I}_{\mathbb{E}[g_s]-2\hat{\mathbb{E}}'_t[g_s]\leq\epsilon/2} \\ \leq \quad & \mathbb{I}_{\hat{\mathbb{E}}'_t[g_s]-\hat{\mathbb{E}}_t[g_s]\geq\epsilon/4} \end{aligned}$$

Taking expectations with respect to the "ghost samples", we have

$$\begin{aligned} \mathbb{I}_{\mathbb{E}[g_s]-2\hat{\mathbb{E}}_t[g_s]\geq\epsilon} \quad & Pr'(\mathbb{E}[g_s] - 2\hat{\mathbb{E}}'_t[g_s] \leq \epsilon/2) \\ \leq \quad & Pr'(\hat{\mathbb{E}}'_t[g_s] - \hat{\mathbb{E}}_t[g_s] \geq \epsilon/4) \end{aligned}$$

According to Inequality (B.11), we have

$$Pr'(\mathbb{E}[g_s] - 2\hat{\mathbb{E}}'_t[g_s] \leq \epsilon/2) \leq \exp\left(-\tfrac{p_{min}^2\epsilon t}{32}\right) \tag{B.13}$$

Hence, by substituting Inequality (B.13) into Inequality (B.13), we have

$$\begin{aligned} & \mathbb{I}_{\mathbb{E}[g_s]-2\hat{\mathbb{E}}_t[g_s]\geq\epsilon}(1 - \exp\left(-\tfrac{p_{min}^2\epsilon t}{32}\right)) \\ \leq \quad & Pr'(\hat{\mathbb{E}}'_t[g_s] - \hat{\mathbb{E}}_t[g_s] \geq \epsilon/4) \end{aligned}$$

Then, taking the expectation with respect to $\{\mathbf{x}_i, f_i, Q_i, p_i\}_{i=1}^t$,

$$\begin{aligned} & Pr(\mathbb{E}[g_s] - 2\hat{\mathbb{E}}_t[g_s] \geq \epsilon)(1 - \exp\left(-\tfrac{p_{min}^2\epsilon t}{32}\right)) \\ \leq \quad & Pr(\hat{\mathbb{E}}'_t[g_s] - \hat{\mathbb{E}}_t[g_s] \geq \epsilon/4) \end{aligned}$$

When $t \geq \frac{32\ln 2}{p_{min}^2\epsilon}$, we have

$$p \leq 2Pr(\sup_{g\in\mathcal{G}}\hat{\mathbb{E}}'_t[g] - \hat{\mathbb{E}}_t[g] \geq \epsilon/4) \tag{B.14}$$

Define a $\epsilon/32$-cover set $\mathcal{G}'$ of $\mathcal{G}$, denoted by $\{g_j : j = 1, ..., \mathcal{M}(\epsilon/32, \mathcal{G})\}$.

Then, $\mathcal{G}$ can be denoted as the union of $\mathcal{M}(\epsilon/32, \mathcal{G})$ subsets. That is, $\mathcal{G} = \mathcal{G}_1 \cup ... \cup \mathcal{G}_{\mathcal{M}(\epsilon/32,\mathcal{G})}$. For each $j \in [1, \mathcal{M}(\epsilon/32, \mathcal{G})]$, the subset G_j is centered at g_j with radius $\epsilon/32$.

Considering an union bound, we have

$$\begin{aligned} & Pr(\sup_{g\in\mathcal{G}}\hat{\mathbb{E}}'_t[g] - \hat{\mathbb{E}}_t[g] \geq \epsilon/4) \\ \leq \quad & \sum_{j=1}^{\mathcal{M}(\epsilon/32,\mathcal{G})} Pr(\sup_{g\in\mathcal{G}_j}\hat{\mathbb{E}}'_t[g] - \hat{\mathbb{E}}_t[g] \geq \epsilon/4) \end{aligned} \tag{B.15}$$

According to the definition of covering number, for any $g \in \mathcal{G}_j$,

$$\tfrac{1}{2t}\sum_{i=1}^{2t}\tfrac{Q_i}{p_i}|g_j(\mathbf{x}_i, f_i) - g(\mathbf{x}_i, f_i)| \leq \epsilon/32 \tag{B.16}$$

According to Inequality (B.16), we have

$$\hat{\mathbb{E}}_t[g] - \hat{\mathbb{E}}_t[g_j] < \epsilon/16. \tag{B.17}$$

$$\hat{\mathbb{E}}'_t[g_j] - \hat{\mathbb{E}}'_t[g] < \epsilon/16. \tag{B.18}$$

Combining (B.17) and (B.18), for any $g \in \mathcal{G}_j$,

$$\hat{\mathbb{E}}_t[g] - \hat{\mathbb{E}}'_t[g] < \hat{\mathbb{E}}_t[g_j] - \hat{\mathbb{E}}_t[g_j] + \epsilon/8$$

It follows

$$\begin{aligned} & Pr(\sup_{g\in\mathcal{G}_j} \hat{\mathbb{E}}'_t[g] - \hat{\mathbb{E}}_t[g] \geq \epsilon/4) \\ \leq\ & Pr(\hat{\mathbb{E}}'_t[g_j] - \hat{\mathbb{E}}_t[g_j] \geq \epsilon/8) \end{aligned} \tag{B.19}$$

By substituting Inequality (B.19) into the right hand side of Inequality (B.15), we have

$$\begin{aligned} & Pr(\sup_{g\in\mathcal{G}} \hat{\mathbb{E}}'_t[g] - \hat{\mathbb{E}}_t[g] \geq \epsilon/4) \\ \leq\ & \sum_{j=1}^{\mathcal{M}(\epsilon/32,\mathcal{G})} Pr(\hat{\mathbb{E}}'_t[g_j] - \hat{\mathbb{E}}_t[g_j] \geq \epsilon/8) \end{aligned} \tag{B.20}$$

By substituting Inequality (B.11) into the right hand side of Inequality (B.20), we have

$$\begin{aligned} & Pr(\sup_{g\in\mathcal{G}} \hat{\mathbb{E}}'_t[g] - \hat{\mathbb{E}}_t[g] \geq \epsilon/4) \\ \leq\ & \mathcal{M}(\epsilon/32,\mathcal{G}) \exp\left(-\tfrac{p_{min}^2 \epsilon t}{128}\right) \end{aligned} \tag{B.21}$$

Then, by substituting Inequality (B.21) into the right hand side of Inequality (B.14) and express p in terms of its definition shown in (B.12), we have

$$\begin{aligned} & Pr(\exists g \in \mathcal{G} \text{ s.t. } \mathbb{E}[g] - 2\hat{\mathbb{E}}_t[g] \geq \epsilon) \\ \leq\ & 2\mathcal{M}(\epsilon/32,\mathcal{G}) \exp\left(-\tfrac{p_{min}^2 \epsilon t}{128}\right), \end{aligned}$$

which can also be written as: for any $g \in \mathcal{G}$, with probability at least $1-\delta$,

$$\mathbb{E}[g] \leq 2\hat{\mathbb{E}}_t[g] + 128 \cdot \frac{\ln \mathcal{M}(\epsilon/32,\mathcal{G}) + \ln \frac{2}{\delta}}{p_{min}^2 t} \tag{B.22}$$

(II) **Upper bounding the regularizer**

In order to obtain our result in Lemma A.1, we need to guarantee that the term $2\hat{\mathbb{E}}_t[g_{\hat{\eta}_t}]$ also has dependence on t of t^{-1}.

Therefore, we proved that $2\hat{\mathbb{E}}_t[g_{\hat{\eta}_t}] = O(t^{-1})$ in the following. Since we know that $\hat{\eta}_t(\cdot)$ is the minimizer of the $J'[\hat{\eta}]$ at the t-th round,

$$J'[\hat{\eta}_t] = \tfrac{1}{2}\sum_{i=1}^t \tfrac{Q_i}{p_i}(\hat{\eta}_t(\mathbf{x}_i) - f_i)^2 + \tfrac{\sigma^2}{2}\|\hat{\eta}_t\|_{\mathcal{H}}^2$$

which can be written as

$$J'[\alpha'] = |\mathbf{f}' - K\alpha'|^2 + \sigma^2 \cdot \alpha'^T K' \alpha'. \tag{B.23}$$

where $\alpha' = \{\frac{Q_1}{\sqrt{p_1}}\alpha_1, ..., \frac{Q_t}{\sqrt{p_t}}\alpha_t\}$, $\mathbf{f}' = \{\frac{Q_1}{\sqrt{p_1}}f_1, ..., \frac{Q_t}{\sqrt{p_t}}f_t\}$ and K' is a $t \times t$ matrix, in which the i-th row and the j-th column equals $Q_i Q_j \sqrt{p_i p_j} k(\mathbf{x}_i, \mathbf{x}_j)$. The minimizer $\hat{\eta}_t(\cdot)$ is achieved by differentiating $J'[\alpha']$ w.r.t. α' and set the derivative be 0. Denote $\hat{\eta}_t(\cdot)$ by $\sum_{i=1}^t \hat{\alpha}_i k(\mathbf{x}_i, \cdot)$. Then we have

$$\sigma^2 K'\alpha' + K(K\alpha' - \mathbf{f}') = 0, \tag{B.24}$$

which can be simplified as

$$|K\alpha' - \mathbf{f}'| = |\frac{\sigma^2 K'\alpha'}{K}| \tag{B.25}$$

Since every entry in K is greater than the corresponding entry in K', the right hand side of the above equation is upper bounded by $|\sigma^2 \mathbf{I}\alpha'|$. Therefore, we have

$$\begin{aligned} & \tfrac{1}{t}\sum_{i=1}^t \tfrac{Q_i}{p_i}(\hat{\eta}_t(\mathbf{x}_i) - f_i)^2 \\ =\ & \tfrac{1}{t}(\mathbf{f}' - K\alpha')^2 \leq \tfrac{\sigma^4}{t}(\mathbf{I}\alpha')^2 \\ =\ & \tfrac{\sigma^4}{t}\sum_{i=1}^t \tfrac{Q_i}{p_i}\hat{\alpha}_i^2 \leq \tfrac{\sigma^4}{p_{min} t}\sum_{i=1}^t \hat{\alpha}_i^2 \\ \leq\ & \tfrac{\sigma^4 A^2}{p_{min} t} \end{aligned}$$

where A is the upper bound on $\|\hat{\alpha}\|$ for any $\hat{\eta} \in \mathcal{F}$.

So we know that

$$\begin{aligned} \hat{\mathbb{E}}_t[g_{\hat{\eta}_t}] &\leq \tfrac{1}{t}\sum_{i=1}^t \tfrac{Q_i}{p_i}(\hat{\eta}_t(\mathbf{x}_i) - f_i)^2 \\ &\leq \tfrac{\sigma^4 A^2}{p_{min} t} \end{aligned} \tag{B.26}$$

By substituting Inequality (B.26) into Inequality (B.22) with respect to $g_{\hat{\eta}_t}$, we have the result of Lemma A.1.

B.2 Proof of Lemma A.2 It is easy to show the correctness of Lemma A.2 as follows.

Proof. Firstly, we have

$$(\mathbb{E}_{\mathbf{x}}[|\hat{\eta}_t(\mathbf{x}) - \eta(\mathbf{x})|])^2 \leq \mathbb{E}_{\mathbf{x}}[(\hat{\eta}_t(\mathbf{x}) - \eta(\mathbf{x}))^2] \tag{B.27}$$

Since

$$\begin{aligned} & (\mathbb{E}_{\mathbf{x},f}[|\hat{\eta}_t(\mathbf{x}) - f|])^2 \\ =\ & (\mathbb{E}_{\mathbf{x}}[|\hat{\eta}_t(\mathbf{x}) - \eta(\mathbf{x})|])^2 \\ =\ & \mathbb{E}_{\mathbf{x}}[(\hat{\eta}_t(\mathbf{x}) - \eta(\mathbf{x}))^2 - (\hat{\eta}^*(\mathbf{x}) - \eta(\mathbf{x}))^2] \\ =\ & \mathbb{E}[g_{\hat{\eta}_t}], \end{aligned} \tag{B.28}$$

we have

$$\mathbb{E}_{\mathbf{x},f}[|\hat{\eta}_t(\mathbf{x}) - f|] \leq \sqrt{\mathbb{E}[g_{\hat{\eta}_t}]} \tag{B.29}$$

Then, we can simply obtain our result from the result of Lemma A.1. So we are done.

B.3 Proof of Lemma A.3 Next, we show the proof of Lemma A.3 as follows.

Proof. We start our proof from the definition of $E(h_t)$. 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 $Pr(\cdot)$. That is,

$$\begin{aligned} & E(h_t) \\ =\ & Pr_{(\mathbf{x},y)\sim P}(y \neq h_t(\mathbf{x})) - Pr_{(\mathbf{x},y)\sim P}(y \neq h^*(\mathbf{x})) \\ =\ & \mathbb{E}_{\mathbf{x}\sim P(X)}[Pr_{y\sim P(Y|X)}(y \neq h_t(\mathbf{x})) \\ & -Pr_{y\sim P(Y|X)}(y \neq h^*(\mathbf{x}))]. \end{aligned}$$

For a certain instance I with feature space $\mathbf{x}$, if $h_t(\mathbf{x}) = h^*(\mathbf{x})$ for pattern $\mathbf{x}$, then $\mathbb{I}_{y\neq h_t(\mathbf{x})} - \mathbb{I}_{y\neq h^*(\mathbf{x})} = 0$; otherwise, since $Pr(y \neq h^*(\mathbf{x})) = \min\{\eta(\mathbf{x}), 1-\eta(\mathbf{x})\}$, $Pr(y \neq h_t(\mathbf{x})) = \max\{\eta(\mathbf{x}), 1-\eta(\mathbf{x})\}$. Therefore, we have

$$E(h_t) \quad = \quad \mathbb{E}[|2\eta(\mathbf{x}) - 1||h_t(\mathbf{x}) - h^*(\mathbf{x})|].$$

Since $|\eta(\mathbf{x}) - \frac{1}{2}| \leq |\eta(\mathbf{x}) - \hat{\eta}(\mathbf{x})|$ when $h_t(\mathbf{x}) \neq h^*(\mathbf{x})$, we can upper bound $E(h_t)$ as follows,

$$\begin{aligned} E(h_t) &\leq \mathbb{E}[2|\eta(\mathbf{x}) - \hat{\eta}_t(\mathbf{x})||h_t(\mathbf{x}) - h^*(\mathbf{x})|] \\ &= 2\mathbb{E}[|\eta(\mathbf{x}) - \hat{\eta}_t(\mathbf{x})|\mathbb{I}_{h_t(\mathbf{x})\neq h^*(\mathbf{x})}] \end{aligned} \tag{B.30}$$

According to Cauchy-Schwarz inequality, we have

$$\begin{aligned} &\mathbb{E}[|\eta(\mathbf{x}) - \hat{\eta}_t(\mathbf{x})|\mathbb{I}_{h_t(\mathbf{x})\neq h^*(\mathbf{x})}] \\ \leq\ &\sqrt{\mathbb{E}[(\eta(\mathbf{x}) - \hat{\eta}_t(\mathbf{x}))^2]}\sqrt{Pr(h_t(\mathbf{x}) \neq h^*(\mathbf{x}))} \end{aligned} \tag{B.31}$$

Notice that the right hand side of the above inequality is the product of $\sqrt{\mathbb{E}[(\eta(\mathbf{x}) - \hat{\eta}_t(\mathbf{x}))^2]}$ and $\sqrt{Pr(h_t(\mathbf{x}) \neq h^*(\mathbf{x}))}$. The first term can be upper bounded according to Lemma A.1. In the following, we consider the second term.

Since h_t is the classifier returned at the t-th round, Lemma A.1's result is achieved. That is, with probability at least $1-\delta$, $\mathbb{E}_{\mathbf{x}}[|\hat{\eta}_t(\mathbf{x}) - \eta(\mathbf{x})|] \leq \Delta_t$, where we use Δ_t to represent the right hand side of the inequality in Lemma A.2 with respect to h_t. That is,

$$\Delta_t = \sqrt{128 \cdot \frac{\ln \mathcal{M}(\epsilon/32, \mathcal{G}) + \ln\frac{2}{\delta}}{p_{min}^2 t} + \frac{2\sigma^4 A^2}{p_{min} t}} \tag{B.32}$$

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

$$\begin{aligned} &Pr(h_t(\mathbf{x}) \neq h^*(\mathbf{x})) \\ =\ &Pr(h_t(\mathbf{x}) \neq h^*(\mathbf{x})|\mathbb{E}_{\mathbf{x}}[|\hat{\eta}_t(\mathbf{x}) - \eta(\mathbf{x})|] \leq \Delta_t) \end{aligned} \tag{B.33}$$

Observe that when $\mathbb{E}_{\mathbf{x}}[|\hat{\eta}_t(\mathbf{x}) - \eta(\mathbf{x})|] \leq \Delta_t$, $h_t(\mathbf{x}) \neq h^*(\mathbf{x})$ only if $|\eta(\mathbf{x}) - \frac{1}{2}| < \Delta_t$. Therefore,

$$\begin{aligned} &Pr(h_t(\mathbf{x}) \neq h^*(\mathbf{x})|\mathbb{E}_{\mathbf{x}}[|\hat{\eta}_t(\mathbf{x}) - \eta(\mathbf{x})|] \leq \Delta_t) \\ \leq\ &Pr(|\eta(\mathbf{x}) - \frac{1}{2}| < \Delta_t) \end{aligned} \tag{B.34}$$

According to the definition of Margin Assumption 1, we can upper bound the right hand side of the above inequality by the term $c \cdot \Delta_t^\gamma$. Combining this result with Inequality (B.33) and Inequality(B.34), we have, with probability at least $1-\delta$,

$$Pr_{\mathbf{x}}(h_t(\mathbf{x}) \neq h^*(\mathbf{x})) \leq c \cdot \Delta_t^\gamma$$

Let us back to consider Inequality (B.31), where we can find that the first term is upper bounded by Δ_t with probability at least $1-\delta$, while the second term is also upper bounded by $\Delta_t^{\gamma/2}$ with probability at least $1-\delta$. Since Lemma A.1 implies Lemma A.2, with probability at least $1-\delta$,

$$E(h_t) \leq c \cdot \Delta_t^{1+\frac{\gamma}{2}} \tag{B.35}$$

By substituting Equation (B.32) (i.e., the definition of Δ_t) into Inequality (B.35), we complete the proof.

B.4 Proof of Theorem 3.1 After we know that Lemma A.1, Lemma A.2 and Lemma A.3 are correct, we show the correctness of Theorem 3.1 as follows.

Proof. Suppose that the total number of rounds equals n. In the following, we use $\mathbb{E}[\cdot]$ to represent $\mathbb{E}_{\{\mathbf{x}_i, f_i\}_{i=1}^{t-1}, \mathbf{x}_t}[\cdot]$ if there is no specification. According to the definition of p_t, we have

$$\begin{aligned} &\mathbb{E}[p_t] \\ =\ &p_{min} \vee \\ &Pr(\hat{\eta}_t(\mathbf{x}_t) \geq 0.5)\mathbb{E}[Pr(\eta(\mathbf{x}_t) < 0.5)|\{\mathbf{x}_i, f_i\}_{i=1}^{t-1},\ \mathbf{x}_t] + \\ &Pr(\hat{\eta}_t(\mathbf{x}_t) < 0.5)\mathbb{E}[Pr(\eta(\mathbf{x}_t) \geq 0.5)|\{\mathbf{x}_i, f_i\}_{i=1}^{t-1},\ \mathbf{x}_t)] \\ \leq\ &p_{min} \vee Pr(\hat{\eta}_t(\mathbf{x}_t) \geq 0.5, \eta(\mathbf{x}_t) < 0.5) \\ &+Pr(\hat{\eta}_t(\mathbf{x}_t) \geq 0.5, \eta(\mathbf{x}_t) < 0.5) \\ =\ &p_{min} \vee Pr(h_t(\mathbf{x}_t) \neq h^*(\mathbf{x}_t)) \end{aligned}$$

Since we proved in Equation (B.33) that with probability at least $1-\delta$, $Pr_{\mathbf{x}}(h_t(\mathbf{x}) \neq h^*(\mathbf{x})) \leq c \cdot (1 \wedge \Delta_t^\gamma)$, where the definition of Δ_t can be found in Equation (B.32). Then, we have $\mathbb{E}[p_t] \leq p_{min} \vee c \cdot (1 \wedge \Delta_t^\gamma)$.

So, the expected total number of label requests equals $\sum_{t=1}^n \mathbb{E}[p_t]$. That is, with probability at least $1-\delta$, the label complexity is upper bounded by

$$T \quad \leq \quad p_{min} \cdot n \vee c \cdot \textstyle\sum_{t=1}^n (1 \wedge \Delta_t^\gamma) \tag{B.36}$$

According to Lemma A.3, we have

$$n \leq \frac{128 \cdot \frac{\ln \mathcal{M}(\epsilon/32, \mathcal{F}) + \ln\frac{2}{\delta}}{p_{min}^2} + \frac{2\sigma^4 A^2}{p_{min}}}{\epsilon^{\frac{4}{2+\gamma}}} \tag{B.37}$$

Let $\theta = 128 \cdot \frac{\ln \mathcal{M}(\epsilon/32, \mathcal{F}) + \ln\frac{2}{\delta}}{p_{min}} + 2\sigma^4 A^2$. Then, according to (B.32) and (B.37), we have $n \leq \frac{\theta}{p_{min}\epsilon^{\frac{4}{2+\gamma}}}$ and $\Delta_t = \sqrt{\frac{\theta}{p_{min} t}}$. After we substitute the above inequality and equation into Inequality (B.36), we have

$$T \leq \frac{\theta}{\epsilon^{\frac{4}{2+\gamma}}} \vee (c \cdot \textstyle\sum_{t=1}^n (1 \wedge (\frac{\theta}{p_{min} t}))^{\gamma/2})$$

which ends the proof.