Journal of Machine Learning Research 14 (2013) 377-397 Submitted 2/12; Revised 8/12; Published 2/13

Learning Theory Approach to Minimum Error Entropy Criterion

Ting Hu TINGHU@WHU.EDU.CN
School of Mathematics and Statistics
Wuhan University
Wuhan 430072, China

Jun Fan JUNFAN2@STUDENT.CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
83 Tat Chee Avenue
Kowloon, Hong Kong, China

Qiang Wu QWU@MTSU.EDU
Department of Mathematical Sciences
Middle Tennessee State University
Murfreesboro, TN 37132, USA

Ding-Xuan Zhou MAZHOU@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
83 Tat Chee Avenue
Kowloon, Hong Kong, China

Editor: Gabor Lugosi

Abstract

We consider the minimum error entropy (MEE) criterion and an empirical risk minimization learning algorithm when an approximation of Rényi's entropy (of order 2) by Parzen windowing is minimized. This learning algorithm involves a Parzen windowing scaling parameter. We present a learning theory approach for this MEE algorithm in a regression setting when the scaling parameter is large. Consistency and explicit convergence rates are provided in terms of the approximation ability and capacity of the involved hypothesis space. Novel analysis is carried out for the generalization error associated with Rényi's entropy and a Parzen windowing function, to overcome technical difficulties arising from the essential differences between the classical least squares problems and the MEE setting. An involved symmetrized least squares error is introduced and analyzed, which is related to some ranking algorithms.

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

1. Introduction

Information theoretical learning is inspired by introducing information theory into a machine learning paradigm. Within this framework algorithms have been developed for several learning tasks, including regression, classification, and unsupervised learning. It attracts more and more attention because of its successful applications in signal processing, system engineering, and data mining.

A systematic treatment and recent development of this area can be found in Principe (2010) and references therein.

Minimum error entropy (MEE) is a principle of information theoretical learning and provides a family of supervised learning algorithms. It was introduced for adaptive system training in Erdogmus and Principe (2002) and has been applied to blind source separation, maximally informative subspace projections, clustering, feature selection, blind deconvolution, and some other topics (Erdogmus and Principe, 2003; Principe, 2010; Silva et al., 2010). The idea of MEE is to extract from data as much information as possible about the data generating systems by minimizing error entropies in various ways. In information theory, entropies are used to measure average information quantitatively. For a random variable E with probability density function p_E, Shannon's entropy of E is defined as

$$H_S(E) = -\mathbb{E}[\log p_E] = -\int p_E(e)\log p_E(e)de$$

while Rényi's entropy of order α ($\alpha > 0$ but $\alpha \neq 1$) is defined as

$$H_{R,\alpha}(E) = \frac{1}{1-\alpha}\log \mathbb{E}[p_E^{\alpha-1}] = \frac{1}{1-\alpha}\log\left(\int (p_E(e))^{\alpha} de\right)$$

satisfying $\lim_{\alpha\to 1} H_{R,\alpha}(E) = H_S(E)$. In supervised learning our target is to predict the response variable Y from the explanatory variable X. Then the random variable E becomes the error variable $E = Y - f(X)$ when a predictor $f(X)$ is used and the MEE principle aims at searching for a predictor $f(X)$ that contains the most information of the response variable by minimizing information entropies of the error variable $E = Y - f(X)$. This principle is a substitution of the classical least squares method when the noise is non-Gaussian. Note that $\mathbb{E}[Y - f(X)]^2 = \int e^2 p_E(e)de$. The least squares method minimizes the variance of the error variable E and is perfect to deal with problems involving Gaussian noise (such as some from linear signal processing). But it only puts the first two moments into consideration, and does not work very well for problems involving heavy tailed non-Gaussian noise. For such problems, MEE might still perform very well in principle since moments of all orders of the error variable are taken into account by entropies. Here we only consider Rényi's entropy of order $\alpha = 2$: $H_R(E) = H_{R,2}(E) = -\log\int (p_E(e))^2 de$. Our analysis does not apply to Rényi's entropy of order $\alpha \neq 2$.

In most real applications, neither the explanatory variable X nor the response variable Y is explicitly known. Instead, in supervised learning, a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ is available which reflects the distribution of the explanatory variable X and the functional relation between X and the response variable Y. With this sample, information entropies of the error variable $E = Y - f(X)$ can be approximated by estimating its probability density function p_E by Parzen (1962) windowing $\widehat{p}_E(e) = \frac{1}{mh}\sum_{i=1}^m G(\frac{(e-e_i)^2}{2h^2})$, where $e_i = y_i - f(x_i)$, $h > 0$ is an MEE scaling parameter, and G is a windowing function. A typical choice for the windowing function $G(t) = \exp\{-t\}$ corresponds to Gaussian windowing. Then approximations of Shannon's entropy and Rényi's entropy of order 2 are given by their empirical versions $-\frac{1}{m}\sum_{i=1}^m \log \widehat{p}_E(e_i)$ and $-\log(\frac{1}{m}\sum_{i=1}^m \widehat{p}_E(e_i))$ as

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

and

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

respectively. The empirical MEE is implemented by minimizing these computable quantities.

Though the MEE principle has been proposed for a decade and MEE algorithms have been shown to be effective in various applications, its theoretical foundation for mathematical error analysis is not well understood yet. There is even no consistency result in the literature. It has been observed in applications that the scaling parameter h should be large enough for MEE algorithms to work well before smaller values are tuned. However, it is well known that the convergence of Parzen windowing requires h to converge to 0. We believe this contradiction imposes difficulty for rigorous mathematical analysis of MEE algorithms. Another technical barrier for mathematical analysis of MEE algorithms for regression is the possibility that the regression function may not be a minimizer of the associated generalization error, as described in detail in Section 3 below. The main contribution of this paper is a consistency result for an MEE algorithm for regression. It does require h to be large and explains the effectiveness of the MEE principle in applications.

In the sequel of this paper, we consider an MEE learning algorithm that minimizes the empirical Rényi's entropy $\widehat{H_R}$ and focus on the regression problem. We will take a learning theory approach and analyze this algorithm in an *empirical risk minimization* (ERM) setting. Assume ρ is a probability measure on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a separable metric space (input space for learning) and $\mathcal{Y} = \mathbb{R}$ (output space). Let ρ_X be its marginal distribution on $\mathcal{X}$ (for the explanatory variable X) and $\rho(\cdot|x)$ be the conditional distribution of Y for given $X = x$. The sample $\mathbf{z}$ is assumed to be drawn from ρ independently and identically distributed. The aim of the regression problem is to predict the conditional mean of Y for given X by learning the regression function defined by

$$f_\rho(x) = \mathbb{E}(Y|X = x) = \int_{\mathcal{X}} y d\rho(y|x), \qquad x \in \mathcal{X}.$$

The minimization of empirical Rényi's entropy cannot be done over all possible measurable functions which would lead to overfitting. A suitable hypothesis space should be chosen appropriately in the ERM setting. The ERM framework for MEE learning is defined as follows. Recall $e_i = y_i - f(x_i)$.

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

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

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) and its applications in signal processing have been described in a vast MEE literature (Principe, 2010; Erdogmus and Principe, 2002, 2003; Silva et al., 2010). For different purposes the MEE scaling parameter h may be chosen to be large or small. It has been observed empirically that the MEE criterion has nice convergence properties when the MEE scaling parameter h is large. The main purpose of this paper is to verify this observation in the ERM setting and show that $f_{\mathbf{z}}$ with a suitable constant adjustment approximates the regression function well with confidence. Note that the requirement of a constant adjustment is natural because any translate $f_{\mathbf{z}} + c$ of a solution $f_{\mathbf{z}}$ to (1) with a constant $c \in \mathbb{R}$ is another solution to (1). So our consistency result for MEE algorithm (1) will be stated in terms of the variance $\mathbf{var}[f_{\mathbf{z}}(X) - f_\rho(X)]$ of the error function $f_{\mathbf{z}} - f_\rho$. Here we use $\mathbf{var}$ to denote the variance of a random variable.

2. Main Results on Consistency and Convergence Rates

Throughout the paper, we assume $h \geq 1$ and that

$$\mathbb{E}[|Y|^q] < \infty \text{ for some } q > 2, \text{ and } f_\rho \in L^\infty_{\rho_X}. \quad \text{Denote } q^* = \min\{q-2, 2\}. \tag{2}$$

We also assume that the windowing function G satisfies

$$G \in C^2[0,\infty),\ G'_+(0) = -1, \text{ and } C_G := \sup_{t\in(0,\infty)} \left\{|(1+t)G'(t)| + |(1+t)G''(t)|\right\} < \infty. \tag{3}$$

The special example $G(t) = \exp\{-t\}$ for the Gaussian windowing satisfies (3).

Consistency analysis for regression algorithms is often carried out in the literature under a decay assumption for Y such as uniform boundedness and exponential decays. A recent study (Audibert and Catoni, 2011) was made under the assumption $\mathbb{E}[|Y|^4] < \infty$. Our assumption (2) is weaker since q may be arbitrarily close to 2. Note that (2) obviously holds when $|Y| \leq M$ almost surely for some constant $M > 0$, in which case we shall denote $q^* = 2$.

Our consistency result, to be proved in Section 5, asserts that when h and m are large enough, the error $\mathbf{var}[f_{\mathbf{z}}(X) - f_\rho(X)]$ of MEE algorithm (1) can be arbitrarily close to the approximation error (Smale and Zhou, 2003) of the hypothesis space $\mathcal{H}$ with respect to the regression function f_ρ.

Definition 2 *The approximation error of the pair $(\mathcal{H}, \rho)$ is defined by*

$$\mathcal{D}_{\mathcal{H}}(f_\rho) = \inf_{f\in\mathcal{H}} \mathbf{var}[f(X) - f_\rho(X)].$$

Theorem 3 *Under assumptions (2) and (3), for any $0 < \varepsilon \leq 1$ and $0 < \delta < 1$, there exist $h_{\varepsilon,\delta} \geq 1$ and $m_{\varepsilon,\delta}(h) \geq 1$ both depending on $\mathcal{H}, G, \rho, \varepsilon, \delta$ such that for $h \geq h_{\varepsilon,\delta}$ and $m \geq m_{\varepsilon,\delta}(h)$, with confidence $1-\delta$, we have*

$$\mathbf{var}[f_{\mathbf{z}}(X) - f_\rho(X)] \leq \mathcal{D}_{\mathcal{H}}(f_\rho) + \varepsilon. \tag{4}$$

Our convergence rates will be stated in terms of the approximation error and the capacity of the hypothesis space $\mathcal{H}$ measured by covering numbers in this paper.

Definition 4 *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 in $C(\mathcal{X})$ with radius ε and centers in $\mathcal{H}$ covering the set $\mathcal{H}$. We shall assume that for some constants $p > 0$ and $A_p > 0$, there holds*

$$\log \mathcal{N}(\mathcal{H}, \varepsilon) \leq A_p \varepsilon^{-p}, \qquad \forall \varepsilon > 0. \tag{5}$$

The behavior (5) of the covering numbers is typical in learning theory. It is satisfied by balls of Sobolev spaces on $\mathcal{X} \subset \mathbb{R}^n$ and reproducing kernel Hilbert spaces associated with Sobolev smooth kernels. See Anthony and Bartlett (1999), Zhou (2002), Zhou (2003) and Yao (2010). We remark that empirical covering numbers might be used together with concentration inequalities to provide shaper error estimates. This is however beyond our scope and for simplicity we adopt the the covering number in $C(\mathcal{X})$ throughout this paper.

The following convergence rates for (1) with large h will be proved in Section 5.

Theorem 5 *Assume (2), (3) and covering number condition (5) for some $p > 0$. Then for any $0 < \eta \leq 1$ and $0 < \delta < 1$, with confidence $1-\delta$ we have*

$$\mathbf{var}[f_{\mathbf{z}}(X) - f_\rho(X)] \leq \widetilde{C}_{\mathcal{H}} \eta^{(2-q)/2} \left(h^{-\min\{q-2,2\}} + h m^{-\frac{1}{1+p}} \right) \log \frac{2}{\delta} + (1+\eta) \mathcal{D}_{\mathcal{H}}(f_\rho). \tag{6}$$

If $|Y| \leq M$ almost surely for some $M > 0$, then with confidence $1-\delta$ we have

$$\mathbf{var}[f_{\mathbf{z}}(X) - f_\rho(X)] \leq \frac{\widetilde{C}_{\mathcal{H}}}{\eta} \left(h^{-2} + m^{-\frac{1}{1+p}} \right) \log \frac{2}{\delta} + (1+\eta) \mathcal{D}_{\mathcal{H}}(f_\rho). \tag{7}$$

Here $\widetilde{C}_{\mathcal{H}}$ is a constant independent of m, δ, η or h (depending on $\mathcal{H}, G, \rho$ given explicitly in the proof).

Remark 6 *In Theorem 5, we use a parameter $\eta > 0$ in error bounds (6) and (7) 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.*

If moment condition (2) with $q \geq 4$ is satisfied and $\eta = 1$, then by taking $h = m^{\frac{1}{3(1+p)}}$, (6) becomes

$$\mathbf{var}[(f_{\mathbf{z}}(X) - f_\rho(X)] \leq 2\widetilde{C}_{\mathcal{H}} \left(\frac{1}{m} \right)^{\frac{2}{3(1+p)}} \log \frac{2}{\delta} + 2\mathcal{D}_{\mathcal{H}}(f_\rho). \tag{8}$$

If $|Y| \leq M$ almost surely, then by taking $h = m^{\frac{1}{2(1+p)}}$ and $\eta = 1$, error bound (7) becomes

$$\mathbf{var}[f_{\mathbf{z}}(X) - f_\rho(X)] \leq 2\widetilde{C}_{\mathcal{H}} m^{-\frac{1}{1+p}} \log \frac{2}{\delta} + 2\mathcal{D}_{\mathcal{H}}(f_\rho). \tag{9}$$

Remark 7 *When the index p in covering number condition (5) is small enough (the case when $\mathcal{H}$ is a finite ball of a reproducing kernel Hilbert space with a smooth kernel), we see that the power indices for the sample error terms of convergence rates (8) and (9) can be arbitrarily close to $2/3$ and 1, respectively. There is a gap in the rates between the case of (2) with large q and the uniform bounded case. This gap is caused by the Parzen windowing process for which our method does not lead to better estimates when $q > 4$. It would be interesting to know whether the gap can be narrowed.*

Note the result in Theorem 5 does not guarantee that $f_{\mathbf{z}}$ itself approximates f_ρ well when the bounds are small. Instead a constant adjustment is required. Theoretically the best constant is $\mathbb{E}[f_{\mathbf{z}}(X) - f_\rho(X)]$. In practice it is usually approximated by the sample mean $\frac{1}{m}\sum_{i=1}^m (f_{\mathbf{z}}(x_i) - y_i)$ in the case of uniformly bounded noise and the approximation can be easily handled. To deal with heavy tailed noise, we project the output values onto the closed interval $[-\sqrt{m}, \sqrt{m}]$ by the projection $\pi_{\sqrt{m}} : \mathbb{R} \to \mathbb{R}$ defined by

$$\pi_{\sqrt{m}}(y) = \begin{cases} y, & \text{if } y \in [-\sqrt{m}, \sqrt{m}], \\ \sqrt{m}, & \text{if } y > \sqrt{m}, \\ -\sqrt{m}, & \text{if } y < -\sqrt{m}, \end{cases}$$

and then approximate $\mathbb{E}[f_{\mathbf{z}}(X) - f_\rho(X)]$ by the computable quantity

$$\frac{1}{m} \sum_{i=1}^m \left[f_{\mathbf{z}}(x_i) - \pi_{\sqrt{m}}(y_i) \right]. \tag{10}$$

The following quantitative result, to be proved in Section 5, tells us that this is a good approximation.

Theorem 8 *Assume $\mathbb{E}[|Y|^2] < \infty$ and covering number condition (5) for some $p > 0$. Then for any $0 < \delta < 1$, with confidence $1-\delta$ we have*

$$\sup_{f\in\mathcal{H}} \left| \frac{1}{m}\sum_{i=1}^{m} \left[f(x_i) - \pi_{\sqrt{m}}(y_i)\right] - \mathbb{E}[f(X) - f_\rho(X)] \right| \le \widetilde{C}'_{\mathcal{H}} m^{-\frac{1}{2+p}} \log\frac{2}{\delta} \tag{11}$$

which implies in particular that

$$\left| \frac{1}{m}\sum_{i=1}^{m} \left[f_{\mathbf{z}}(x_i) - \pi_{\sqrt{m}}(y_i)\right] - \mathbb{E}[f_{\mathbf{z}}(X) - f_\rho(X)] \right| \le \widetilde{C}'_{\mathcal{H}} m^{-\frac{1}{2+p}} \log\frac{2}{\delta}, \tag{12}$$

where $\widetilde{C}'_{\mathcal{H}}$ is the constant given by

$$\widetilde{C}'_{\mathcal{H}} = 7\sup_{f\in\mathcal{H}} \|f\|_\infty + 4 + 7\sqrt{\mathbb{E}[|Y|^2]} + \mathbb{E}[|Y|^2] + A_p^{\frac{1}{2+p}}.$$

Replacing the mean $\mathbb{E}[f_{\mathbf{z}}(X) - f_\rho(X)]$ by the quantity (10), we define an estimator of f_ρ as

$$\widetilde{f}_{\mathbf{z}} = f_{\mathbf{z}} - \frac{1}{m}\sum_{i=1}^{m} \left[f_{\mathbf{z}}(x_i) - \pi_{\sqrt{m}}(y_i)\right].$$

Putting (12) and the bounds from Theorem 5 into the obvious error expression

$$\left\|\widetilde{f}_{\mathbf{z}} - f_\rho\right\|_{L^2_{\rho_X}} \le \left| \frac{1}{m}\sum_{i=1}^{m} \left[f_{\mathbf{z}}(x_i) - \pi_{\sqrt{m}}(y_i)\right] - \mathbb{E}[f_{\mathbf{z}}(X) - f_\rho(X)] \right| + \sqrt{\mathbf{var}[(f_{\mathbf{z}}(X) - f_\rho(X)]}, \tag{13}$$

we see that $\widetilde{f}_{\mathbf{z}}$ is a good estimator of f_ρ: the power index $\frac{1}{2+p}$ in (12) is greater than $\frac{1}{2(1+p)}$, the power index appearing in the last term of (13) when the variance term is bounded by (9), even in the uniformly bounded case.

To interpret our main results better we present a corollary and an example below.

If there is a constant c_ρ such that $f_\rho + c_\rho \in \mathcal{H}$, we have $\mathcal{D}_{\mathcal{H}}(f_\rho) = 0$. In this case, the choice $\eta = 1$ in Theorem 5 yields the following learning rate. Note that (2) implies $\mathbb{E}[|Y|^2] < \infty$.

Corollary 9 *Assume (5) with some $p > 0$ and $f_\rho + c_\rho \in \mathcal{H}$ for some constant $c_\rho \in \mathbb{R}$. Under conditions (2) and (3), by taking $h = m^{\frac{1}{(1+p)\min\{q-1,3\}}}$, we have with confidence $1-\delta$,*

$$\left\|\widetilde{f}_{\mathbf{z}} - f_\rho\right\|_{L^2_{\rho_X}} \le \left(\widetilde{C}'_{\mathcal{H}} + \sqrt{2\widetilde{C}_{\mathcal{H}}}\right) m^{-\frac{\min\{q-2,2\}}{2(1+p)\min\{q-1,3\}}} \log\frac{2}{\delta}.$$

If $|Y| \le M$ almost surely, then by taking $h = m^{\frac{1}{2(1+p)}}$, we have with confidence $1-\delta$,

$$\left\|\widetilde{f}_{\mathbf{z}} - f_\rho\right\|_{L^2_{\rho_X}} \le \left(\widetilde{C}'_{\mathcal{H}} + \sqrt{2\widetilde{C}_{\mathcal{H}}}\right) m^{-\frac{1}{2(1+p)}} \log\frac{2}{\delta}.$$

This corollary states that $\widetilde{f}_{\mathbf{z}}$ can approximate the regression function very well. Note, however, this happens when the hypothesis space is chosen appropriately and the parameter h tends to infinity.

A special example of the hypothesis space is a ball of a Sobolev space $H^s(\mathcal{X})$ with index $s > \frac{n}{2}$ on a domain $\mathcal{X} \subset \mathbb{R}^n$ which satisfies (5) 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 convergence rate can be arbitrarily close to $\frac{1}{3}$ when $\mathbb{E}[|Y|^4] < \infty$, and $\frac{1}{2}$ when $|Y| \le M$ almost surely.

Example 1 *Let $\mathcal{X}$ be a bounded domain of $\mathbb{R}^n$ with Lipschitz boundary. Assume $f_\rho \in H^s(X)$ for some $s > \frac{n}{2}$ and take $\mathcal{H} = \{f \in H^s(X) : \|f\|_{H^s(X)} \le R\}$ with $R \ge \|f_\rho\|_{H^s(X)}$ and $R \ge 1$. If $\mathbb{E}[|Y|^4] < \infty$, then by taking $h = m^{\frac{1}{3(1+n/s)}}$, we have with confidence $1-\delta$,*

$$\left\| \widetilde{f}_{\mathbf{z}} - f_\rho \right\|_{L^2_{\rho_X}} \le C_{s,n,\rho} R^{\frac{n}{2(s+n)}} m^{-\frac{1}{3(1+n/s)}} \log \frac{2}{\delta}.$$

If $|Y| \le M$ almost surely, then by taking $h = m^{\frac{1}{2(1+n/s)}}$, with confidence $1-\delta$,

$$\left\| \widetilde{f}_{\mathbf{z}} - f_\rho \right\|_{L^2_{\rho_X}} \le C_{s,n,\rho} R^{\frac{n}{2(s+n)}} m^{-\frac{1}{2+2n/s}} \log \frac{2}{\delta}.$$

Here the constant $C_{s,n,\rho}$ is independent of R.

Compared to the analysis of least squares methods, our consistency results for the MEE algorithm require a weaker condition by allowing heavy tailed noise, while the convergence rates are comparable but slightly worse than the optimal one $O(m^{-\frac{1}{2+n/s}})$. Further investigation of error analysis for the MEE algorithm is required to achieve the optimal rate, which is beyond the scope of this paper.

3. Technical Difficulties in MEE and Novelties

The MEE algorithm (1) involving sample pairs like quadratic forms is different from most classical ERM learning algorithms (Vapnik, 1998; Anthony and Bartlett, 1999) constructed by sums of independent random variables. But as done for some ranking algorithms (Agarwal and Niyogi, 2009; Clemencon et al., 2005), one can still follow the same line to define a functional called generalization error or *information error* (related to information potential defined on page 88 of Principe, 2010) associated with the windowing function G over the space of measurable functions on $\mathcal{X}$ as

$$\mathcal{E}^{(h)}(f) = \int_{\mathcal{Z}} \int_{\mathcal{Z}} -h^2 G\left(\frac{[(y-f(x)) - (y'-f(x'))]^2}{2h^2} \right) d\rho(x,y) d\rho(x',y').$$

An essential barrier for our consistency analysis is an observation made by numerical simulations (Erdogmus and Principe, 2003; Silva et al., 2010) and verified mathematically for Shannon's entropy in Chen and Principe (2012) that the regression function f_ρ may not be a minimizer of $\mathcal{E}^{(h)}$. It is totally different from the classical least squares generalization error $\mathcal{E}^{ls}(f) = \int_{\mathcal{Z}} (f(x)-y)^2 d\rho$ which satisfies a nice identity $\mathcal{E}^{ls}(f) - \mathcal{E}^{ls}(f_\rho) = \|f - f_\rho\|^2_{L^2_{\rho_X}} \ge 0$. This barrier leads to three technical difficulties in our error analysis which will be overcome by our novel approaches making full use of the special feature that the MEE scaling parameter h is large in this paper.

3.1 Approximation of Information Error

The first technical difficulty we meet in our mathematical analysis for MEE algorithm (1) is the varying form depending on the windowing function G. Our novel approach here is an approximation of the information error in terms of the variance $\mathbf{var}[f(X) - f_\rho(X)]$ when h is large. This is achieved by showing that $\mathcal{E}^{(h)}$ is closely related to the following *symmetrized least squares error* which has appeared in the literature of ranking algorithms (Clemencon et al., 2005; Agarwal and Niyogi, 2009).

Definition 10 *The symmetrized least squares error is defined on the space $L^2_{\rho_X}$ by*

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

To give the approximation of $\mathcal{E}^{(h)}$, we need a simpler form of $\mathcal{E}^{sls}$.

Lemma 11 *If $\mathbb{E}[Y^2] < \infty$, then by denoting $C_\rho = \int_Z \left[y - f_\rho(x)\right]^2 d\rho$, we have*

$$\mathcal{E}^{sls}(f) = 2\mathbf{var}[f(X) - f_\rho(X)] + 2C_\rho, \qquad \forall f \in L^2_{\rho_X}.$$

Proof Recall that for two independent and identically distributed samples ξ and ξ' of a random variable, one has the identity

$$\mathbb{E}[(\xi - \xi')^2] = 2[\mathbb{E}(\xi - \mathbb{E}\xi)^2] = 2\mathbf{var}(\xi).$$

Then we have

$$\mathcal{E}^{sls}(f) = \mathbb{E}\left[\Big((y - f(x)) - \big(y' - f(x')\big)\Big)^2\right] = 2\mathbf{var}[Y - f(X)].$$

By the definition $\mathbb{E}[Y|X] = f_\rho(X)$, it is easy to see that $C_\rho = \mathbf{var}(Y - f_\rho(X))$ and the covariance between $Y - f_\rho(X)$ and $f_\rho(X) - f(X)$ vanishes. So $\mathbf{var}[Y - f(X)] = \mathbf{var}[Y - f_\rho(X)] + \mathbf{var}[f(X) - f_\rho(X)]$. This proves the desired identity. ■

We are in a position to present the approximation of $\mathcal{E}^{(h)}$ for which a large scaling parameter h plays an important role. Since $\mathcal{H}$ is a compact subset of $C(X)$, we know that the number $\sup_{f\in\mathcal{H}} \|f\|_\infty$ is finite.

Lemma 12 *Under assumptions (2) and (3), for any essentially bounded measurable function f on X, we have*

$$\left|\mathcal{E}^{(h)}(f) + h^2 G(0) - C_\rho - \mathbf{var}[f(X) - f_\rho(X)]\right| \le 5 \cdot 2^7 C_G \left(\left(\mathbb{E}[|Y|^q]\right)^{\frac{q^*+2}{q}} + \|f\|_\infty^{q^*+2} \right) h^{-q^*}.$$

In particular,

$$\left|\mathcal{E}^{(h)}(f) + h^2 G(0) - C_\rho - \mathbf{var}[f(X) - f_\rho(X)]\right| \le C'_{\mathcal{H}} h^{-q^*}, \qquad \forall f \in \mathcal{H},$$

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

$$C'_{\mathcal{H}} = 5 \cdot 2^7 C_G \left(\left(\mathbb{E}[|Y|^q]\right)^{(q^*+2)/q} + \left(\sup_{f\in\mathcal{H}} \|f\|_\infty \right)^{q^*+2} \right).$$

Proof Observe that $q^*+2=\min\{q,4\}\in(2,4]$. By the Taylor expansion and the mean value theorem, we have

$$|G(t)-G(0)-G'_+(0)t|\le\begin{cases}\frac{\|G''\|_\infty}{2}t^2\le\frac{\|G''\|_\infty}{2}t^{(q^*+2)/2}, & \text{if } 0\le t\le 1,\\ 2\|G'\|_\infty t\le 2\|G'\|_\infty t^{(q^*+2)/2}, & \text{if } t>1.\end{cases}$$

So $|G(t)-G(0)-G'_+(0)t|\le\left(\frac{\|G''\|_\infty}{2}+2\|G'\|_\infty\right)t^{(q^*+2)/2}$ for all $t\ge 0$, and by setting $t=\frac{[(y-f(x))-(y'-f(x'))]^2}{2h^2}$, we know that

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

This together with Lemma 11, the normalization assumption $G'_+(0)=-1$ and Hölder's inequality applied when $q>4$ proves the desired bound and hence our conclusion. ■

Applying Lemma 12 to a function $f\in\mathcal{H}$ and $f_\rho\in L^\infty_{\rho_X}$ yields the following fact on the excess generalization error $\mathcal{E}^{(h)}(f)-\mathcal{E}^{(h)}(f_\rho)$.

Theorem 13 *Under assumptions (2) and (3), we have*

$$\left|\mathcal{E}^{(h)}(f)-\mathcal{E}^{(h)}(f_\rho)-\mathbf{var}[f(X)-f_\rho(X)]\right|\le C''_{\mathcal{H}}h^{-q^*},\qquad\forall f\in\mathcal{H},$$

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

$$C''_{\mathcal{H}}=5\cdot 2^8C_G\left((\mathbb{E}[|Y|^q])^{(q^*+2)/q}+\left(\sup_{f\in\mathcal{H}}\|f\|_\infty\right)^{q^*+2}+\|f_\rho\|_\infty^{q^*+2}\right).$$

3.2 Functional Minimizer and Best Approximation

As f_ρ may not be a minimizer of $\mathcal{E}^{(h)}$, the second technical difficulty in our error analysis is the diversity of two ways to define a *target function* in $\mathcal{H}$, one to minimize the information error and the other to minimize the variance $\mathbf{var}[f(X)-f_\rho(X)]$. These possible candidates for the target function are defined as

$$\begin{aligned}f_{\mathcal{H}}&:=\arg\min_{f\in\mathcal{H}}\mathcal{E}^{(h)}(f),\\ f_{approx}&:=\arg\min_{f\in\mathcal{H}}\mathbf{var}[f(X)-f_\rho(X)].\end{aligned}$$

Our novelty to overcome the technical difficulty is to show that when the MEE scaling parameter h is large, these two functions are actually very close.

Theorem 14 *Under assumptions (2) and (3), we have*

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

and

$$\mathbf{var}[f_{\mathcal{H}}(X) - f_\rho(X)] \le \mathbf{var}[f_{approx}(X) - f_\rho(X)] + 2C''_{\mathcal{H}} h^{-q^*}.$$

Proof By Theorem 13 and the definitions of $f_{\mathcal{H}}$ and f_{approx}, we have

$$\begin{aligned}
&\mathcal{E}^{(h)}(f_{\mathcal{H}}) - \mathcal{E}^{(h)}(f_\rho) \le \mathcal{E}^{(h)}(f_{approx}) - \mathcal{E}^{(h)}(f_\rho) \le \mathbf{var}[f_{approx}(X) - f_\rho(X)] + C''_{\mathcal{H}} h^{-q^*} \\
&\le \mathbf{var}[f_{\mathcal{H}}(X) - f_\rho(X)] + C''_{\mathcal{H}} h^{-q^*} \le \mathcal{E}^{(h)}(f_{\mathcal{H}}) - \mathcal{E}^{(h)}(f_\rho) + 2C''_{\mathcal{H}} h^{-q^*} \\
&\le \mathbf{var}[f_{approx}(X) - f_\rho(X)] + 3C''_{\mathcal{H}} h^{-q^*}.
\end{aligned}$$

Then the desired inequalities follow. ■

Moreover, Theorem 13 yields the following error decomposition for our algorithm.

Lemma 15 *Under assumptions (2) and (3), we have*

$$\mathbf{var}[f_{\mathbf{z}}(X) - f_\rho(X)] \le \left\{ \mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}}) \right\} + \mathbf{var}[f_{approx}(X) - f_\rho(X)] + 2C''_{\mathcal{H}} h^{-q^*}. \tag{14}$$

Proof By Theorem 13,

$$\begin{aligned}
\mathbf{var}[f_{\mathbf{z}}(X) - f_\rho(X)] &\le \mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_\rho) + C''_{\mathcal{H}} h^{-q^*} \\
&\le \left\{ \mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}}) \right\} + \mathcal{E}^{(h)}(f_{\mathcal{H}}) - \mathcal{E}^{(h)}(f_\rho) + C''_{\mathcal{H}} h^{-q^*}.
\end{aligned}$$

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

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

Applying Theorem 13 to the above bound implies

$$\mathbf{var}[f_{\mathbf{z}}(X) - f_\rho(X)] \le \left\{ \mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}}) \right\} + \mathbf{var}[f_{approx}(X) - f_\rho(X)] + 2C''_{\mathcal{H}} h^{-q^*}.$$

Then desired error decomposition (14) follows. ■

Error decomposition has been a standard technique to analyze least squares ERM regression algorithms (Anthony and Bartlett, 1999; Cucker and Zhou, 2007; Smale and Zhou, 2009; Ying, 2007). In error decomposition (14) for MEE learning algorithm (1), the first term on the right side is the sample error, the second term $\mathbf{var}[f_{approx}(X) - f_\rho(X)]$ is the approximation error, while the last extra term $2C''_{\mathcal{H}} h^{-q^*}$ is caused by the Parzen windowing and is small when h is large. The quantity $\mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}})$ of the sample error term will be bounded in the following discussion.

3.3 Error Decomposition by U-statistics and Special Properties

We shall decompose the sample error term $\mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}})$ further by means of U-statistics defined for $f \in \mathcal{H}$ and the sample $\mathbf{z}$ as

$$V_f(\mathbf{z}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i} U_f(z_i, z_j),$$

where U_f is a kernel given with $z = (x,y), z' = (x',y') \in \mathcal{Z}$ by

$$U_f(z,z') = -h^2 G\left(\frac{[(y-f(x))-(y'-f(x'))]^2}{2h^2}\right) + h^2 G\left(\frac{\left[\left(y-f_\rho(x)\right)-\left(y'-f_\rho(x')\right)\right]^2}{2h^2}\right). \quad (15)$$

It is easy to see that $\mathbb{E}[V_f] = \mathcal{E}^{(h)}(f) - \mathcal{E}^{(h)}(f_\rho)$ and $U_f(z,z) = 0$. Then

$$\mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}}) = \mathbb{E}\left[V_{f_{\mathbf{z}}}\right] - \mathbb{E}\left[V_{f_{\mathcal{H}}}\right] = \mathbb{E}\left[V_{f_{\mathbf{z}}}\right] - V_{f_{\mathbf{z}}} + V_{f_{\mathbf{z}}} - V_{f_{\mathcal{H}}} + V_{f_{\mathcal{H}}} - \mathbb{E}\left[V_{f_{\mathcal{H}}}\right].$$

By the definition of $f_{\mathbf{z}}$, we have $V_{f_{\mathbf{z}}} - V_{f_{\mathcal{H}}} \leq 0$. Hence

$$\mathcal{E}^{(h)}(f_{\mathbf{z}}) - \mathcal{E}^{(h)}(f_{\mathcal{H}}) \leq \mathbb{E}\left[V_{f_{\mathbf{z}}}\right] - V_{f_{\mathbf{z}}} + V_{f_{\mathcal{H}}} - \mathbb{E}\left[V_{f_{\mathcal{H}}}\right]. \quad (16)$$

The above bound will be estimated by a uniform ratio probability inequality. A technical difficulty we meet here is the possibility that $\mathbb{E}[V_f] = \mathcal{E}^{(h)}(f) - \mathcal{E}^{(h)}(f_\rho)$ might be negative since f_ρ may not be a minimizer of $\mathcal{E}^{(h)}$. It is overcome by the following novel observation which is an immediate consequence of Theorem 13.

Lemma 16 *Under assumptions (2) and (3), if* $\varepsilon \geq C''_{\mathcal{H}} h^{-q^*}$, *then*

$$\mathbb{E}[V_f] + 2\varepsilon \geq \mathbb{E}[V_f] + C''_{\mathcal{H}} h^{-q^*} + \varepsilon \geq \mathbf{var}[f(X) - f_\rho(X)] + \varepsilon \geq \varepsilon, \qquad \forall f \in \mathcal{H}. \quad (17)$$

4. Sample Error Estimates

In this section, we follow (16) and estimate the sample error by a uniform ratio probability inequality based on the following Hoeffding's probability inequality for U-statistics (Hoeffding, 1963).

Lemma 17 *If U is a symmetric real-valued function on $\mathcal{Z} \times \mathcal{Z}$ satisfying $a \leq U(z,z') \leq b$ almost surely and* $\mathrm{var}[U] = \sigma^2$, *then for any* $\varepsilon > 0$,

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

To apply Lemma 17 we need to bound σ^2 and $b-a$ for the kernel U_f defined by (15). Our novelty for getting sharp bounds is to use a Taylor expansion involving a C^2 function $\widetilde{G}$ on $\mathbb{R}$:

$$\widetilde{G}(w) = \widetilde{G}(0) + \widetilde{G}'(0)w + \int_0^w (w-t)\widetilde{G}''(t)dt, \qquad \forall w \in \mathbb{R}. \quad (18)$$

Denote a constant $A_{\mathcal{H}}$ depending on ρ, G, q and $\mathcal{H}$ as

$$A_{\mathcal{H}} = 9 \cdot 2^8 C_G^2 \sup_{f \in \mathcal{H}} \|f - f_\rho\|_\infty^{\frac{4}{q}} \left((\mathbb{E}[|Y|^q])^{\frac{2}{q}} + \|f_\rho\|_\infty^2 + \sup_{f \in \mathcal{H}} \|f - f_\rho\|_\infty^2 \right).$$

Lemma 18 *Assume (2) and (3).*

(a) For any $f, g \in \mathcal{H}$, we have

$$\left|U_f\right| \le 4C_G\|f-f_\rho\|_\infty h \quad \text{and} \quad \left|U_f-U_g\right| \le 4C_G\|f-g\|_\infty h$$

and

$$\mathrm{var}[U_f] \le A_{\mathcal{H}}\left(\mathrm{var}[f(X)-f_\rho(X)]\right)^{(q-2)/q}.$$

(b) If $|Y| \le M$ almost surely for some constant $M > 0$, then we have almost surely

$$\left|U_f\right| \le A'_{\mathcal{H}}\left|(f(x)-f_\rho(x))-(f(x')-f_\rho(x'))\right|, \quad \forall f \in \mathcal{H} \tag{19}$$

and

$$\left|U_f-U_g\right| \le A'_{\mathcal{H}}\left|(f(x)-g(x))-(f(x')-g(x'))\right|, \quad \forall f,g \in \mathcal{H}, \tag{20}$$

where $A'_{\mathcal{H}}$ is a constant depending on ρ, G and $\mathcal{H}$ given by

$$A'_{\mathcal{H}} = 36C_G\left(M+\sup_{f\in\mathcal{H}}\|f\|_\infty\right).$$

Proof Define a function $\widetilde{G}$ on $\mathbb{R}$ by

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

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

$$U_f(z,z') = -h^2\widetilde{G}\left(\frac{(y-f(x))-(y'-f(x'))}{h}\right)+h^2\widetilde{G}\left(\frac{\big(y-f_\rho(x)\big)-\big(y'-f_\rho(x')\big)}{h}\right).$$

(a) We apply the mean value theorem and see that $|U_f(z,z')| \le 2h\|\widetilde{G}'\|_\infty\|f-f_\rho\|_\infty$. The inequality for $|U_f-U_g|$ is obtained when f_ρ is replaced by g. Note that $\|\widetilde{G}'\|_\infty = \|tG'(t^2/2)\|_\infty$. Then the bounds for U_f and U_f-U_g are verified by noting $\|tG'(t^2/2)\|_\infty \le 2C_G$.

To bound the variance, we apply (18) to the two points $w_1 = \frac{(y-f(x))-(y'-f(x'))}{h}$ and $w_2 = \frac{\big(y-f_\rho(x)\big)-\big(y'-f_\rho(x')\big)}{h}$. Writing w_2-t as $w_2-w_1+w_1-t$, we see from $\widetilde{G}'(0)=0$ that

$$\begin{aligned} U_f(z,z') &= h^2\left(\widetilde{G}(w_2)-\widetilde{G}(w_1)\right) = h^2\widetilde{G}'(0)(w_2-w_1) \\ &\quad +h^2\int_0^{w_2}(w_2-t)\widetilde{G}''(t)dt - h^2\int_0^{w_1}(w_1-t)\widetilde{G}''(t)dt \\ &= h^2\int_0^{w_2}(w_2-w_1)\widetilde{G}''(t)dt + h^2\int_{w_1}^{w_2}(w_1-t)\widetilde{G}''(t)dt. \end{aligned}$$

It follows that

$$\begin{aligned} \left|U_f(z,z')\right| &\le \|\widetilde{G}''\|_\infty\left|\big(y-f_\rho(x)\big)-\big(y'-f_\rho(x')\big)\right|\ \left|\big(f(x)-f_\rho(x)\big)-\big(f(x')-f_\rho(x')\big)\right| \\ &\quad +\|\widetilde{G}''\|_\infty\left|\big(f(x)-f_\rho(x)\big)-\big(f(x')-f_\rho(x')\big)\right|^2. \end{aligned} \tag{21}$$

Since $\mathbb{E}[|Y|^q] < \infty$, we apply Hölder's inequality and see that

$$\begin{aligned}
&\int_{\mathcal{Z}}\int_{\mathcal{Z}} \left|\big(y-f_\rho(x)\big)-\big(y'-f_\rho(x')\big)\right|^2 \left|\big(f(x)-f_\rho(x)\big)-\big(f(x')-f_\rho(x')\big)\right|^2 d\rho(z)d\rho(z') \\
&\le \left\{\int_{\mathcal{Z}}\int_{\mathcal{Z}} \left|\big(y-f_\rho(x)\big)-\big(y'-f_\rho(x')\big)\right|^q d\rho(z)d\rho(z')\right\}^{2/q} \\
&\qquad \left\{\int_{\mathcal{Z}}\int_{\mathcal{Z}} \left|\big(f(x)-f_\rho(x)\big)-\big(f(x')-f_\rho(x')\big)\right|^{2q/(q-2)} d\rho(z)d\rho(z')\right\}^{1-2/q} \\
&\le \left\{4^{q+1}(\mathbb{E}[|Y|^q]+\|f_\rho\|_\infty^q)\right\}^{2/q} \left\{\|f-f_\rho\|_\infty^{4/(q-2)} 2\mathrm{var}[f(X)-f_\rho(X)]\right\}^{(q-2)/q}.
\end{aligned}$$

Here we have separated the power index $2q/(q-2)$ into the sum of $4/(q-2)$ and 2. Then

$$\begin{aligned}
\mathrm{var}[U_f] \quad &\le \quad \mathbb{E}[U_f^2] \le 2\|\widetilde{G}''\|_\infty^2 2^{\frac{5q+3}{q}} (\mathbb{E}[|Y|^q]+\|f_\rho\|_\infty^q)^{\frac{2}{q}} \|f-f_\rho\|_\infty^{\frac{4}{q}} \left(\mathrm{var}[f(X)-f_\rho(X)]\right)^{\frac{q-2}{q}} \\
&\quad +2\|\widetilde{G}''\|_\infty^2 4\|f-f_\rho\|_\infty^2 2\mathrm{var}[f(X)-f_\rho(X)].
\end{aligned}$$

Hence the desired inequality holds true since $\|\widetilde{G}''\|_\infty \le \|G'\|_\infty + \|t^2 G''(t^2/2)\|_\infty \le 3C_G$ and $\mathrm{var}[f(X)-f_\rho(X)] \le \|f-f_\rho\|_\infty^2$.

(b) If $|Y| \le M$ almost surely for some constant $M > 0$, then we see from (21) that almost surely $\left|U_f(z,z')\right| \le 4\|\widetilde{G}''\|_\infty (M+\|f_\rho\|_\infty+\|f-f_\rho\|_\infty)\left|\big(f(x)-f_\rho(x)\big)-\big(f(x')-f_\rho(x')\big)\right|$. Hence (19) holds true almost surely. Replacing f_ρ by g in (21), we see immediately inequality (20). The proof of Lemma 18 is complete. ■

With the above preparation, we can now give the uniform ratio probability inequality for U-statistics to estimate the sample error, following methods in the learning theory literature (Haussler et al., 1994; Koltchinskii, 2006; Cucker and Zhou, 2007).

Lemma 19 *Assume (2), (3) and $\varepsilon \ge C''_{\mathcal{H}} h^{-q^*}$. Then we have*

$$\mathrm{Prob}\left\{\sup_{f\in\mathcal{H}} \frac{\left|V_f-\mathbb{E}[V_f]\right|}{(\mathbb{E}[V_f]+2\varepsilon)^{(q-2)/q}} > 4\varepsilon^{2/q}\right\} \le 2\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{4C_G h}\right) \exp\left\{-\frac{(m-1)\varepsilon}{A''_{\mathcal{H}} h}\right\},$$

where $A''_{\mathcal{H}}$ is the constant given by

$$A''_{\mathcal{H}} = 4A_{\mathcal{H}}(C''_{\mathcal{H}})^{-2/q} + 12C_G \sup_{f\in\mathcal{H}} \|f-f_\rho\|_\infty.$$

If $|Y| \le M$ almost surely for some constant $M > 0$, then we have

$$\mathrm{Prob}\left\{\sup_{f\in\mathcal{H}} \frac{\left|V_f-\mathbb{E}[V_f]\right|}{\sqrt{\mathbb{E}[V_f]+2\varepsilon}} > 4\sqrt{\varepsilon}\right\} \le 2\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{2A'_{\mathcal{H}}}\right) \exp\left\{-\frac{(m-1)\varepsilon}{A''_{\mathcal{H}}}\right\},$$

where $A''_{\mathcal{H}}$ is the constant given by

$$A''_{\mathcal{H}} = 8A'_{\mathcal{H}} + 6A'_{\mathcal{H}} \sup_{f\in\mathcal{H}} \|f-f_\rho\|_\infty.$$

Proof If $\|f-f_j\|_\infty \le \frac{\varepsilon}{4C_G h}$, Lemma 18 (a) implies $|\mathbb{E}[V_f]-\mathbb{E}[V_{f_j}]| \le \varepsilon$ and $|V_f - V_{f_j}| \le \varepsilon$ almost surely. These in connection with Lemma 16 tell us that

$$\frac{\big|V_f-\mathbb{E}[V_f]\big|}{(\mathbb{E}[V_f]+2\varepsilon)^{(q-2)/q}} > 4\varepsilon^{2/q} \quad \Longrightarrow \quad \frac{\big|V_{f_j}-\mathbb{E}[V_{f_j}]\big|}{(\mathbb{E}[V_{f_j}]+2\varepsilon)^{(q-2)/q}} > \varepsilon^{2/q}.$$

Thus by taking $\{f_j\}_{j=1}^N$ to be an $\frac{\varepsilon}{4C_G h}$ net of the set $\mathcal{H}$ with N being the covering number $\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{4C_G h}\right)$, we find

$$\begin{aligned}
&\text{Prob}\left\{\sup_{f\in\mathcal{H}} \frac{\big|V_f-\mathbb{E}[V_f]\big|}{(\mathbb{E}[V_f]+2\varepsilon)^{(q-2)/q}} > 4\varepsilon^{2/q}\right\} \le \text{Prob}\left\{\sup_{j=1,\ldots,N} \frac{\big|V_{f_j}-\mathbb{E}[V_{f_j}]\big|}{(\mathbb{E}[V_{f_j}]+2\varepsilon)^{(q-2)/q}} > \varepsilon^{2/q}\right\} \\
&\le \sum_{j=1,\ldots,N} \text{Prob}\left\{\frac{\big|V_{f_j}-\mathbb{E}[V_{f_j}]\big|}{(\mathbb{E}[V_{f_j}]+2\varepsilon)^{(q-2)/q}} > \varepsilon^{2/q}\right\}.
\end{aligned}$$

Fix $j \in \{1,\ldots,N\}$. Apply Lemma 17 to $U = U_{f_j}$ satisfying $\frac{1}{m(m-1)}\sum_{i=1}^m \sum_{j\neq i} U(z_i,z_j) - \mathbb{E}[U] = V_{f_j} - \mathbb{E}[V_{f_j}]$. By the bounds for $|U_{f_j}|$ and $\text{var}[U_{f_j}]$ from Part (b) of Lemma 18, we know by taking $\widetilde{\varepsilon} = \varepsilon^{2/q}(\mathbb{E}[V_{f_j}]+2\varepsilon)^{(q-2)/q}$ that

$$\begin{aligned}
&\text{Prob}\left\{\frac{\big|V_{f_j}-\mathbb{E}[V_{f_j}]\big|}{(\mathbb{E}[V_{f_j}]+2\varepsilon)^{(q-2)/q}} > \varepsilon^{2/q}\right\} = \text{Prob}\left\{\big|V_{f_j}-\mathbb{E}[V_{f_j}]\big| > \widetilde{\varepsilon}\right\} \\
&\le 2\exp\left\{-\frac{(m-1)\widetilde{\varepsilon}^2}{4A_{\mathcal{H}}\left(\text{var}[f_j(X)-f_\rho(X)]\right)^{(q-2)/q} + 12C_G\|f_j-f_\rho\|_\infty h\widetilde{\varepsilon}}\right\} \\
&\le 2\exp\left\{-\frac{(m-1)\varepsilon^{4/q}(\mathbb{E}[V_{f_j}]+2\varepsilon)^{(q-2)/q}}{4A_{\mathcal{H}}+12C_G\|f_j-f_\rho\|_\infty h\varepsilon^{2/q}}\right\},
\end{aligned}$$

where in the last step we have used the important relation (17) to the function $f = f_j$ and bounded $\left(\text{var}[f_j(X)-f_\rho(X)]\right)^{(q-2)/q}$ by $\left\{(\mathbb{E}[V_{f_j}]+2\varepsilon)\right\}^{(q-2)/q}$. This together with the notation $N = \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{4C_G h}\right)$ and the inequality $\|f_j-f_\rho\|_\infty \le \sup_{f\in\mathcal{H}}\|f-f_\rho\|_\infty$ gives the first desired bound, where we have observed that $\varepsilon \ge C''_{\mathcal{H}} h^{-q^*}$ and $h \ge 1$ imply $\varepsilon^{-2/q} \le (C''_{\mathcal{H}})^{-2/q} h$.

If $|Y| \le M$ almost surely for some constant $M > 0$, then we follows the same line as in our above proof. According to Part (b) of Lemma 18, we should replace $4C_G h$ by $2A'_{\mathcal{H}}$, q by 4, and bound the variance $\text{var}[U_{f_j}]$ by $2A'_{\mathcal{H}}\text{var}[f_j(X)-f_\rho(X)] \le 2A'_{\mathcal{H}}(\mathbb{E}[V_{f_j}]+2\varepsilon)$. Then the desired estimate follows. The proof of Lemma 19 is complete. ■

We are in a position to bound the sample error. To unify the two estimates in Lemma 19, we denote $A'_{\mathcal{H}} = 2C_G$ in the general case. For $m \in \mathbb{N}, 0 < \delta < 1$, let $\varepsilon_{m,\delta}$ be the smallest positive solution to the inequality

$$\log \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{2A'_{\mathcal{H}}}\right) - \frac{(m-1)\varepsilon}{A''_{\mathcal{H}}} \le \log\frac{\delta}{2}. \tag{22}$$

Proposition 20 *Let* $0<\delta<1, 0<\eta\leq 1$. *Under assumptions (2) and (3), we have with confidence of* $1-\delta$,

$$\mathbf{var}[f_{\mathbf{z}}(X)-f_\rho(X)]\leq(1+\eta)\mathbf{var}[f_{approx}(X)-f_\rho(X)]+12\left(2+24^{\frac{q-2}{2}}\right)\eta^{\frac{2-q}{2}}(h\varepsilon_{m,\delta}+2C''_{\mathcal{H}}h^{-q^*}).$$

If $|Y|\leq M$ *almost surely for some* $M>0$, *then with confidence of* $1-\delta$, *we have*

$$\mathbf{var}[f_{\mathbf{z}}(X)-f_\rho(X)]\leq(1+\eta)\mathbf{var}[f_{approx}(X)-f_\rho(X)]+\frac{278}{\eta}(\varepsilon_{m,\delta}+2C''_{\mathcal{H}}h^{-2}).$$

Proof Denote $\tau=(q-2)/q$ and $\varepsilon_{m,\delta,h}=\max\{h\varepsilon_{m,\delta},C''_{\mathcal{H}}h^{-q^*}\}$ in the general case with some $q>2$, while $\tau=1/2$ and $\varepsilon_{m,\delta,h}=\max\{\varepsilon_{m,\delta},C''_{\mathcal{H}}h^{-2}\}$ when $|Y|\leq M$ almost surely. Then by Lemma 19, we know that with confidence $1-\delta$, there holds

$$\sup_{f\in\mathcal{H}}\frac{\left|V_f-\mathbb{E}[V_f]\right|}{(\mathbb{E}[V_f]+2\varepsilon_{m,\delta,h})^\tau}\leq 4\varepsilon_{m,\delta,h}^{1-\tau}$$

which implies

$$\mathbb{E}\left[V_{f_{\mathbf{z}}}\right]-V_{f_{\mathbf{z}}}+V_{f_{\mathcal{H}}}-\mathbb{E}\left[V_{f_{\mathcal{H}}}\right]\leq 4\varepsilon_{m,\delta,h}^{1-\tau}(\mathbb{E}[V_{f_{\mathbf{z}}}]+2\varepsilon_{m,\delta,h})^\tau+4\varepsilon_{m,\delta,h}^{1-\tau}(\mathbb{E}[V_{f_{\mathcal{H}}}]+2\varepsilon_{m,\delta,h})^\tau.$$

This together with Lemma 15 and (16) yields

$$\mathbf{var}[f_{\mathbf{z}}(X)-f_\rho(X)]\leq 4\mathcal{S}+16\varepsilon_{m,\delta,h}+\mathbf{var}[f_{approx}(X)-f_\rho(X)]+2C''_{\mathcal{H}}h^{-q^*},\tag{23}$$

where

$$\mathcal{S}:=\varepsilon_{m,\delta,h}^{1-\tau}(\mathbb{E}[V_{f_{\mathbf{z}}}])^\tau+\varepsilon_{m,\delta,h}^{1-\tau}(\mathbb{E}[V_{f_{\mathcal{H}}}])^\tau=(\frac{24}{\eta})^\tau\varepsilon_{m,\delta,h}^{1-\tau}\left(\frac{\eta}{24}\mathbb{E}[V_{f_{\mathbf{z}}}]\right)^\tau+(\frac{12}{\eta})^\tau\varepsilon_{m,\delta,h}^{1-\tau}\left(\frac{\eta}{12}\mathbb{E}[V_{f_{\mathcal{H}}}]\right)^\tau.$$

Now we apply Young's inequality

$$a\cdot b\leq(1-\tau)a^{1/(1-\tau)}+\tau b^{1/\tau},\qquad a,b\geq 0$$

and find

$$\mathcal{S}\leq\left(\frac{24}{\eta}\right)^{\tau/(1-\tau)}\varepsilon_{m,\delta,h}+\frac{\eta}{24}\mathbb{E}[V_{f_{\mathbf{z}}}]+\left(\frac{12}{\eta}\right)^{\tau/(1-\tau)}\varepsilon_{m,\delta,h}+\frac{\eta}{12}\mathbb{E}[V_{f_{\mathcal{H}}}].$$

Combining this with (23), Theorem 13 and the identity $\mathbb{E}[V_f]=\mathcal{E}^{(h)}(f)-\mathcal{E}^{(h)}(f_\rho)$ gives

$$\mathbf{var}[f_{\mathbf{z}}(X)-f_\rho(X)]\leq\frac{\eta}{6}\mathbf{var}[f_{\mathbf{z}}(X)-f_\rho(X)]+(1+\frac{\eta}{3})\mathbf{var}[f_{approx}(X)-f_\rho(X)]+\mathcal{S}',$$

where $\mathcal{S}':=(16+8(24/\eta)^{\tau/(1-\tau)})\varepsilon_{m,\delta,h}+3C''_{\mathcal{H}}h^{-q^*}$. Since $1/(1-\frac{\eta}{6})\leq 1+\frac{\eta}{3}$ and $(1+\frac{\eta}{3})^2\leq 1+\eta$, we see that

$$\mathbf{var}[f_{\mathbf{z}}(X)-f_\rho(X)]\leq(1+\eta)\mathbf{var}[f_{approx}(X)-f_\rho(X)]+\frac{4}{3}\mathcal{S}'.$$

Then the desired estimates follow, and the proposition is proved. ■

5. Proof of Main Results

We are now in a position to prove our main results stated in Section 2.

5.1 Proof of Theorem 3

Recall $\mathcal{D}_{\mathcal{H}}(f_\rho) = \mathbf{var}[f_{approx}(X) - f_\rho(X)]$. Take $\eta = \min\{\varepsilon/(3\mathcal{D}_{\mathcal{H}}(f_\rho)), 1\}$. Then

$$\eta \mathbf{var}[f_{approx}(X) - f_\rho(X)] \le \varepsilon/3.$$

Now we take

$$h_{\varepsilon,\delta} = \left(72\left(2+24^{(q-2)/2}\right)\eta^{(2-q)/2}C''_{\mathcal{H}}/\varepsilon\right)^{1/q^*}.$$

Set $\widetilde{\varepsilon} := \varepsilon/\left(36\left(2+24^{(q-2)/2}\right)\eta^{(2-q)/2}\right)$. We choose

$$m_{\varepsilon,\delta}(h) = \frac{hA''_{\mathcal{H}}}{\widetilde{\varepsilon}}\left(\log \mathcal{N}\left(\mathcal{H}, \frac{\widetilde{\varepsilon}}{2hA'_{\mathcal{H}}}\right) - \log\frac{\delta}{2}\right) + 1.$$

With this choice, we know that whenever $m \ge m_{\varepsilon,\delta}(h)$, the solution $\varepsilon_{m,\delta}$ to inequality (22) satisfies $\varepsilon_{m,\delta} \le \widetilde{\varepsilon}/h$. Combining all the above estimates and Proposition 20, we see that whenever $h \ge h_{\varepsilon,\delta}$ and $m \ge m_{\varepsilon,\delta}(h)$, error bound (4) holds true with confidence $1-\delta$. This proves Theorem 3. ■

5.2 Proof of Theorem 5

We apply Proposition 20. By covering number condition (5), we know that $\varepsilon_{m,\delta}$ is bounded by $\widetilde{\varepsilon}_{m,\delta}$, the smallest positive solution to the inequality

$$A_p\left(\frac{2A'_{\mathcal{H}}}{\varepsilon}\right)^p - \frac{(m-1)\varepsilon}{A''_{\mathcal{H}}} \le \log\frac{\delta}{2}.$$

This inequality written as $\varepsilon^{1+p} - \frac{A''_{\mathcal{H}}}{m-1}\log\frac{2}{\delta}\varepsilon^p - A_p\left(2A'_{\mathcal{H}}\right)^p\frac{A''_{\mathcal{H}}}{m-1} \ge 0$ is well understood in learning theory (e.g., Cucker and Zhou, 2007) and its solution can be bounded as

$$\widetilde{\varepsilon}_{m,\delta} \le \max\left\{2\frac{A''_{\mathcal{H}}}{m-1}\log\frac{2}{\delta}, \left(2A_pA''_{\mathcal{H}}(2A'_{\mathcal{H}})^p\right)^{1/(1+p)}(m-1)^{-\frac{1}{1+p}}\right\}.$$

If $\mathbb{E}[|Y|^q] < \infty$ for some $q > 2$, then the first part of Proposition 20 verifies (6) with the constant $\widetilde{C}_{\mathcal{H}}$ given by

$$\widetilde{C}_{\mathcal{H}} = 24\left(2+24^{(q-2)/2}\right)\left(2A''_{\mathcal{H}} + \left(2A_pA''_{\mathcal{H}}(2A'_{\mathcal{H}})^p\right)^{1/(1+p)} + 2C''_{\mathcal{H}}\right).$$

If $|Y| \le M$ almost surely for some $M > 0$, then the second part of Proposition 20 proves (7) with the constant $\widetilde{C}_{\mathcal{H}}$ given by

$$\widetilde{C}_{\mathcal{H}} = 278\left(2A''_{\mathcal{H}} + \left(2A_pA''_{\mathcal{H}}(2A'_{\mathcal{H}})^p\right)^{1/(1+p)} + 2C''_{\mathcal{H}}\right).$$

This completes the proof of Theorem 5. ■

5.3 Proof of Theorem 8

Note

$$\left|\frac{1}{m}\sum_{i=1}^{m}\left[f(x_i)-\pi_{\sqrt{m}}(y_i)\right]-\frac{1}{m}\sum_{i=1}^{m}\left[g(x_i)-\pi_{\sqrt{m}}(y_i)\right]\right|\le\|f-g\|_\infty$$

and

$$\left|\mathbb{E}[f(X)-\pi_{\sqrt{m}}(Y)]-\mathbb{E}[g(X)-\pi_{\sqrt{m}}(Y)]\right|\le\|f-g\|_\infty.$$

So by taking $\{f_j\}_{j=1}^N$ to be an $\frac{\varepsilon}{4}$ net of the set $\mathcal{H}$ with $N=\mathcal{N}\left(\mathcal{H},\frac{\varepsilon}{4}\right)$, we know that for each $f\in\mathcal{H}$ there is some $j\in\{1,\ldots,N\}$ such that $\|f-f_j\|_\infty\le\frac{\varepsilon}{4}$. Hence

$$\begin{aligned}&\left|\frac{1}{m}\sum_{i=1}^{m}\left[f(x_i)-\pi_{\sqrt{m}}(y_i)\right]-\mathbb{E}[f(X)-\pi_{\sqrt{m}}(Y)]\right|>\varepsilon\\ \Longrightarrow\quad&\left|\frac{1}{m}\sum_{i=1}^{m}\left[f_j(x_i)-\pi_{\sqrt{m}}(y_i)\right]-\mathbb{E}[f_j(X)-\pi_{\sqrt{m}}(Y)]\right|>\frac{\varepsilon}{2}.\end{aligned}$$

It follows that

$$\begin{aligned}&\mathrm{Prob}\left\{\sup_{f\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}\left[f(x_i)-\pi_{\sqrt{m}}(y_i)\right]-\mathbb{E}[f(X)-\pi_{\sqrt{m}}(Y)]\right|>\varepsilon\right\}\\ \le\quad&\mathrm{Prob}\left\{\sup_{j=1,\ldots,N}\left|\frac{1}{m}\sum_{i=1}^{m}\left[f_j(x_i)-\pi_{\sqrt{m}}(y_i)\right]-\mathbb{E}[f_j(X)-\pi_{\sqrt{m}}(Y)]\right|>\frac{\varepsilon}{2}\right\}\\ \le\quad&\sum_{j=1}^{N}\mathrm{Prob}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\left[f_j(x_i)-\pi_{\sqrt{m}}(y_i)\right]-\mathbb{E}[f_j(X)-\pi_{\sqrt{m}}(Y)]\right|>\frac{\varepsilon}{2}\right\}.\end{aligned}$$

For each fixed $j\in\{1,\ldots,N\}$, we apply the classical Bernstein probability inequality to the random variable $\xi=f_j(X)-\pi_{\sqrt{m}}(Y)$ on (Z,ρ) bounded by $\widetilde{M}=\sup_{f\in\mathcal{H}}\|f\|_\infty+\sqrt{m}$ with variance $\sigma^2(\xi)\le\mathbb{E}[|f_j(X)-\pi_{\sqrt{m}}(Y)|^2]\le 2\sup_{f\in\mathcal{H}}\|f\|_\infty^2+2\mathbb{E}[|Y|^2]=:\sigma_{\mathcal{H}}^2$ and know that

$$\begin{aligned}&\mathrm{Prob}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\left[f_j(x_i)-\pi_{\sqrt{m}}(y_i)\right]-\mathbb{E}[f_j(X)-\pi_{\sqrt{m}}(Y)]\right|>\frac{\varepsilon}{2}\right\}\\ &\le 2\exp\left\{-\frac{m(\varepsilon/2)^2}{\frac{2}{3}\widetilde{M}\varepsilon/2+2\sigma^2(\xi)}\right\}\le 2\exp\left\{-\frac{m\varepsilon^2}{\frac{4}{3}\widetilde{M}\varepsilon+8\sigma_{\mathcal{H}}^2}\right\}.\end{aligned}$$

The above argument together with covering number condition (5) yields

$$\begin{aligned}&\mathrm{Prob}\left\{\sup_{f\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}\left[f(x_i)-\pi_{\sqrt{m}}(y_i)\right]-\mathbb{E}[f(X)-\pi_{\sqrt{m}}(Y)]\right|>\varepsilon\right\}\\ &\le 2N\exp\left\{-\frac{m\varepsilon^2}{\frac{4}{3}\widetilde{M}\varepsilon+8\sigma_{\mathcal{H}}^2}\right\}\le 2\exp\left\{A_p\left(\frac{4}{\varepsilon}\right)^p-\frac{m\varepsilon^2}{\frac{4}{3}\widetilde{M}\varepsilon+8\sigma_{\mathcal{H}}^2}\right\}.\end{aligned}$$

Bounding the right-hand side above by δ is equivalent to the inequality

$$\varepsilon^{2+p}-\frac{4}{3m}\widetilde{M}\log\frac{2}{\delta}\varepsilon^{1+p}-\frac{8}{m}\sigma_{\mathcal{H}}^2\log\frac{2}{\delta}\varepsilon^p-\frac{A_p4^p}{m}\ge 0.$$

By taking $\widetilde{\varepsilon}_{m,\delta}$ to be the smallest solution to the above inequality, we see from Cucker and Zhou (2007) as in the proof of Theorem 5 that with confidence at least $1-\delta$,

$$\begin{aligned}
&\sup_{f\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}\left[f(x_i)-\pi_{\sqrt{m}}(y_i)\right]-\mathbb{E}[f(X)-\pi_{\sqrt{m}}(Y)]\right| \\
&\le\widetilde{\varepsilon}_{m,\delta}\le\max\left\{\frac{4\widetilde{M}}{m}\log\frac{2}{\delta},\sqrt{\frac{24\sigma^2_{\mathcal{H}}}{m}\log\frac{2}{\delta}},\left(\frac{A_p4^p}{m}\right)^{\frac{1}{2+p}}\right\} \\
&\le\left\{7\sup_{f\in\mathcal{H}}\|f\|_\infty+4+7\sqrt{\mathbb{E}[|Y|^2]}+4A_p^{\frac{1}{2+p}}\right\}m^{-\frac{1}{2+p}}\log\frac{2}{\delta}.
\end{aligned}$$

Moreover, since $\pi_{\sqrt{m}}(y)-y=0$ for $|y|\le\sqrt{m}$ while $|\pi_{\sqrt{m}}(y)-y|\le|y|\le\frac{|y|^2}{\sqrt{m}}$ for $|y|>\sqrt{m}$, we know that

$$\begin{aligned}
&\left|\mathbb{E}[\pi_{\sqrt{m}}(Y)]-\mathbb{E}[f_\rho(X)]\right|=\left|\int_X\int_Y\pi_{\sqrt{m}}(y)-yd\rho(y|x)d\rho_X(x)\right| \\
&=\left|\int_X\int_{|y|>\sqrt{m}}\pi_{\sqrt{m}}(y)-yd\rho(y|x)d\rho_X(x)\right|\le\int_X\int_{|y|>\sqrt{m}}\frac{|y|^2}{\sqrt{m}}d\rho(y|x)d\rho_X(x)\le\frac{\mathbb{E}[|Y|^2]}{\sqrt{m}}.
\end{aligned}$$

Therefore, (11) holds with confidence at least $1-\delta$. The proof of Theorem 8 is complete. ■

6. Conclusion and Discussion

In this paper we have proved the consistency of an MEE algorithm associated with Rényi's entropy of order 2 by letting the scaling parameter h in the kernel density estimator tends to infinity at an appropriate rate. This result explains the effectiveness of the MEE principle in empirical applications where the parameter h is required to be large enough before smaller values are tuned. However, the motivation of the MEE principle is to minimize error entropies approximately, and requires small h for the kernel density estimator to converge to the true probability density function. Therefore, our consistency result seems surprising.

As far as we know, our result is the first rigorous consistency result for MEE algorithms. There are many open questions in mathematical analysis of MEE algorithms. For instance, can MEE algorithm (1) be consistent by taking $h\to 0$? Can one carry out error analysis for the MEE algorithm if Shannon's entropy or Rényi's entropy of order $\alpha\neq 2$ is used? How can we establish error analysis for other learning settings such as those with non-identical sampling processes (Smale and Zhou, 2009; Hu, 2011)? These questions require further research and will be our future topics.

It might be helpful to understand our theoretical results by relating MEE algorithms to ranking algorithms. Note that MEE algorithm (1) essentially minimizes the empirical version of the information error which, according to our study in Section 2, differs from the symmetrized least squares error used in some ranking algorithms by an extra term which vanishes when $h\to\infty$. Our study may shed some light on analysis of some ranking algorithms.

Table 1: NOTATIONS

notation	meaning	pages
p_E	probability density function of a random variable E	378
$H_S(E)$	Shannon's entropy of a random variable E	378
$H_{R,\alpha}(E)$	Rényi's entropy of order α	378
X	explanatory variable for learning	378
Y	response variable for learning	378
$E = Y - f(X)$	error random variable associated with a predictor $f(X)$	378
$H_R(E)$	Rényi's entropy of order $\alpha = 2$	378
$\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$	a sample for learning	378
G	windowing function	378, 379, 380
h	MEE scaling parameter	378, 379
$\widehat{p_E}$	Parzen windowing approximation of p_E	378
$\widehat{H_S}$	empirical Shannon entropy	378
$\widehat{H_R}$	empirical Rényi's entropy of order 2	378
f_ρ	the regression function of ρ	379
$f_{\mathbf{z}}$	output function of the MEE learning algorithm (1)	379
$\mathcal{H}$	the hypothesis space for the ERM algorithm	379
var	the variance of a random variable	379
$q, q^* = \min\{q-2, 2\}$	power indices in condition (2) for $\mathbb{E}[\lvert Y\rvert^q] < \infty$	380
C_G	constant for decay condition (3) of G	380
$\mathcal{D}_{\mathcal{H}}(f_\rho)$	approximation error of the pair $(\mathcal{H}, \rho)$	380
$\mathcal{N}(\mathcal{H}, \varepsilon)$	covering number of the hypothesis space $\mathcal{H}$	380
p	power index for covering number condition (5)	380
$\pi_{\sqrt{m}}$	projection onto the closed interval $[-\sqrt{m}, \sqrt{m}]$	381
$\widetilde{f_{\mathbf{z}}}$	estimator of f_ρ	382
$\mathcal{E}^{(h)}(f)$	generalization error associated with G and h	383
$\mathcal{E}^{ls}(f)$	least squares generalization error $\mathcal{E}^{ls}(f) = \int_{\mathcal{Z}} (f(x) - y)^2 d\rho$	383
C_ρ	constant $C_\rho = \int_{\mathcal{Z}} \left[y - f_\rho(x)\right]^2 d\rho$ associated with ρ	384
$f_{\mathcal{H}}$	minimizer of $\mathcal{E}^{(h)}(f)$ in $\mathcal{H}$	385
f_{approx}	minimizer of $\mathbf{var}[f(X) - f_\rho(X)]$ in $\mathcal{H}$	385
U_f	kernel for the U statistics V_f	387
$\widetilde{G}$	an intermediate function defined by $\widetilde{G}(t) = G(t^2/2)$	388

Acknowledgments

We would like to thank the referees for their constructive suggestions and comments. The work described in this paper was supported by National Science Foundation of China under Grants (No. 11201348 and 11101403) and by a grant from the Research Grants Council of Hong Kong [Project No. CityU 103709]. The corresponding author is Ding-Xuan Zhou.

References

S. Agarwal and P. Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10:441–474, 2009.

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

J. Y. Audibert and O. Catoni. Robust linear least squares regression. *Annals of Statistics*, 39:2766–2794, 2011.

B. Chen and J. C. Principe. Some further results on the minimum error entropy estimation. *Entropy*, 14:966–977, 2012.

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

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

D. Erdogmus and J. C. Principe. An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems. *IEEE Transactions on Signal Processing*, 50:1780–1786, 2002.

D. Erdogmus and J. C. Principe. Convergence properties and data efficiency of the minimum error entropy criterion in adaline training. *IEEE Transactions on Signal Processing*, 51:1966–1978, 2003.

D. Haussler, M. Kearns, and R. Schapire. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. *Machine Learning*, 14:83–114, 1994.

T. Hu. Online regression with varying Gaussians and non-identical distributions. *Analysis and Applications*, 9:395–408, 2011.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34:2593–2656, 2006.

E. Parzen. On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, 33:1049–1051, 1962.

J. C. Principe. *Information Theoretic Learning: Rényi's Entropy and Kernel Perspectives*. Springer, New York, 2010.

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

S. Smale and D. X. Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1:17–41, 2003.

S. Smale and D. X. Zhou. Online learning with Markov sampling. *Analysis and Applications*, 7:87–113, 2009.

V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

Y. Yao. On complexity issue of online learning algorithms. *IEEE Transactions on Information Theory*, 56:6470–6481, 2010.

Y. Ying. Convergence analysis of online algorithms. *Advances in Computational Mathematics*, 27:273–291, 2007.

D. X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18:739–767, 2002.

D. X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49:1743-1752, 2003.