Learning Theory Approach to Minimum Error Entropy Criterion†

Ting Hu, Jun Fan, Qiang Wu, and Ding-Xuan Zhou

Abstract

We consider the minimum error entropy (MEE) criterion and an empirical risk minimization learning algorithm in a regression setting. A learning theory approach is presented for this MEE algorithm and explicit error bounds are provided in terms of the approximation ability and capacity of the involved hypothesis space when the MEE scaling parameter is large. Novel asymptotic analysis is conducted for the generalization error associated with Renyi's entropy and a Parzen window function, to overcome technical difficulties arisen from the essential differences between the classical least squares problems and the MEE setting. A semi-norm and the involved symmetrized least squares error are introduced, which is related to some ranking algorithms.

Keywords: minimum error entropy, learning theory, Renyi's entropy, empirical risk minimization, approximation error

1 Introduction

Least squares method is a fundamental computational tool in various fields of science and engineering. It has been well understood mathematically due to the quadratic form of its related least squares loss function which is perfect to deal with problems involving Gaussian

† The work described in this paper is supported by a grant from the Research Grants Council of Hong Kong [Project No. CityU 103709]. Ting Hu (tinghuwhu.edu.cn) is with School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China. Jun Fan (junfan2@student.cityu.edu.hk) and Ding-Xuan Zhou (mazhou@cityu.edu.hk) are with Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China. Qiang Wu (wuqiangmath@gmail.com) is with Department of Mathematical Sciences, Middle Tennessee State University, Box 34 Murfreesboro, TN 37132-0001, USA.

noise (such as some from linear signal processing). The least squares method has many extensions or alternatives for different purposes. An information theoretic alternative, *minimum error entropy* (MEE) criterion [5], is based on entropy, a measurement for average information, defined in various forms such as Shannon's entropy and Renyi's entropy.

Renyi's entropy (of order 2) for a random variable e is defined in terms of its probability density function (pdf) f_e as $H(e) = -\log \mathbb{E}[f_e(e)] = -\log \int (f_e(e))^2 de$. The pdf is often unknown. Instead, to estimate the entropy, one needs to learn the density from a sample $\{e_i\}_{i=1}^m$. A practical way for approximating f_e is Parzen [8] windowing $\frac{1}{mh}\sum_{i=1}^m G(\frac{(e-e_i)^2}{2h^2})$ by means of a window function $G : \mathbb{R}_+ \to \mathbb{R}$ with $G(t) = \exp\{-t\}$ to be a typical example corresponding to Gaussian windowing. Then Renyi's entropy can be estimated through its discretized version called empirical Renyi's entropy defined by

$$\widehat{H} = -\log \frac{1}{m^2 h} \sum_{i=1}^m \sum_{j=1}^m G\left(\frac{(e_i - e_j)^2}{2h^2}\right).$$

Minimizing this computable quantity with e being an error random variable in various ways leads to different MEE algorithms. In this paper, we study an MEE learning algorithm for regression in an *empirical risk minimization* (ERM) setting.

The *regression* problem aims at learning a regression function defined on a separable metric space X (input space for learning) with values in $Y = \mathbb{R}$ (output space). To model the learning problem, we assume that ρ is a Borel probability measure on $Z := X \times Y$ and $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ is a sample independently drawn according to ρ. With a test function f on X, the error random variable e on Z for Renyi's entropy takes the form $e = y - f(x)$. Putting this into the empirical Renyi's entropy $\widehat{H}$ leads to our MEE learning algorithm in an ERM setting.

Definition 1. *Let G be a continuous function defined on $[0, \infty)$ and $h > 0$. Let $\mathcal{H}$ be a compact subset of $C(X)$. Then MEE learning algorithm associated with $\mathcal{H}$ is defined by*

$$f_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}} \left\{ -\log \frac{1}{m^2 h} \sum_{i=1}^m \sum_{j=1}^m G\left(\frac{[(y_i - f(x_i)) - (y_j - f(x_j))]^2}{2h^2}\right)\right\}. \tag{1.1}$$

The set $\mathcal{H}$ is called the hypothesis space for learning. Its compactness ensures the existence of a minimizer $f_{\mathbf{z}}$. Computational methods for solving optimization problem (1.1) and its applications in signal processing have been described in a vast MEE literature [9, 5, 6, 10]. Asymptotic behaviors of $f_{\mathbf{z}}$ for small or large MEE Scaling parameter h have also been discussed for different purposes. It has been observed that the MEE criterion has nice convergence properties when the MEE Scaling parameter h becomes large. The first purpose of

this paper is to verify this observation in the ERM setting and show that $f_{\mathbf{z}}$ approximates the regression function well with confidence. Here the *regression function* f_ρ is defined by

$$f_\rho(x) = \int_Y y d\rho(y|x), \qquad x \in X,$$

where $\rho(\cdot|x)$ is the conditional distribution of ρ at $x \in X$.

Our mathematical analysis for the convergence of $f_{\mathbf{z}}$ to f_ρ is stated in terms of the approximation ability of the hypothesis space $\mathcal{H}$ and its capacity. The approximation ability is measured by the approximation error. We assume $f_\rho \in L^2_{\rho_X}$.

Definition 2. *Define a semi-norm* $|||\cdot|||_{L^2_{\rho_X}}$ *on the space* $L^2_{\rho_X}$ *as*

$$|||f|||_{L^2_{\rho_X}} = \min_{c\in\mathbb{R}} \|f - c\|_{L^2_{\rho_X}}, \qquad f \in L^2_{\rho_X}. \tag{1.2}$$

The approximation error of the pair $(\mathcal{H}, f_\rho)$ *is defined by*

$$\mathcal{D}_{\mathcal{H}}(f_\rho) = \inf_{f\in\mathcal{H}} |||f - f_\rho|||^2_{L^2_{\rho_X}} = \inf_{f\in\mathcal{H}} \min_{c\in\mathbb{R}} \|f - f_\rho - c\|^2_{L^2_{\rho_X}}. \tag{1.3}$$

The minimizer in (1.2) is achieved by the constant $c^* = \int_X f(x) d\rho_X$ and in (1.3) by the constant $\int_X f(x) - f_\rho(x) d\rho_X$. The approximation error for the least squares ERM regression was studied in [11]. An essential difference between that and the approximation error here is an additional constant function, which is similar to an offset in support vector machines [13].

The capacity of the hypothesis space $\mathcal{H}$ is measured by covering numbers in this paper.

Definition 3. *For* $\varepsilon > 0$, *the covering number* $\mathcal{N}(\mathcal{H}, \varepsilon)$ *is defined to be the smallest integer* $l \in \mathbb{N}$ *such that there exist* l *disks with radius* ε *in* $C(X)$ *covering the set* $\mathcal{H}$. *We shall assume that for some constants* $p > 0$ *and* $A_p > 0$, *there holds*

$$\log \mathcal{N}(\mathcal{H}, \varepsilon) \le A_p \varepsilon^{-p}, \qquad \forall \varepsilon > 0. \tag{1.4}$$

The asymptotic behavior (1.4) of the covering numbers is typical in learning theory. It is satisfied by balls of Sobolev spaces on $X \subset \mathbb{R}^n$ and reproducing kernel Hilbert spaces associated with Sobolev smooth kernels. See [2, 16, 17, 14].

Throughout the paper we assume that

$$G \in C^2[0,\infty),\ G'_+(0) = -1,\ \|G''\|_\infty < \infty,\ \|tG'(t^2/2)\|_\infty := \sup_{t\in\mathbb{R}} |tG'(t^2/2)| < \infty.$$

Note that the special example $G(t) = \exp\{-t\}$ for the Gaussian windowing satisfies the above assumption with $\|G''\|_\infty = 1$ and $\|tG'(t^2/2)\|_\infty = e^{-1/2} < \infty$.

We also assume $\int_Z y^4 d\rho < \infty$ and $f_\rho \in L^\infty_{\rho_X}$. The following error bound for (1.1) with large h will be proved in Section 5.

Theorem 1. *Assume covering number condition (1.4) with some $p > 0$. Define $f_{\mathbf{z}}$ by (1.1) with $h \geq 1$ and $m > 1$. Then for any $0 < \eta \leq 1$ and $0 < \delta < 1$, with confidence $1 - \delta$ we have*

$$\||f_{\mathbf{z}} - f_\rho|\|^2_{L^2_{\rho_X}} \leq \frac{\widetilde{C}_{\mathcal{H}}}{\eta}\left(\frac{1}{h^2} + \frac{h^2}{m} + \frac{h^{\frac{2+p}{1+p}}}{m^{\frac{1}{1+p}}}\right)\log\frac{2}{\delta} + (1+\eta)\mathcal{D}_{\mathcal{H}}(f_\rho), \tag{1.5}$$

where $\widetilde{C}_{\mathcal{H}}$ is a constant independent of m, δ or h (depending on $\mathcal{H}$). In particular, if $h = m^{\frac{1}{4+3p}}$, we have

$$\||f_{\mathbf{z}} - f_\rho|\|^2_{L^2_{\rho_X}} \leq \frac{3\widetilde{C}_{\mathcal{H}}}{\eta}\left(\frac{1}{m}\right)^{\frac{2}{4+3p}}\log\frac{2}{\delta} + (1+\eta)\mathcal{D}_{\mathcal{H}}(f_\rho). \tag{1.6}$$

Remark 1. *In Theorem 1, we use a parameter $\eta > 0$ in the error bounds (1.5) and (1.6) to show that the bounds consist of two terms, one of which is essentially the approximation error $\mathcal{D}_{\mathcal{H}}(f_\rho)$ since η can be arbitrarily small. The reader can simply set $\eta = 1$ to get the main ideas of our analysis.*

When $f_\rho + c_\rho \in \mathcal{H}$ for some constant $c_\rho \in \mathbb{R}$, we know that $\mathcal{D}_{\mathcal{H}}(f_\rho) = 0$. In this case, the choice $\eta = 1$ in Theorem 1 yields the following learning rate.

Corollary 1. *Assume (1.4) with some $p > 0$ and $f_\rho + c_\rho \in \mathcal{H}$ for some constant $c_\rho \in \mathbb{R}$. Define $f_{\mathbf{z}}$ by (1.1) with $h = m^{\frac{1}{4+3p}}$ and $m > 1$. Then with confidence $1 - \delta$ we have*

$$\||f_{\mathbf{z}} - f_\rho|\|^2_{L^2_{\rho_X}} = \|f_{\mathbf{z}} - f_\rho - \int_X f_{\mathbf{z}}(x) - f_\rho(x) d\rho_X\|^2_{L^2_{\rho_X}} \leq 3\widetilde{C}_{\mathcal{H}}\left(\frac{1}{m}\right)^{\frac{2}{4+3p}}\log\frac{2}{\delta}.$$

A special example of the hypothesis space is a ball of a Sobolev space $H^s(X)$ with index $s > \frac{n}{2}$ on a domain $X \subset \mathbb{R}^n$ which satisfies (1.4) with $p = \frac{n}{s}$. When s is large enough, the positive index $\frac{n}{s}$ can be arbitrarily small. Then the power exponent of the following learning rate can be arbitrarily close to $\frac{1}{2}$.

Example 1. *Let X be a bounded domain of $\mathbb{R}^n$ with Lipschitz boundary. If $f_\rho \in H^s(X)$ for some $s > \frac{n}{2}$ and $\mathcal{H} = \{f \in H^s(X) : \|f\|_{H^s(X)} \leq R\}$ with $R \geq \|f_\rho\|_{H^s(X)}$, then for $f_{\mathbf{z}}$ defined by (1.1) with $h = m^{\frac{1}{4+3n/s}}$ and $m > 1$, with confidence $1 - \delta$ we have*

$$\||f_{\mathbf{z}} - f_\rho|\|^2_{L^2_{\rho_X}} \leq 3\widetilde{C}_{\mathcal{H}}\left(\frac{1}{m}\right)^{\frac{2}{4+3n/s}}\log\frac{2}{\delta}.$$

Our error analysis for algorithm (1.1) is based on asymptotic behaviors of the involved generalization error associated with the window function G. The Taylor expansion $G(t) \approx G(0) + G'_+(0)t$ leads us to consider the following *symmetrized least squares error* which has appeared in the literature of ranking algorithms [3, 1].

Definition 4. *The symmetrized least squares error is defined on the space $L^2_{\rho_X}$ by*

$$\mathcal{E}^{sls}(f) = \int_Z \int_Z \left[(y - f(x)) - (v - f(u))\right]^2 d\rho(x, y) d\rho(u, v), \qquad f \in L^2_{\rho_X}. \tag{1.7}$$

The second purpose of this paper is to reveal the following relation between the symmetrized least squares error and the square of the semi-norm $\|\|\cdot\|\|_{L^2_{\rho_X}}$, to be proved in the next section. We expect that this result can be applied to error analysis of some ranking algorithms.

Theorem 2. *If $\int_Z y^2 d\rho < \infty$, then*

$$\mathcal{E}^{sls}(f) = 2\|\|f - f_\rho\|\|^2_{L^2_{\rho_X}} + 2\int_Z \left[y - f_\rho(x)\right]^2 d\rho, \qquad \forall f \in L^2_{\rho_X}. \tag{1.8}$$

2 Information Error and Its Asymptotic Analysis

In this section we study a functional called *information error* or generalization error associated with the window function G defined over the space of measurable functions on X as

$$\mathcal{E}^{(h)}(f) = \int_Z \int_Z -h^2 G\left(\frac{\left[(y - f(x)) - (v - f(u))\right]^2}{2h^2}\right) d\rho(x, y) d\rho(u, v)$$

and investigate its asymptotic behavior as h tends to infinity.

Denote a constant C_ρ associated with the measure ρ as

$$C_\rho = \int_Z \left[y - f_\rho(x)\right]^2 d\rho.$$

Theorem 3. *If $G \in C^2[0, \infty)$ with $\|G''\|_\infty = \sup_{0<t<\infty} |G''(t)| < \infty$ and $\int_Z y^4 d\rho < \infty$, then for any essentially bounded measurable function f on X, we have*

$$\left|\mathcal{E}^{(h)}(f) + h^2 G(0) + G'_+(0) C_\rho + G'_+(0) \|\|f - f_\rho\|\|^2_{L^2_{\rho_X}}\right| \le \frac{64\|G''\|_\infty}{h^2} \left\{\int_Z y^4 d\rho + \|f\|^4_\infty\right\}. \tag{2.1}$$

In particular,

$$\left|\mathcal{E}^{(h)}(f) + h^2 G(0) + G'_+(0) C_\rho + G'_+(0) |||f - f_\rho|||^2_{L^2_{\rho_X}}\right| \le \frac{C'_{\mathcal{H}}}{h^2}, \qquad \forall f \in \mathcal{H},$$

where $C'_{\mathcal{H}}$ is the constant depending on ρ, h and f given by

$$C'_{\mathcal{H}} = 64\|G''\|_\infty \left\{ \int_Z y^4 d\rho + \left(\sup_{f \in \mathcal{H}} \|f\|_\infty \right)^4 \right\}.$$

Proof. By the Taylor expansion $|G(t) - G(0) - G'_+(0)t| \le \frac{\|G''\|_\infty}{2} t^2$ for $t \ge 0$, we know that

$$\begin{aligned}
&\left| \mathcal{E}^{(h)}(f) + h^2 G(0) + \int_Z \int_Z G'_+(0) \frac{[(y - f(x)) - (v - f(u))]^2}{2} d\rho(x, y) d\rho(u, v) \right| \\
&\le \frac{\|G''\|_\infty}{8h^2} \int_Z \int_Z [(y - f(x)) - (v - f(u))]^4 \, d\rho(x, y) d\rho(u, v) \\
&\le \frac{4\|G''\|_\infty}{h^2} \int_Z [y - f(x)]^4 \, d\rho \le \frac{4\|G''\|_\infty}{h^2} \int_Z [y - f(x)]^4 \, d\rho \\
&\le \frac{64\|G''\|_\infty}{h^2} \left\{ \int_Z y^4 d\rho + \|f\|^4_\infty \right\}.
\end{aligned}$$

This together with Theorem 2 proves bound (2.1) and hence our conclusion. □

Applying Theorem 3 to a function $f \in \mathcal{H}$ and $f_\rho \in L^\infty_{\rho_X}$ yields the following relation on the excess generalization error $\mathcal{E}^{(h)}(f) - \mathcal{E}^{(h)}(f_\rho)$.

Corollary 2. *Under the condition of Theorem 3, if $f_\rho \in L^\infty_{\rho_X}$, we have*

$$\left|\mathcal{E}^{(h)}(f) - \mathcal{E}^{(h)}(f_\rho) + G'_+(0) |||f - f_\rho|||^2_{L^2_{\rho_X}}\right| \le \frac{C''_{\mathcal{H}}}{h^2}, \qquad \forall f \in \mathcal{H},$$

where

$$C''_{\mathcal{H}} = C'_{\mathcal{H}} + 64\|G'''\|_\infty \left\{ \int_Z y^4 d\rho + \|f_\rho\|^4_\infty \right\}.$$

We end this section by proving Theorem 2 stated in the introduction.

Proof of Theorem 2. By the definition of the regression function, we have

$$\begin{aligned}
\mathcal{E}^{sls}(f) &= \int_Z \int_Z \left[(y - f(x)) - (v - f(u))\right]^2 d\rho(x,y) d\rho(u,v) \\
&= \int_Z \left\{ \int_X \left[f_\rho(x) - f(x) - (v - f(u))\right]^2 d\rho_X(x) + \int_Z \left[y - f_\rho(x)\right]^2 d\rho(x,y) \right\} d\rho(u,v) \\
&= \int_X \int_X \left[f_\rho(x) - f(x) - (f_\rho(u) - f(u))\right]^2 d\rho_X(x) d\rho_X(u) \\
&\quad + \int_X \int_Z \left[v - f_\rho(u)\right]^2 d\rho(u,v) d\rho_X(x) + \int_Z \int_Z \left[y - f_\rho(x)\right]^2 d\rho(x,y) d\rho(u,v) \\
&= \int_X \int_X \left[f(x) - f_\rho(x) - (f(u) - f_\rho(u))\right]^2 d\rho_X(x) d\rho_X(u) + 2 \int_Z \left[y - f_\rho(x)\right]^2 d\rho.
\end{aligned}$$

Let c^* be the best approximation in $L^2_{\rho_X}$ of $f - f_\rho$ from the subspace of the constant functions. That is,

$$c^* = \arg\min_{c \in \mathbb{R}} \|f - f_\rho - c\|_{L^2_{\rho_X}}.$$

Then the function $f - f_\rho - c^*$ is orthogonal to any constant function in the Hilbert space $L^2_{\rho_X}$, and $\|f - f_\rho - c^*\|_{L^2_{\rho_X}} = \||f - f_\rho|\|_{L^2_{\rho_X}}$. It follows that

$$\int_X \left(f(x) - f_\rho(x) - c\right)^2 d\rho_X(x) = |c - c^*|^2 + \||f - f_\rho|\|^2, \qquad \forall c \in \mathbb{R}.$$

With $c = f(u) - f_\rho(u) \in \mathbb{R}$, this implies that for any fixed $u \in X$, there holds

$$\int_X \left[f(x) - f_\rho(x) - (f(u) - f_\rho(u))\right]^2 d\rho_X(x) = |f(u) - f_\rho(u) - c^*|^2 + \||f - f_\rho|\|^2_{L^2_{\rho_X}}.$$

Hence

$$\begin{aligned}
&\int_X \int_X \left[f(x) - f_\rho(x) - (f(u) - f_\rho(u))\right]^2 d\rho_X(x) d\rho_X(u) \\
&= \int_X |f(u) - f_\rho(u) - c^*|^2 d\rho_X(u) + \||f - f_\rho|\|^2_{L^2_{\rho_X}} = 2\||f - f_\rho|\|^2_{L^2_{\rho_X}}.
\end{aligned}$$

Then the desired equality follows. This proves Theorem 2. □

3 Error Decomposition for the ERM Algorithm

Error decomposition has been a standard technique to analyze least squares ERM regression algorithms [2, 4, 12, 15]. It decomposes the error $\|f - f_\rho\|^2_{L^2_{\rho_X}}$ into the sum of $\mathcal{E}^{ls}(f) - \mathcal{E}^{ls}(f^{ls}_{\mathcal{H}})$

(sample error) and $\mathcal{E}^{ls}(f_{\mathcal{H}}^{ls}) - \mathcal{E}^{ls}(f_\rho) = \|f_{\mathcal{H}}^{ls} - f_\rho\|_{L^2_{\rho_X}}^2$ (approximation error) where $\mathcal{E}^{ls}(f) = \int_Z (f(x)-y)^2 d\rho$ and $f_{\mathcal{H}}^{ls}$ is a minimizer (called target function) of $\mathcal{E}^{ls}(f)$ in $\mathcal{H}$. A technical difficulty arises for the error decomposition of ERM algorithm (1.1) since there might be two ways to define a *target function* in $\mathcal{H}$, one to minimize the information error and the other to minimize the distance to f_ρ under the semi-norm $|||\cdot|||_{L^2_{\rho_X}}$. These possible candidates for the target function are defined as

$$f_{\mathcal{H}} := \arg\min_{f\in\mathcal{H}} \mathcal{E}^{(h)}(f), \tag{3.1}$$

$$f_{approx} := \arg\min_{f\in\mathcal{H}} |||f - f_\rho|||_{L^2_{\rho_X}} = \arg\min_{f\in\mathcal{H}} \min_{c\in\mathbb{R}} \|f - f_\rho - c\|_{L^2_{\rho_X}}. \tag{3.2}$$

The first technical novelty of this paper is to show that then the MEE scaling parameter h is large, these two functions are actually very close.

Theorem 4. *Under the condition of Corollary 2, if $G'_+(0) < 0$, then*

$$\mathcal{E}^{(h)}(f_{approx}) \le \mathcal{E}^{(h)}(f_{\mathcal{H}}) + \frac{2C''_{\mathcal{H}}}{h^2}$$

and

$$|||f_{\mathcal{H}} - f_\rho|||_{L^2_{\rho_X}}^2 \le |||f_{approx} - f_\rho|||_{L^2_{\rho_X}}^2 + \frac{2C''_{\mathcal{H}}}{-G'_+(0)h^2}.$$

Proof. By Corollary 2 and the definitions of $f_{\mathcal{H}}$ and f_{approx}, we have

$$\begin{aligned}
&\mathcal{E}^{(h)}(f_{\mathcal{H}}) - \mathcal{E}^{(h)}(f_\rho) \le \mathcal{E}^{(h)}(f_{approx}) - \mathcal{E}^{(h)}(f_\rho) \le -G'_+(0)|||f_{approx} - f_\rho|||_{L^2_{\rho_X}}^2 + \frac{C''_{\mathcal{H}}}{h^2} \\
&\le -G'_+(0)|||f_{\mathcal{H}} - f_\rho|||_{L^2_{\rho_X}}^2 + \frac{C''_{\mathcal{H}}}{h^2} \le \mathcal{E}^{(h)}(f_{\mathcal{H}}) - \mathcal{E}^{(h)}(f_\rho) + \frac{2C''_{\mathcal{H}}}{h^2} \\
&\le -G'_+(0)|||f_{approx} - f_\rho|||_{L^2_{\rho_X}}^2 + \frac{3C''_{\mathcal{H}}}{h^2}.
\end{aligned}$$

Then the desired inequalities follow. □

Corollary 2 actually yields the following error decomposition for our algorithm.

Lemma 1. *Under the condition of Corollary 2, if $G'_+(0) < 0$, then*

$$|||f_{\mathbf{z}} - f_\rho|||_{L^2_{\rho_X}}^2 \le \frac{1}{-G'_+(0)} \left\{\mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}})\right\} + |||f_{approx} - f_\rho|||_{L^2_{\rho_X}}^2 + \frac{2C''_{\mathcal{H}}}{-G'_+(0)h^2}. \tag{3.3}$$

Proof. By Corollary 2,

$$\begin{aligned}
-G'_+(0)|||f_{\mathbf{z}} - f_\rho|||^2 \quad &\le \mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_\rho) + \frac{C''_{\mathcal{H}}}{h^2} \\
&\le \left\{\mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}})\right\} + \mathcal{E}^{(h)}(f_{\mathcal{H}}) - \mathcal{E}^{(h)}(f_\rho) + \frac{C''_{\mathcal{H}}}{h^2}.
\end{aligned}$$

Since $f_{approx} \in \mathcal{H}$, the definition of $f_{\mathcal{H}}$ tells us that

$$\mathcal{E}^{(h)}(f_{\mathcal{H}}) - \mathcal{E}^{(h)}(f_\rho) \leq \mathcal{E}^{(h)}(f_{approx}) - \mathcal{E}^{(h)}(f_\rho).$$

Applying Corollary 2 to the above bound implies

$$-G'_+(0)|||f_{\mathbf{z}} - f_\rho|||^2 \leq \left\{\mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}})\right\} - G'_+(0)|||f_{approx} - f_\rho|||^2 + \frac{2C''_{\mathcal{H}}}{h^2}.$$

Then desired error decomposition (3.3) follows. □

4 Sample Error Estimates

In this section, we estimate the sample error $\mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}})$. In this step, we demonstrate our second technical novelty. Define the empirical information error for measurable functions f on X as

$$\mathcal{E}^{(h)}_{\mathbf{z}}(f) = -\frac{h^2}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i} G\left(\frac{[(y_i - f(x_i)) - (y_j - f(x_j))]^2}{2h^2}\right).$$

Then

$$\mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}}) = \mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}_{\mathbf{z}}(f_{\mathbf{z}}) + \mathcal{E}^{(h)}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}_{\mathbf{z}}(f_{\mathcal{H}}) + \mathcal{E}^{(h)}_{\mathbf{z}}(f_{\mathcal{H}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}}).$$

By the definition of $f_{\mathbf{z}}$, we have $\mathcal{E}^{(h)}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}_{\mathbf{z}}(f_{\mathcal{H}}) \leq 0$. Hence

$$\mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}}) \leq \mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}_{\mathbf{z}}(f_{\mathbf{z}}) + \mathcal{E}^{(h)}_{\mathbf{z}}(f_{\mathcal{H}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}}) = \mathcal{S}_1 + \mathcal{S}_2, \tag{4.1}$$

where

$$\begin{aligned} \mathcal{S}_1 : &= \left[\mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_\rho)\right] - \left[\mathcal{E}^{(h)}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}_{\mathbf{z}}(f_\rho)\right], \\ \mathcal{S}_2 : &= \left[\mathcal{E}^{(h)}_{\mathbf{z}}(f_{\mathcal{H}}) - \mathcal{E}^{(h)}_{\mathbf{z}}(f_\rho)\right] - \left[\mathcal{E}^{(h)}(f_{\mathcal{H}}) - \mathcal{E}^{(h)}(f_\rho)\right] \end{aligned}$$

We use Hoeffding's probability inequality for U-statistics [7] to bound $\mathcal{S}_1$ and $\mathcal{S}_2$.

Lemma 2. *If U is a symmetric real-valued function on $Z \times Z$ satisfying $a \leq U(z, z') \leq b$ almost surely and $\mathrm{var}(U) = \sigma^2$, then for any $\varepsilon > 0$,*

$$\mathrm{Prob}\left\{\frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i} U(z_i, z_j) - \mathbb{E}U \geq \varepsilon\right\} \leq \exp\left\{-\frac{(m-1)\varepsilon^2}{4\sigma^2 + (4/3)(b-a)\varepsilon}\right\}.$$

Lemma 3. *For any $0 < \delta < 1$, with confidence of $1 - \frac{\delta}{2}$, there holds*

$$\mathcal{S}_2 \leq \|tG'(t^2/2)\|_\infty \log \frac{2}{\delta} \left\{ \frac{6}{m-1} + \frac{4\||f_{\mathcal{H}} - f_\rho|\|_{L^2_{\rho_X}}}{\sqrt{m-1}} \right\} h.$$

Proof. Let U be the symmetric real-valued function on $Z \times Z$ defined in terms of two variables $z = (x, y), z' = (u, v) \in Z$ as

$$U(z, z') = -h^2 G\left(\frac{[(y - f_{\mathcal{H}}(x)) - (v - f_{\mathcal{H}}(u))]^2}{2h^2} \right) + h^2 G\left(\frac{[(y - f_\rho(x)) - (v - f_\rho(u))]^2}{2h^2} \right).$$

Define a function g on $\mathbb{R}$ by

$$g(t) = G(t^2/2), \qquad t \in \mathbb{R}. \tag{4.2}$$

We see that $g \in C^2(\mathbb{R})$, $g(0) = G(0)$, $g'(t) = tG'(t^2/2)$ with $g'(0) = 0$. Moreover,

$$U(z, z') \;=\; -h^2 g\left(\frac{(y - f_{\mathcal{H}}(x)) - (v - f_{\mathcal{H}}(u))}{h} \right) + h^2 g\left(\frac{(y - f_\rho(x)) - (v - f_\rho(u))}{h} \right).$$

It follows that

$$|U(z, z')| \leq h^2 \|g'\|_\infty \frac{|[f_{\mathcal{H}}(u) - f_\rho(u)] - [f_{\mathcal{H}}(x) - f_\rho(x)]|}{h}.$$

This tells us that

$$|U(z, z')| \leq 2\|tG'(t^2/2)\|_\infty \|f_{\mathcal{H}} - f_\rho\|_\infty h$$

almost surely. Moreover, putting the constant

$$c^* = \arg\min_{c \in \mathbb{R}} \|f_{\mathcal{H}} - f_\rho - c\|_{L^2_{\rho_X}}$$

into

$$[f_{\mathcal{H}}(u) - f_\rho(u)] - [f_{\mathcal{H}}(x) - f_\rho(x)] = [f_{\mathcal{H}}(u) - f_\rho(u) - c^*] - [f_{\mathcal{H}}(x) - f_\rho(x) - c^*]$$

we find that

$$\mathrm{var}(U) \leq \mathbb{E}(U^2) \leq 4\|tG'(t^2/2)\|^2_\infty \|f_{\mathcal{H}} - f_\rho - c^*\|^2_{L^2_{\rho_X}} h^2 = 4\|tG'(t^2/2)\|^2_\infty \||f_{\mathcal{H}} - f_\rho|\|^2_{L^2_{\rho_X}} h^2.$$

Note that $\mathbb{E}U = \mathcal{E}^{(h)}(f_{\mathcal{H}}) - \mathcal{E}^{(h)}(f_\rho)$ and $\frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i} U(z_i, z_j) = \mathcal{E}^{(h)}_{\mathbf{z}}(f_{\mathcal{H}}) - \mathcal{E}^{(h)}_{\mathbf{z}}(f_\rho)$. Then by Lemma 2 we have

$$\mathrm{Prob}\left\{ \mathcal{S}_2 \geq \varepsilon \right\} \leq \exp\left\{ -\frac{(m-1)\varepsilon^2}{16\|tG'(t^2/2)\|^2_\infty \||f_{\mathcal{H}} - f_\rho|\|^2_{L^2_{\rho_X}} h^2 + 6\|tG'(t^2/2)\|_\infty \|f_{\mathcal{H}} - f_\rho\|_\infty h\varepsilon} \right\}.$$

By setting the above probability bound to be $\delta/2$ and solving the corresponding quadratic equality, we know that with confidence at least $1-\frac{\delta}{2}$, we have

$$\mathcal{S}_2 \le \frac{6\|tG'(t^2/2)\|_\infty \|f_{\mathcal{H}}-f_\rho\|_\infty h}{m-1}\log\frac{2}{\delta} + \frac{4\|tG'(t^2/2)\|_\infty \|\|f_{\mathcal{H}}-f_\rho\|\|_{L^2_{\rho_X}} h}{\sqrt{m-1}}\log\frac{2}{\delta}.$$

Then our desired estimate follows. □

Estimating $\mathcal{S}_1$ is more involved. One possible way to get tight bounds is by Hoeffding's decomposition, as done for ranking algorithms in [3]. We take another way of applying a ratio probability inequality. Since $f_{\mathbf{z}}$ depends on the sample $\mathbf{z}$, we use covering numbers of the set $\mathcal{H}$ to give our bounds.

A key difficulty in estimating $\mathcal{S}_1$ is caused by an essential difference between the information error $\mathcal{E}^{(h)}(f)$ associated with the entropy and the generalization error associated with the least squares loss: $\mathcal{E}^{(h)}(f_j)-\mathcal{E}^{(h)}(f_\rho)$ is not equal to $\|f_j-f_\rho\|^2_{L^2_{\rho_X}}$. The second technical novelty of this paper is to apply Corollary 2 to bound a variance term (σ^2 in Lemma 2) by the following inequality

$$\text{if } \varepsilon \ge \frac{C''_{\mathcal{H}}}{h^2}, \text{then } \mathcal{E}^{(h)}(f)-\mathcal{E}^{(h)}(f_\rho)+\varepsilon \ge \mathcal{E}^{(h)}(f)-\mathcal{E}^{(h)}(f_\rho)+\frac{C''_{\mathcal{H}}}{h^2} \ge -G'_+(0)\|\|f-f_\rho\|\|^2_{L^2_{\rho_X}}. \quad (4.3)$$

While this inequality is a easy consequence of Corollary 2, its special role in our estimation of $\mathcal{S}_1$ needs to be demonstrated with details in the following proof.

Lemma 4. *Under conditions of Corollary 2, if* $G'_+(0)<0$, $\|tG'(t^2/2)\|_\infty<\infty$ *and*

$$\varepsilon \ge \frac{C''_{\mathcal{H}}}{h^2}, \quad (4.4)$$

then we have

$$\text{Prob}\left\{\sup_{f\in\mathcal{H}}\frac{\left[\mathcal{E}^{(h)}(f)-\mathcal{E}^{(h)}(f_\rho)\right]-\left[\mathcal{E}^{(h)}_{\mathbf{z}}(f)-\mathcal{E}^{(h)}_{\mathbf{z}}(f_\rho)\right]}{\sqrt{\mathcal{E}^{(h)}(f)-\mathcal{E}^{(h)}(f_\rho)+\varepsilon}} > 4\sqrt{\varepsilon}\right\} \le \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{2\|tG'(t^2/2)\|_\infty h}\right)$$
$$\exp\left\{-\frac{(m-1)\varepsilon}{16\|tG'(t^2/2)\|^2_\infty h^2/(-G'_+(0))+6\|tG'(t^2/2)\|_\infty \sup_{f\in\mathcal{H}}\|f-f_\rho\|_\infty h}\right\}.$$

Proof. From the argument in the proof of Lemma 3, we see that for any $f,g\in\mathcal{H}$, we have

$$\left|\mathcal{E}^{(h)}(f)-\mathcal{E}^{(h)}(g)\right| \le 2\|tG'(t^2/2)\|_\infty\|f-g\|_\infty h$$

and almost surely

$$\left|\mathcal{E}^{(h)}_{\mathbf{z}}(f)-\mathcal{E}^{(h)}_{\mathbf{z}}(g)\right| \le 2\|tG'(t^2/2)\|_\infty\|f-g\|_\infty h$$

Thus, if $\|f - f_j\|_\infty \leq \frac{\varepsilon}{2\|tG'(t^2/2)\|_\infty h}$, then

$$\frac{\left[\mathcal{E}^{(h)}(f) - \mathcal{E}^{(h)}(f_\rho)\right] - \left[\mathcal{E}_{\mathbf{z}}^{(h)}(f) - \mathcal{E}_{\mathbf{z}}^{(h)}(f_\rho)\right]}{\sqrt{\mathcal{E}^{(h)}(f) - \mathcal{E}^{(h)}(f_\rho) + \varepsilon}} > 4\sqrt{\varepsilon}$$

implies

$$\frac{\left[\mathcal{E}^{(h)}(f_j) - \mathcal{E}^{(h)}(f_\rho)\right] - \left[\mathcal{E}_{\mathbf{z}}^{(h)}(f_j) - \mathcal{E}_{\mathbf{z}}^{(h)}(f_\rho)\right]}{\sqrt{\mathcal{E}^{(h)}(f_j) - \mathcal{E}^{(h)}(f_\rho) + \varepsilon}} > \sqrt{\varepsilon}.$$

Thus by taking $\{f_j\}_{j=1}^N$ to be an $\frac{\varepsilon}{2\|tG'(t^2/2)\|_\infty h}$ net of the set $\mathcal{H}$ with N being the covering number $\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{2\|tG'(t^2/2)\|_\infty h}\right)$, we find

$$\begin{aligned}
&\text{Prob}\left\{\sup_{f\in\mathcal{H}} \frac{\left[\mathcal{E}^{(h)}(f) - \mathcal{E}^{(h)}(f_\rho)\right] - \left[\mathcal{E}_{\mathbf{z}}^{(h)}(f) - \mathcal{E}_{\mathbf{z}}^{(h)}(f_\rho)\right]}{\sqrt{\mathcal{E}^{(h)}(f) - \mathcal{E}^{(h)}(f_\rho) + \varepsilon}} > 4\sqrt{\varepsilon}\right\} \\
&\leq \text{Prob}\left\{\sup_{j=1,\ldots,N} \frac{\left[\mathcal{E}^{(h)}(f_j) - \mathcal{E}^{(h)}(f_\rho)\right] - \left[\mathcal{E}_{\mathbf{z}}^{(h)}(f_j) - \mathcal{E}_{\mathbf{z}}^{(h)}(f_\rho)\right]}{\sqrt{\mathcal{E}^{(h)}(f_j) - \mathcal{E}^{(h)}(f_\rho) + \varepsilon}} > \sqrt{\varepsilon}\right\} \\
&\leq \sum_{j=1,\ldots,N} \text{Prob}\left\{\frac{\left[\mathcal{E}^{(h)}(f_j) - \mathcal{E}^{(h)}(f_\rho)\right] - \left[\mathcal{E}_{\mathbf{z}}^{(h)}(f_j) - \mathcal{E}_{\mathbf{z}}^{(h)}(f_\rho)\right]}{\sqrt{\mathcal{E}^{(h)}(f_j) - \mathcal{E}^{(h)}(f_\rho) + \varepsilon}} > \sqrt{\varepsilon}\right\}.
\end{aligned}$$

Fix $j \in \{1, \ldots, N\}$. Consider the function

$$U(z, z') = h^2 G\left(\frac{[(y - f_j(x)) - (v - f_j(u))]^2}{2h^2}\right) - h^2 G\left(\frac{[(y - f_\rho(x)) - (v - f_\rho(u))]^2}{2h^2}\right).$$

It satisfies

$$|U(z, z')| \leq 2\|tG'(t^2/2)\|_\infty \|f_j - f_\rho\|_\infty h$$

almost surely and

$$\text{var}(U) \leq 4\|tG'(t^2/2)\|_\infty^2 \||f_j - f_\rho|\|_{L^2_{\rho_X}}^2 h^2.$$

Also, $-\mathbb{E}U = \mathcal{E}^{(h)}(f_\mathcal{H}) - \mathcal{E}^{(h)}(f_\rho)$ and $\frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j\neq i} U(z_i, z_j) = -\left\{\mathcal{E}_{\mathbf{z}}^{(h)}(f_j) - \mathcal{E}_{\mathbf{z}}^{(h)}(f_\rho)\right\}$. Set

$$\widetilde{\varepsilon} = \mathcal{E}^{(h)}(f_j) - \mathcal{E}^{(h)}(f_\rho) + \varepsilon, \tag{4.5}$$

by Lemma 2 we find

$$\begin{aligned}
&\text{Prob}\left\{\frac{\left[\mathcal{E}^{(h)}(f_j)-\mathcal{E}^{(h)}(f_\rho)\right]-\left[\mathcal{E}^{(h)}_{\mathbf{z}}(f_j)-\mathcal{E}^{(h)}_{\mathbf{z}}(f_\rho)\right]}{\sqrt{\mathcal{E}^{(h)}(f_j)-\mathcal{E}^{(h)}(f_\rho)+\varepsilon}}>\sqrt{\varepsilon}\right\}\\
&=\text{Prob}\left\{\left[\mathcal{E}^{(h)}(f_j)-\mathcal{E}^{(h)}(f_\rho)\right]-\left[\mathcal{E}^{(h)}_{\mathbf{z}}(f_j)-\mathcal{E}^{(h)}_{\mathbf{z}}(f_\rho)\right]>\sqrt{\varepsilon}\sqrt{\widetilde{\varepsilon}}\right\}\\
&\le\exp\left\{-\frac{(m-1)\varepsilon\widetilde{\varepsilon}}{16\|tG'(t^2/2)\|_\infty^2\||f_j-f_\rho|\|_{L^2_{\rho_X}}^2h^2+6\|tG'(t^2/2)\|_\infty\|f_j-f_\rho\|_\infty h\sqrt{\varepsilon}\sqrt{\widetilde{\varepsilon}}}\right\}.
\end{aligned}$$

Now we apply the important relation (4.3) to the function $f=f_j$ and find by noting the definition (4.5) for $\widetilde{\varepsilon}$ that

$$\||f_j-f_\rho|\|_{L^2_{\rho_X}}^2\le\frac{\widetilde{\varepsilon}}{-G'_+(0)}=\frac{\mathcal{E}^{(h)}(f_j)-\mathcal{E}^{(h)}(f_\rho)+\varepsilon}{-G'_+(0)}.$$

This together with the inequalities $\frac{\sqrt{\varepsilon}\sqrt{\widetilde{\varepsilon}}}{\widetilde{\varepsilon}}\le 1$ and $\|f_j-f_\rho\|_\infty\le\sup_{f\in\mathcal{H}}\|f-f_\rho\|_\infty$ gives

$$\begin{aligned}
&\text{Prob}\left\{\frac{\left[\mathcal{E}^{(h)}(f_j)-\mathcal{E}^{(h)}(f_\rho)\right]-\left[\mathcal{E}^{(h)}_{\mathbf{z}}(f_j)-\mathcal{E}^{(h)}_{\mathbf{z}}(f_\rho)\right]}{\sqrt{\mathcal{E}^{(h)}(f_j)-\mathcal{E}^{(h)}(f_\rho)+\varepsilon}}>\sqrt{\varepsilon}\right\}\\
&\le\exp\left\{-\frac{(m-1)\varepsilon}{16\|tG'(t^2/2)\|_\infty^2h^2/(-G'_+(0))+6\|tG'(t^2/2)\|_\infty\sup_{f\in\mathcal{H}}\|f-f_\rho\|_\infty h}\right\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\text{Prob}\left\{\sup_{f\in\mathcal{H}}\frac{\left[\mathcal{E}^{(h)}(f)-\mathcal{E}^{(h)}(f_\rho)\right]-\left[\mathcal{E}^{(h)}_{\mathbf{z}}(f)-\mathcal{E}^{(h)}_{\mathbf{z}}(f_\rho)\right]}{\sqrt{\mathcal{E}^{(h)}(f)-\mathcal{E}^{(h)}(f_\rho)+\varepsilon}}>4\sqrt{\varepsilon}\right\}\\
&\le N\exp\left\{-\frac{(m-1)\varepsilon}{16\|tG'(t^2/2)\|_\infty^2h^2/(-G'_+(0))+6\|tG'(t^2/2)\|_\infty\sup_{f\in\mathcal{H}}\|f-f_\rho\|_\infty h}\right\}.
\end{aligned}$$

This proves the required inequality. □

We are in a position to bound $\mathcal{S}_1$ and hence the sample error.

Proposition 1. *Under conditions of Corollary 2, if $G'_+(0)<0$, $\|tG'(t^2/2)\|_\infty<\infty$ and covering number condition (1.4) is satisfied, then with confidence of $1-\delta$, we have*

$$\mathcal{E}^{(h)}(f_{\mathbf{z}})-\mathcal{E}^{(h)}(f_{\mathcal{H}})\le C_{\mathcal{H},G,\rho}\max\left\{h^{-2},\ \frac{h^2+h}{m-1}\log\frac{2}{\delta},\ \frac{\||f_{\mathcal{H}}-f_\rho|\|_{L^2_{\rho_X}}h}{\sqrt{m-1}},\ \frac{h+h^{(2+p)/(1+p)}}{(m-1)^{\frac{1}{1+p}}}\right\},$$

where $C_{\mathcal{H},G,\rho}$ is a constant independent of m,δ or h.

Proof. By condition (1.4), the probability bound in Lemma 4 is at most

$$\exp\left\{A_p\left(\frac{2\|tG'(t^2/2)\|_\infty h}{\varepsilon}\right)^p - \frac{(m-1)\varepsilon}{C_h}\right\},$$

where

$$C_h = 16\|tG'(t^2/2)\|_\infty^2 h^2/(-G'_+(0)) + 6\|tG'(t^2/2)\|_\infty \sup_{f\in\mathcal{H}}\|f-f_\rho\|_\infty h.$$

Requiring this bound to be at most $\delta/2$ is equivalent to the inequality

$$\varepsilon^{1+p} - \frac{C_h}{m-1}\log\frac{2}{\delta}\varepsilon^p - A_p\frac{(2\|tG'(t^2/2)\|_\infty h)^p\, C_h}{m-1} \geq 0.$$

By Lemma 7.2 in [4], we know that the above inequality is satisfied as long as

$$\varepsilon \geq \max\left\{\frac{2C_h}{m-1}\log\frac{2}{\delta},\ \left(A_p\left(2\|tG'(t^2/2)\|_\infty h\right)^p C_h\right)^{1/(1+p)}(m-1)^{-\frac{1}{1+p}}\right\}.$$

Now let us take a constant

$$C'''_{\mathcal{H}} = C''_{\mathcal{H}} + 32\|tG'(t^2/2)\|_\infty^2/(-G'_+(0)) + 12\|tG'(t^2/2)\|_\infty \sup_{f\in\mathcal{H}}\|f-f_\rho\|_\infty + 16A_p^{1/(1+p)}$$
$$\left(\|tG'(t^2/2)\|_\infty^{(2+p)/(1+p)}(-G'_+(0))^{-1/(1+p)} + \|tG'(t^2/2)\|_\infty \sup_{f\in\mathcal{H}}\|f-f_\rho\|_\infty^{1/(1+p)}\right)$$

and take ε to be ε^* given by

$$\varepsilon^* = C'''_{\mathcal{H}}\max\left\{h^{-2},\ \frac{h^2+h}{m-1}\log\frac{2}{\delta},\ (h+h^{(2+p)/(1+p)})(m-1)^{-\frac{1}{1+p}}\right\}.$$

With this choice, by Lemma 4, we know that with confidence at least $1-\delta/2$, we have

$$\sup_{f\in\mathcal{H}}\frac{\left[\mathcal{E}^{(h)}(f)-\mathcal{E}^{(h)}(f_\rho)\right]-\left[\mathcal{E}^{(h)}_{\mathbf{z}}(f)-\mathcal{E}^{(h)}_{\mathbf{z}}\right]}{\sqrt{\mathcal{E}^{(h)}(f)-\mathcal{E}^{(h)}(f_\rho)+\varepsilon^*}} \leq 4\sqrt{\varepsilon^*},$$

which implies in particular

$$\left[\mathcal{E}^{(h)}(f_{\mathbf{z}})-\mathcal{E}^{(h)}(f_\rho)\right]-\left[\mathcal{E}^{(h)}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}^{(h)}_{\mathbf{z}}(f_\rho)\right] \leq 4\sqrt{\varepsilon^*}\sqrt{\mathcal{E}^{(h)}(f_{\mathbf{z}})-\mathcal{E}^{(h)}(f_\rho)+\varepsilon^*}.$$

This together with the elementary inequality $4\sqrt{a}\sqrt{b} \leq \frac{a}{2}+8b$ that

$$\mathcal{S}_1 \leq \frac{1}{2}\left(\mathcal{E}^{(h)}(f_{\mathbf{z}})-\mathcal{E}^{(h)}(f_\rho)\right)+12\varepsilon^*.$$

We combine this estimate with Lemma 3 and (4.1), and see that with confidence of $1-\delta$,

$$\mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}}) \leq \frac{1}{2}\left(\mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_\rho)\right) + 12\varepsilon^* \\ + \|tG'(t^2/2)\|_\infty \log\frac{2}{\delta} h \left\{ \frac{6}{m-1} + \frac{4|||f_{\mathcal{H}} - f_\rho|||_{L^2_{\rho_X}}}{\sqrt{m-1}} \right\}$$

and hence

$$\mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}}) \leq 24\varepsilon^* + \|tG'(t^2/2)\|_\infty \log\frac{2}{\delta} h \left\{ \frac{12}{m-1} + \frac{8|||f_{\mathcal{H}} - f_\rho|||_{L^2_{\rho_X}}}{\sqrt{m-1}} \right\}.$$

By setting the constant $C_{\mathcal{H},G,\rho}$ as

$$C_{\mathcal{H},G,\rho} = 24C'''_{\mathcal{H}} + 20\|tG'(t^2/2)\|_\infty,$$

we verify the conclusion of Proposition 1. □

5 Proof of the Error Bounds

We are now in a position to prove our error bounds for algorithm (1.1) stated in the introduction.

Proof of Theorem 1. Since $G'_+(0) = -1$ and $|||f_{approx} - f_\rho|||^2_{L^2_{\rho_X}} = \mathcal{D}_{\mathcal{H}}(f_\rho)$, by Lemma 1, we have

$$|||f_{\mathbf{z}} - f_\rho|||^2_{L^2_{\rho_X}} \leq \left\{\mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}})\right\} + \mathcal{D}_{\mathcal{H}}(f_\rho) + \frac{2C''_{\mathcal{H}}}{h^2}.$$

Combining this with the bound in Proposition 1 for the sample error and the restriction $h \geq 1$ tells us that with confidence of $1-\delta$,

$$|||f_{\mathbf{z}} - f_\rho|||^2_{L^2_{\rho_X}} \leq C_{\mathcal{H},G,\rho} \max\left\{ h^{-2},\ \frac{4h^2}{m}\log\frac{2}{\delta},\ \frac{2|||f_{\mathcal{H}} - f_\rho|||_{L^2_{\rho_X}} h}{\sqrt{m}},\ \frac{4h^{\frac{2+p}{1+p}}}{m^{\frac{1}{1+p}}} \right\} + \mathcal{D}_{\mathcal{H}}(f_\rho) + \frac{2C''_{\mathcal{H}}}{h^2}.$$

But

$$C_{\mathcal{H},G,\rho}\frac{2|||f_{\mathcal{H}} - f_\rho|||_{L^2_{\rho_X}} h}{\sqrt{m}} \leq \eta|||f_{\mathcal{H}} - f_\rho|||^2_{L^2_{\rho_X}} + \frac{C^2_{\mathcal{H},G,\rho}h^2}{\eta m}$$

and according to Theorem 4, $|||f_{\mathcal{H}} - f_\rho|||^2_{L^2_{\rho_X}} \leq \mathcal{D}_{\mathcal{H}}(f_\rho) + \frac{2C''_{\mathcal{H}}}{h^2}$. It follows that with confidence

of $1-\delta$,

$$\begin{aligned}
|||f_{\mathbf{z}} - f_\rho|||^2_{L^2_{\rho_X}} &\leq C_{\mathcal{H},G,\rho} \max\left\{h^{-2},\ \frac{h^2}{m}\left(2\log\frac{2}{\delta} + \frac{C_{\mathcal{H},G,\rho}}{\eta}\right),\ \frac{4h^{\frac{2+p}{1+p}}}{m^{\frac{1}{1+p}}}\right\} \\
&\quad +(1+\eta)\mathcal{D}_{\mathcal{H}}(f_\rho) + \frac{(2+2\eta)C''_{\mathcal{H}}}{h^2} \\
&\leq \frac{\widetilde{C}_{\mathcal{H}}}{\eta}\left(\frac{1}{h^2} + \frac{h^2}{m} + \frac{h^{\frac{2+p}{1+p}}}{m^{\frac{1}{1+p}}}\right)\log\frac{2}{\delta} + (1+\eta)\mathcal{D}_{\mathcal{H}}(f_\rho),
\end{aligned}$$

where

$$\widetilde{C}_{\mathcal{H}} = \max\left\{C_{\mathcal{H},G,\rho} + 4C''_{\mathcal{H}},\ C_{\mathcal{H},G,\rho}\left(2 + C_{\mathcal{H},G,\rho}\right),\ 4C_{\mathcal{H},G,\rho}\right\}.$$

This proves (1.5) and thereby Theorem 1. □

References

[1] S. Agarwal and P. Niyogi, Generalization bounds for ranking algorithms via algorithmic stability, J. Machine Learning Research **10** (2009), 441–474.

[2] M. Anthony and P. Bartlett, Neural Network Learning: Theoretical Foundations, Cambridge University Press, 1999.

[3] S. Clemencon, G. Lugosi, and N. Vayatis, Ranking and scoring using empirical risk minimization, Proceedings of COLT 2005, in LNCS Computational Learning Theory, vol. 3559, pp.1–15, Springer-Verlag, Berlin, Heidelberg.

[4] F. Cucker and D. X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, 2007.

[5] D. Erdogmus and J. C. Principe, An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems, IEEE Trans. Signal Process. **50** (2002), 1780-1786.

[6] D. Erdogmus and J. C. Principe, Convergence properties and data efficiency of the minimum error entropy criterion in adaline training, IEEE Trans. Signal Process. **51** (2003), 1966-1978.

[7] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Amer. Stat. Assoc. **58** (1963), 13–30.

[8] E. Parzen, On the estimation of a probability density function and the mode, Ann. Math. Stat. **33** (1962), 1049-1051.

[9] J. C. Principe, Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives, Springer, New York, 2010.

[10] L. M. Silva, J. M. de Sá, and L. A. Alexandre, The MEE principle in data classification: a perceptrop-based analysis, Neural Comput. **22** (2010), 2698–2728.

[11] S. Smale and D. X. Zhou, Estimating the approximation error in learning theory, Anal. Appl. **1** (2003) 17–41.

[12] S. Smale and D.X. Zhou, Online learning with Markov sampling, Anal. Appl. **7** (2009) 87–113.

[13] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, 1998.

[14] Y. Yao, On complexity issue of online learning algorithms, IEEE Trans. Inform. Theory **56** (2010) 6470–6481.

[15] Y. Ying, Convergence analysis of online algorithms, Adv. Comput. Math. **27** (2007), 273–291.

[16] D. X. Zhou, The covering number in learning theory, J. Complexity **18** (2002), 739–767.

[17] D. X. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory **49** (2003), 1743-1752.