ARTICLE IN PRESS YACHA:1017

Appl. Comput. Harmon. Anal. • • • (• • • •) • • •–• • •

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Consistency analysis of an empirical minimum error entropy algorithm ☆

Jun Fan [a], Ting Hu [b], Qiang Wu [c,*], Ding-Xuan Zhou [d]

[a] *Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706, USA*
[b] *School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China*
[c] *Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, TN 37132, USA*
[d] *Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China*

ARTICLE INFO

Article history:
Received 19 September 2014
Accepted 16 December 2014
Available online xxxx
Communicated by Qingtang Jiang

Keywords:
Minimum error entropy
Learning theory
Rényi's entropy
Error entropy consistency
Regression consistency

ABSTRACT

In this paper we study the consistency of an empirical minimum error entropy (MEE) algorithm in a regression setting. We introduce two types of consistency. The error entropy consistency, which requires the error entropy of the learned function to approximate the minimum error entropy, is shown to be always true if the bandwidth parameter tends to 0 at an appropriate rate. The regression consistency, which requires the learned function to approximate the regression function, however, is a complicated issue. We prove that the error entropy consistency implies the regression consistency for homoskedastic models where the noise is independent of the input variable. But for heteroskedastic models, a counterexample is used to show that the two types of consistency do not coincide. A surprising result is that the regression consistency is always true, provided that the bandwidth parameter tends to infinity at an appropriate rate. Regression consistency of two classes of special models is shown to hold with fixed bandwidth parameter, which further illustrates the complexity of regression consistency of MEE. Fourier transform plays crucial roles in our analysis.

Published by Elsevier Inc.

1. Introduction

Information theoretical learning (ITL) is an important research area in signal processing and machine learning. It uses concepts of entropies and divergences from information theory to substitute the conventional statistical descriptors of variances and covariances. The idea dates back at least to [12] while its blossom

☆ The work described in this paper is supported by National Natural Science Foundation of China under Grants (No. 11201079, 11201348, 11101403, and 11461161006) and by a grant from the Research Grants Council of Hong Kong [Project No. CityU 104012].
* Corresponding author. Fax: +1 615 898 5422.
E-mail addresses: junfan@stat.wisc.edu (J. Fan), tinghu@whu.edu.cn (T. Hu), qwu@mtsu.edu (Q. Wu), mazhou@cityu.edu.hk (D.-X. Zhou).

http://dx.doi.org/10.1016/j.acha.2014.12.005
1063-5203/Published by Elsevier Inc.

was inspired by a series of works of Principe and his collaborators. In [4] the minimum error entropy (MEE) principle was introduced to regression problems. Later on its computational properties were studied and its applications in feature extraction, clustering, and blind source separation were developed [5,7,3,6]. More recently the MEE principle was applied to classification problems [16,17]. For a comprehensive survey and more recent advances on ITL and the MEE principle, see [15] and the references therein.

The main purpose of this paper is rigorous consistency analysis of an empirical MEE algorithm for regression. Note that the ultimate goal of regression problems is the prediction on unobserved data or forecasting the future. Consistency analysis in terms of predictive powers is deemed to be important to interpret the effectiveness of a regression algorithm. The empirical MEE has been developed and successfully applied in various fields for more than a decade and there are some theoretical studies in the literature which provide good understanding of computational complexity of the empirical MEE and its parameter choice strategy. However, the consistency of the MEE algorithm, especially from a prediction perspective, is lacking. In our earlier work [8], we proved the consistency of the MEE algorithm in a special situation, where we require the algorithm to utilize a large bandwidth parameter. The motivation of the MEE algorithm (to be describe below) is to minimize the error entropy which requires a small bandwidth parameter. The result in [8] is somewhat contradictory to this motivation. An interesting question is whether the MEE algorithm is consistent in terms of predictive powers if a small bandwidth parameter is chosen as implied by its motivation. Unfortunately, this is not a simple 'yes' or 'no' question. Instead, the consistency of the MEE algorithm is a very complicated issue. In this paper we will try to depict a full picture on it – establishing the relationship between the error entropy and an L_2 metric measuring the predictive powers, and providing conditions for the MEE algorithm to be predictively consistent.

In statistics a regression problem is usually modeled as the estimation of a target function f^* from a metric space $\mathcal{X}$ to the another metric space $\mathcal{Y} \subset \mathbb{R}$ for which a set of observations $(x_i, y_i),\ i = 1, \ldots, n$, are obtained from a model

$$Y = f^*(X) + \epsilon, \qquad \mathbf{E}(\epsilon|X) = 0. \tag{1.1}$$

In the statistical learning context [18], the regression setting is usually described as the learning of the regression function which is defined as conditional mean $\mathbf{E}(Y|X)$ of the output variable Y for given input variable X under the assumption that there is an unknown joint probability measure ρ on the product space $\mathcal{X} \times \mathcal{Y}$. These two settings are equivalent by noticing that

$$f^*(x) = \mathbf{E}(Y|X = x).$$

A learning algorithm for regression produces a function $f_{\mathbf{z}}$ from the observations $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n$ as an approximation of f^*. The goodness of this approximation can be measured by certain distance between $f_{\mathbf{z}}$ and f^*, for instance, $\|f_{\mathbf{z}} - f^*\|_{L^2_{\rho_X}}$, the L^2 distance with respect to the marginal distribution ρ_X of ρ on $\mathcal{X}$.

MEE algorithms for regression are motivated by minimizing some entropies of the error random variable $E = E(f) = Y - f(X)$, where $f : X \to \mathbb{R}$ is a hypothesis function. In this paper we focus on the Rényi's entropy of order 2 defined as

$$\mathscr{R}(f) = -\log\big(\mathbf{E}[p_{\scriptscriptstyle E}]\big) = -\log\bigg(\int_{\mathbb{R}} \big(p_{\scriptscriptstyle E}(e)\big)^2\, de\bigg). \tag{1.2}$$

Here and in the sequel, $p_{\scriptscriptstyle E}$ is the probability density function of E. Since ρ is unknown, we need an empirical estimate of $p_{\scriptscriptstyle E}$. Denote $e_i = y_i - f(x_i)$. Then $p_{\scriptscriptstyle E}$ can be estimated from the sample $\mathbf{z}$ by a kernel density estimator by using a Gaussian kernel $G_h(t) = \frac{1}{\sqrt{2\pi}h}\exp(-\frac{t^2}{2h^2})$ with bandwidth parameter h:

Please cite this article in press as: J. Fan et al., Consistency analysis of an empirical minimum error entropy algorithm, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.12.005

$$p_{E,\mathbf{z}}(e) = \frac{1}{n}\sum_{j=1}^{n} G_h(e-e_j) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{\sqrt{2\pi}h}\exp\left(-\frac{(e-e_j)^2}{2h^2}\right).$$

The MEE algorithm produces an appropriate $f_{\mathbf{z}}$ from a set $\mathcal{H}$ of continuous functions on $\mathcal{X}$ called the hypothesis space by minimizing the empirical version of the Rényi's entropy

$$\mathscr{R}_{\mathbf{z}}(f) = -\log\left(\frac{1}{n}\sum_{i=1}^{n} p_{E,\mathbf{z}}(e_i)\right) = -\log\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} G_h(e_i-e_j)\right).$$

That is, $f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}} \mathscr{R}_{\mathbf{z}}(f)$. It is obvious that minimizers of $\mathscr{R}$ and $\mathscr{R}_{\mathbf{z}}$ are not unique because $\mathscr{R}(f) = \mathscr{R}(f+b)$ and $\mathscr{R}_{\mathbf{z}}(f) = \mathscr{R}_{\mathbf{z}}(f+b)$ for any constant b. Taking this into account, $f_{\mathbf{z}}$ should be adjusted by a constant when it is used as an approximation of the regression function f^*.

The empirical entropy $\mathscr{R}_{\mathbf{z}}(f)$ involves an empirical mean $\frac{1}{n}\sum_{i=1}^{n} p_{E,\mathbf{z}}(e_i)$ which makes it look like an M-estimator. However, the density estimator $p_{E,\mathbf{z}}$ itself is data dependent, making the MEE algorithm different from standard M-estimations, with two summation indices involved. This can be seen from our earlier work [8] where we used U-statistics for the error analysis in the case of large parameter h.

To study the asymptotic behavior of the MEE algorithm we define two types of consistency as follows:

Definition 1.1. The MEE algorithm is **consistent with respect to the Rényi's error entropy** if $\mathscr{R}(f_{\mathbf{z}})$ converges to $\mathscr{R}^* = \inf_{f:\mathcal{X}\to\mathbb{R}} \mathscr{R}(f)$ in probability as $n \to \infty$, i.e.,

$$\lim_{n\to\infty} \mathbf{P}\big(\mathscr{R}(f_{\mathbf{z}}) - \mathscr{R}^* > \varepsilon\big) = 0, \quad \forall \varepsilon > 0.$$

The MEE algorithm is **consistent with respect to the regression function** if $f_{\mathbf{z}}$ plus a suitable constant adjustment converges to f^* in probability with the convergence measured in the $L^2_{\rho_X}$ sense, i.e., there is a constant $b_{\mathbf{z}}$ such that $f_{\mathbf{z}} + b_{\mathbf{z}}$ converges to f^* in probability, i.e.,

$$\lim_{n\to\infty} \mathbf{P}\big(\|f_{\mathbf{z}} + b_{\mathbf{z}} - f^*\|^2_{L^2_{\rho_X}} > \varepsilon\big) = 0, \quad \forall \varepsilon > 0.$$

Note that the error entropy consistency ensures the learnability of the minimum error entropy, as is expected from the motivation of empirical MEE algorithms. However, the error entropy itself is not a metric that directly measures the predictive powers of the algorithm. (We assume that a metric d measuring the predictive powers should be monotone in the sense that $|E(f_1)| \le |E(f_2)|$ implies $d(f_1) \le d(f_2)$. Error entropy is clearly not such a metric.)

To measure the predictive consistency, one may choose different metrics. In the definition of regression function consistency we have adopted the L_2 distance to the true regression function f^*, the target function of the regression problem. The regression consistency guarantees good approximations of the regression target function f^* and thus serves as a good measure for predictive powers.

Our main results, stated in several theorems in Section 2 below, involve two main contributions. (i) We characterize the relationship between the error entropy consistency and regression consistency. We prove that the error entropy consistency implies the regression function consistency only for very special cases, for instance, the homoskedastic models, while in general this is not true. For heteroskedastic models, a counterexample is used to show that the error entropy consistency and regression consistency is not necessary to coincide. (ii) We prove a variety of consistency results for the MEE algorithm. Firstly we prove that the error entropy consistency is always true by choosing the bandwidth parameter h to tend to 0 slowly enough. As a result, the regression function consistency holds for the homoskedastic models. Secondly, for heteroskedastic models, regression consistency is shown to be incorrect if the bandwidth parameter is chosen to be small. But we restate the result from [8] which shows that the empirical MEE is always consistent with respect to the regression function if the bandwidth parameter is allowed to be chosen large enough. Lastly, we consider two classes of special regression models for which the regression consistency can be true with fixed choices of the bandwidth parameter h. These results indicate that the consistency of the empirical MEE is a very complicated issue.

2. Main results

We state our main results in this section while giving their proofs later. We need to make some assumptions for analysis purposes. Two main assumptions, on the regression model and the hypothesis class respectively, will be used throughout the paper.

For the regression model, we assume some natural regularity conditions.

Definition 2.1. The regression model (1.1) is **MEE admissible** if

(i) the density function $p_{\epsilon|X}$ of the noise variable ϵ for given $X = x \in \mathcal{X}$ exists and is uniformly bounded by a constant M_p;
(ii) the regression function f^* is bounded by a constant $M > 0$;
(iii) the minimum of $\mathscr{R}(f)$ is achieved by a measurable function $f^*_{\mathscr{R}}$.

Note that we do not require the boundedness or exponential decay of the noise term as in the usual setting of learning theory. It is in fact an advantage of MEE to allow heavy tailed noises. Also, it is easy to see that if $f^*_{\mathscr{R}}$ is a minimizer, then for any constant b, $f^*_{\mathscr{R}} + b$ is also a minimizer. So we cannot assume the uniqueness of $f^*_{\mathscr{R}}$. Also, no obvious relationship between f^* and $f^*_{\mathscr{R}}$ exists. Figuring out such a non-trivial relationship is one of our tasks below. We also remark that some results below may hold under relaxed conditions, but for simplifying our statements, we will not discuss them in detail.

Our second assumption is on the hypothesis space which is required to be a learnable class and have good approximation ability with respect to the target function.

Definition 2.2. We say the hypothesis space $\mathcal{H}$ is **MEE admissible** if

(i) $\mathcal{H}$ is uniformly bounded, i.e., there is a constant M such that $|f(x)| \le M$ for all $f \in \mathcal{H}$ and all $x \in \mathcal{X}$;
(ii) the ℓ_2-norm empirical cover number $\mathcal{N}_2(\mathcal{H}, \varepsilon)$ (see Appendix A or [2,19] for its definition) satisfies $\log \mathcal{N}_2(\mathcal{H}, \varepsilon) \le c\varepsilon^{-s}$ for some constant $c > 0$ and some index $0 < s < 2$;
(iii) a minimizer $f^*_{\mathscr{R}}$ of $\mathscr{R}(f)$ and the regression function f^* are in $\mathcal{H}$.

The first condition in Definition 2.2 is common in the literature and is natural since we do not expect to learn unbounded functions. The second condition ensures $\mathcal{H}$ is a learnable class so that overfitting will not happen. This is often imposed in learning theory. It is also easily fulfilled by many commonly used function classes. The third condition guarantees the target function can be well approximated by $\mathcal{H}$ for otherwise no algorithm is able to learn the target function well from $\mathcal{H}$. Although this condition can be relaxed to that the target function can be approximated by function sequences in $\mathcal{H}$, we will not adopt this relaxed situation for simplicity.

Throughout the paper, we assume that both the regression model (1.1) and the hypothesis space $\mathcal{H}$ are MEE admissible. Our first main result is to verify the error entropy consistency.

Theorem 2.3. *If the bandwidth parameter $h = h(n)$ is chosen to satisfy*

$$\lim_{n\to\infty} h(n) = 0, \qquad \lim_{n\to\infty} h^2\sqrt{n} = +\infty, \tag{2.1}$$

then $\mathscr{R}(f_{\mathbf{z}})$ converges to $\mathscr{R}^*$ in probability.

If, in addition, the derivative of $p_{\epsilon|X}$ exists and is uniformly bounded by a constant M' independent of X, then by choosing $h(n) \sim n^{-\frac{1}{6}}$, for any $0 < \delta < 1$, with probability at least $1-\delta$, we have

$$\mathscr{R}(f_{\mathbf{z}}) - \mathscr{R}^* = O\big(\sqrt{\log(2/\delta)}n^{-\frac{1}{6}}\big).$$

In the literature of practical implementations of MEE, the optimal choice of h is suggested to be $h(n) \sim n^{-\frac{1}{5}}$ (see e.g. [15]). We see this choice satisfies our condition for the error entropy consistency. Deriving the optimal rate for MEE is certainly of critical importance but out of the scope of this paper.

The error entropy consistency in Theorem 2.3 states the minimum error entropy can be approximated with a suitable choice of the bandwidth parameter. This is a somewhat expected result because empirical MEE algorithms are motivated by minimizing the sample version of the error entropy risk functional. However, later we will show that this does not necessarily imply the consistency with respect to the regression function. Instead, the regression consistency is a complicated problem. Let us discuss it in two different situations.

Definition 2.4. The regression mode (1.1) is **homoskedastic** if the noise ϵ is independent of X. Otherwise it is said to be **heteroskedastic**.

Our second main result states the regression consistency for homoskedastic models.

Theorem 2.5. *If the regression model is homoskedastic, then the following holds.*

(i) *$\mathscr{R}^* = \mathscr{R}(f^*)$. As a result, for any constant b, $f^*_{\mathscr{R}} = f^* + b$ is a minimizer of $\mathscr{R}(f)$;*
(ii) *there is a constant C depending on $\rho, \mathcal{H}$ and M such that, for any $f \in \mathcal{H}$,*

$$\big\| f + \mathbf{E}_x\big[f^*(x) - f(x)\big] - f^* \big\|^2_{L^2_{\rho_X}} \le C\big(\mathscr{R}(f) - \mathscr{R}^*\big);$$

(iii) *if (2.1) is true, then $f_{\mathbf{z}} + \mathbf{E}_x[f^*(x) - f_{\mathbf{z}}(x)]$ converges to f^* in probability;*
(iv) *if, in addition, the derivative of $p_{\epsilon|X}$ exists and is uniformly bounded by a constant M' independent of X, then, for any $0 < \delta < 1$, by choosing $h \sim n^{-\frac{1}{6}}$ we have, with confidence $1-\delta$, that*

$$\big\| f_{\mathbf{z}} + \mathbf{E}_x\big[f^*(x) - f_{\mathbf{z}}(x)\big] - f^* \big\|^2_{L^2_{\rho_X}} = O\big(\sqrt{\log(2/\delta)}n^{-\frac{1}{6}}\big).$$

Theorem 2.5 (iii) shows the regression consistency for homoskedastic models. It is a corollary of error entropy consistency stated in Theorem 2.3 and the relationship between the $L^2_{\rho_X}$ distance and the excess error entropy stated in Theorem 2.5 (ii). Thus the homoskedastic model is a special case for which the error entropy consistency and regression consistency coincide with each other.

Things are much more complicated for heteroskedastic models. Our third main result illustrates the incoincidence of the minimizer $f^*_{\mathscr{R}}$ and the regression function f^* by Example 5.1 in Section 5.

Proposition 2.6. *There exists a heteroskedastic model such that the regression function f^* is not a minimizer of $\mathscr{R}(f)$ and the regression consistency fails even if the error entropy consistency is true.*

This result shows that, in general, the error entropy consistency does not imply the regression consistency. Therefore, these two types of consistency do not coincide for heteroskedastic models.

However, this observation does not mean the empirical MEE algorithm cannot be consistent with respect to the regression function. In fact, in [8] we proved the regression consistency for large bandwidth parameter h and derived learning rate when h is of the form $h = n^\theta$ for some $\theta > 0$.

Please cite this article in press as: J. Fan et al., Consistency analysis of an empirical minimum error entropy algorithm, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.12.005

Our fourth main result in this paper is to verify the regression consistency for a more general choice of large bandwidth parameter h.

Theorem 2.7. *Choosing the bandwidth parameter $h = h(n)$ such that*

$$\lim_{n\to\infty} h(n) = +\infty, \qquad \lim_{n\to\infty} \frac{h^2}{\sqrt{n}} = 0, \tag{2.2}$$

we have $f_{\mathbf{z}} + \mathbf{E}_x[f^*(x) - f_{\mathbf{z}}(x)]$ converges to f^* in probability. A convergence rate of order $O(\sqrt{\log(2/\delta)}n^{-\frac{1}{4}})$ can be obtained with confidence $1-\delta$ for $\|f_{\mathbf{z}} + \mathbf{E}_x[f^*(x) - f_{\mathbf{z}}(x)] - f^*\|^2_{L^2_{\rho_X}}$ by taking $h \sim n^{\frac{1}{8}}$.

Such a result looks surprising. Note that the empirical MEE algorithm is motivated by minimizing an empirical version of the error entropy. This empirical error entropy approximates the true one when h tends to zero. But the regression consistency is in general true as h tends to infinity, a condition under which the error entropy consistency may not be true. From this point of view, the regression consistency of the empirical MEE algorithm does not justify its motivation.

Observe that the regression consistency in Theorem 2.5 and Theorem 2.7 suggests the constant adjustment to be $b = \mathbf{E}_x[f^*(x) - f_{\mathbf{z}}(x)]$. In practice the constant adjustment is usually taken as $\frac{1}{n}\sum_{i=1}^n (y_i - f_{\mathbf{z}}(x_i))$ which is exactly the sample mean of b.

The last two main results of this paper are about the regression consistency of two special classes of regression models. We show that the bandwidth parameter h can be chosen as a fixed positive constant to make MEE consistent in these situations. Moreover the convergence rate is of order $O(n^{-1/2})$, much higher than previous general cases. Throughout this paper, we use i to denote the imaginary unit and $\bar{a}$ the conjugate of a complex number a. The Fourier transform $\widehat{f}$ is defined for an integrable function f on $\mathbb{R}$ as $\widehat{f}(\xi) = \int_{\mathbb{R}} f(x)\mathrm{e}^{-\mathrm{i}x\xi}dx$. Recall the inverse Fourier transform is given by $f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{f}(\xi)\mathrm{e}^{\mathrm{i}x\xi}d\xi$ when f is square integrable. Fourier transform plays crucial roles in our analysis.

Definition 2.8. A univariate function f is **unimodal** if for some $t \in \mathbb{R}$, the function is monotonically increasing on $(-\infty, t]$ and monotonically decreasing on $[t, \infty)$.

Definition 2.9. We define $\mathcal{P}_1$ to be the set of probability measures ρ on $\mathcal{X} \times \mathcal{Y}$ satisfying the following conditions:

(i) $p_{\epsilon|X=x}$ is symmetric (i.e. even) and unimodal for every $x \in \mathcal{X}$;
(ii) the Fourier transform $\widehat{p_{\epsilon|X=x}}$ is nonnegative on $\mathbb{R}$ for every $x \in \mathcal{X}$;
(iii) there exist two constants $c_0 > 0$ and $C_0 > 0$ such that $\widehat{p_{\epsilon|X=x}}(\xi) \geq C_0$ for $\xi \in [-c_0, c_0]$ and every $x \in \mathcal{X}$.

We define $\mathcal{P}_2$ to be the set of probability measures ρ on $\mathcal{X} \times \mathcal{Y}$ such that $p_{\epsilon|X=x}$ is symmetric for every $x \in \mathcal{X}$ and there exists some constant $\widetilde{M} > 0$ such that $p_{\epsilon|X=x}$ is supported on $[-\widetilde{M}, \widetilde{M}]$ for every $x \in \mathcal{X}$.

The boundedness assumption on the noise for the family $\mathcal{P}_2$ is very natural in regression setting. For the family $\mathcal{P}_1$, the conditions look complicated, but the following two examples tell that they are also common in statistical modeling.

Example 2.10 *(Symmetric α-stable Lévy distributions).* A distribution is said to be symmetric α-stable Lévy distributions [14] if it is symmetric and its Fourier transform is represented in the form $e^{-\gamma^\alpha|\xi|^\alpha}$, with $\gamma > 0$ and $0 < \alpha \leq 2$. Obviously, Gaussian distribution with mean zero is a special case with $\alpha = 2$. Cauchy distribution with median zero is another special case with $\alpha = 1$. Every distribution in this set is unimodal

Please cite this article in press as: J. Fan et al., Consistency analysis of an empirical minimum error entropy algorithm, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.12.005

[11]. If we choose a subset of these distributions with $\gamma \le C$ (C is a constant), then the Fourier transform is positive and $\exists c_0 = 1/C$ and $C_0 = e^{-1}$ such that $\forall \xi \in [-c_0, c_0]$, $\widehat{p_{\epsilon|X}}(\xi) \ge C_0$.

Example 2.11 *(Linnik distributions).* A Linnik distribution is also referred to as a symmetric geometric stable distribution [10]. A distribution is said to be Linnik distribution if it is symmetric and its Fourier transform is represented in the form $\frac{1}{1+\lambda^\alpha|\xi|^\alpha}$, with $\lambda > 0$ and $0 < \alpha \le 2$. Obviously, Laplace distribution with mean zero is a special case with $\alpha = 2$. Every distribution in this set is unimodal [11]. If we choose a subset of these distributions with $\lambda \le C$ (C is a constant), then the Fourier transform is positive and $\exists c_0 = 1/C$ and $C_0 = \frac{1}{2}$ such that $\forall \xi \in [-c_0, c_0]$, $\widehat{p_{\epsilon|X}}(\xi) \ge C_0$.

Corresponding to the definition of the empirical Rényi's entropy $\mathscr{R}_{\mathbf{z}}(f)$, after removing the logarithm, we define information error of a measurable function $f : \mathcal{X} \to \mathbb{R}$ as

$$\mathcal{E}_h(f) = -\int_{\mathbb{R}}\int_{\mathbb{R}} G_h(e - e')p_E(e)p_E(e')\,de\,de'$$
$$= -\int_{\mathcal{Z}}\int_{\mathcal{Z}} G_h\big((y - f(x)) - (y' - f(x'))\big)d\rho(x, y)d\rho(x', y').$$

Theorem 2.12. *If ρ belongs to $\mathcal{P}_1$, then $f^* + b$ is a minimizer of $\mathcal{E}_h(f)$ for any constant b and any fixed $h > 0$. Moreover, we have $f_{\mathbf{z}} + \mathbf{E}_x[f^*(x) - f_{\mathbf{z}}(x)]$ converges to f^* in probability. Convergence rate of order $O(\sqrt{\log(2/\delta)}n^{-\frac{1}{2}})$ can be obtained with confidence $1 - \delta$ for $\|f_{\mathbf{z}} + \mathbf{E}_x[f^*(x) - f_{\mathbf{z}}(x)] - f^*\|^2_{L^2_{\rho_X}}$.*

Theorem 2.13. *If ρ belongs to $\mathcal{P}_2$, then there exists some $h_{\rho,\mathcal{H}} > 0$ such that $f^* + b$ is a minimizer of $\mathcal{E}_h(f)$ for any fixed $h > h_{\rho,\mathcal{H}}$ and constant b. Also $f_{\mathbf{z}} + \mathbf{E}_x[f^*(x) - f_{\mathbf{z}}(x)]$ converges to f^* in probability. Convergence rate of order $O(\sqrt{\log(2/\delta)}n^{-\frac{1}{2}})$ can be obtained with confidence $1 - \delta$ for $\|f_{\mathbf{z}} + \mathbf{E}_x[f^*(x) - f_{\mathbf{z}}(x)] - f^*\|^2_{L^2_{\rho_X}}$.*

3. Error entropy consistency

In this section we will prove that $\mathscr{R}(f_{\mathbf{z}})$ converges to $\mathscr{R}^*$ in probability when $h = h(n)$ tends to zero slowly satisfying (2.1). Several useful lemmas are needed to prove our first main result (Theorem 2.3).

Lemma 3.1. *For any measurable function f on $\mathcal{X}$, the probability density function for the error variable $E = Y - f(X)$ is given as*

$$p_E(e) = \int_{\mathcal{X}} p_{\epsilon|X}\big(e + f(x) - f^*(x)|x\big)d\rho_X(x). \tag{3.1}$$

As a result, we have $|p_E(e)| \le M$ for every $e \in \mathbb{R}$.

Proof. Eq. (3.1) follows from the fact that

$$\epsilon = Y - f^*(X) = E + f(X) - f^*(X).$$

The inequality $|p_E(e)| \le M$ follows from the assumption $|p_{\epsilon|X}(t)| \le M$. □

Denote by B_L and B_U the lower bound and upper bound of $\mathbf{E}[p_E]$ over $\mathcal{H}$, i.e.,

$$B_L = \inf_{f\in\mathcal{H}} \int_{\mathbb{R}} \big(p_E(e)\big)^2 de \quad \text{and} \quad B_U = \sup_{f\in\mathcal{H}} \int_{\mathbb{R}} \big(p_E(e)\big)^2 de.$$

Lemma 3.2. *We have* $0 < B_L$ *and* $B_U \leq M_p$*.*

Proof. Since $\int_{\mathcal{X}} \int_{-\infty}^{\infty} p_{\epsilon|X}(t|x)dtd\rho_X(x) = 1$, there is some constant $0 < A < +\infty$ such that

$$a = \int_{\mathcal{X}} \int_{-A}^{A} p_{\epsilon|X}(t|x)dtd\rho_X(x) > \frac{1}{2}.$$

For any $f \in \mathcal{H}$, by the fact $|f| \leq M$ and $|f^*| \leq M$, it is easy to check from (3.1) that

$$\begin{aligned}\int_{-(A+2M)}^{A+2M} p_E(e)de &= \int_{\mathcal{X}} \int_{-(A+2M)}^{A+2M} p_{\epsilon|X}\big(e + f(x) - f^*(x)|x\big)ded\rho_X(x) \\ &= \int_{\mathcal{X}} \int_{-(A+2M)+f(x)-f^*(x)}^{A+2M+f(x)-f^*(x)} p_{\epsilon|X}(t|x)dtd\rho_X(x) \\ &\geq \int_{\mathcal{X}} \int_{-A}^{A} p_{\epsilon|X}(t|x)dtd\rho_X(x) = a.\end{aligned}$$

Then by the Schwartz inequality we have

$$\begin{aligned}a &\leq \int_{-(A+2M)}^{A+2M} p_E(e)de \leq \Bigg(\int_{-(A+2M)}^{A+2M} \big(p_E(e)\big)^2 de\Bigg)^{\frac{1}{2}} \Bigg(\int_{-(A+2M)}^{A+2M} de\Bigg)^{\frac{1}{2}} \\ &\leq \sqrt{2A+4M}\Bigg(\int_{\mathbb{R}} \big(p_E(e)\big)^2 de\Bigg)^{\frac{1}{2}}.\end{aligned}$$

This gives

$$\int_{\mathbb{R}} \big(p_E(e)\big)^2 de \geq \frac{a^2}{2A+4M} \geq \frac{1}{8A+16M}$$

for any $f \in \mathcal{H}$. Hence $B_L \geq \frac{1}{8A+16M} > 0$.

The second inequality follows from the fact that p_E is a density function and uniformly bounded by M_p. This proves Lemma 3.2. □

It helps our analysis to remove the logarithm from the Rényi's entropy (1.2) and define

$$V(f) = -\mathbf{E}[p_E] = -\int \big(p_E(e)\big)^2 de. \tag{3.2}$$

Then $\mathscr{R}(f) = -\log(-V(f))$. Since $-\log(-t)$ is strictly increasing for $t \leq 0$, minimizing $\mathscr{R}(f)$ is equivalent to minimizing $V(f)$. As a result, their minimizers are the same. Denote $V^* = \inf_{f:\mathcal{X}\to\mathbb{R}} V(f)$. Then $V^*(f) = -\log(-\mathscr{R}^*)$, and we have the following lemma.

Lemma 3.3. *For any* $f \in \mathcal{H}$ *we have*

$$\frac{1}{B_U}\big(V(f) - V^*\big) \leq \mathscr{R}(f) - \mathscr{R}^* \leq \frac{1}{B_L}\big(V(f) - V^*\big).$$

Proof. Since the derivative of the function $-\log(-t)$ is $-\frac{1}{t}$, by the mean value theorem we get

$$\mathscr{R}(f) - \mathscr{R}^* = \mathscr{R}(f) - \mathscr{R}(f^*_{\mathscr{R}}) = -\log(-V(f)) - [-\log(-V(f^*_{\mathscr{R}}))] = -\frac{1}{\xi}(V(f) - V(f^*_{\mathscr{R}}))$$

for some $\xi \in [V(f^*_{\mathscr{R}}), V(f)] \subset [-B_U, -B_L]$. This leads to the conclusion. □

From Lemma 3.3 we see that, to prove Theorem 2.3, it is equivalent to prove the convergence of $V(f_{\mathbf{z}})$ to V^*. To this end we define an empirical version of the generalization error $\mathcal{E}_{h,\mathbf{z}}(f)$ as

$$\mathcal{E}_{h,\mathbf{z}}(f) = -\frac{1}{n^2}\sum_{i,j=1}^{n} G_h(e_i - e_j) = -\frac{1}{n^2}\sum_{i,j=1}^{n} G_h\big((y_i - f(x_i)) - (y_j - f(x_j))\big).$$

Again we see the equivalence between minimizing $\mathscr{R}_{\mathbf{z}}(f)$ and minimizing $\mathcal{E}_{h,\mathbf{z}}(f)$. So $f_{\mathbf{z}}$ is also a minimizer of $\mathcal{E}_{h,\mathbf{z}}$ over the hypothesis class $\mathcal{H}$. We then can bound $V(f_{\mathbf{z}}) - V^*$ by an error decomposition as

$$\begin{aligned} V(f_{\mathbf{z}}) - V^* &= \big(V(f_{\mathbf{z}}) - \mathcal{E}_{h,\mathbf{z}}(f_{\mathbf{z}})\big) + \big(\mathcal{E}_{h,\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{h,\mathbf{z}}(f^*_{\mathscr{R}})\big) + \big(\mathcal{E}_{h,\mathbf{z}}(f^*_{\mathscr{R}}) - V(f^*_{\mathscr{R}})\big) \\ &\le 2\sup_{f\in\mathcal{H}}\big|\mathcal{E}_{h,\mathbf{z}}(f) - V(f)\big| \le 2\mathcal{S}_{\mathbf{z}} + 2\mathcal{A}_h. \end{aligned}$$

where $\mathcal{S}_{\mathbf{z}}$ is called the sample error defined by $\mathcal{S}_{\mathbf{z}} = \sup_{f\in\mathcal{H}}|\mathcal{E}_{h,\mathbf{z}}(f) - \mathcal{E}_h(f)|$ and $\mathcal{A}_h$ is called approximation error defined by $\sup_{f\in\mathcal{H}}|\mathcal{E}_h(f) - V(f)|$.

The sample error $\mathcal{S}_{\mathbf{z}}$ depends on the sample, and can be estimated by the following proposition.

Proposition 3.4. *There is a constant $B > 0$ depending on M, c and s (in Definition 2.2) such that for every $\varepsilon_1 > 0$,*

$$\mathbf{P}\left(\mathcal{S}_{\mathbf{z}} > \varepsilon_1 + \frac{B}{h^2\sqrt{n}}\right) \le \exp(-2nh^2\varepsilon_1^2).$$

This proposition implies that $\mathcal{S}_{\mathbf{z}}$ is bounded by $O(\frac{1}{h^2\sqrt{n}} + \frac{1}{h\sqrt{n}})$ with large probability. The proof of this proposition is long and complicated. But it is rather standard in the context of learning theory. So we leave it in Appendix A where the constant B will be given explicitly.

The approximation error is small when h tends to zero, as shown in next proposition.

Proposition 3.5. *We have $\lim_{h\to 0}\mathcal{A}_h = 0$. If the derivative of $p_{\epsilon|X}$ is uniformly bounded by a constant M', then $\mathcal{A}_h \le M'h$.*

Proof. Since $G_h(t) = \frac{1}{h}G_1(\frac{t}{h})$, by changing the variable e' to $\tau = \frac{e-e'}{h}$, we have

$$\begin{aligned} \mathcal{A}_h &= \sup_{f\in\mathcal{H}}\left|\int_{\mathbb{R}}\int_{\mathbb{R}}\frac{1}{h}G_1\left(\frac{e-e'}{h}\right)p_E(e)p_E(e')de\,de' - \int_{\mathbb{R}}(p_E(e))^2 de\right| \\ &= \sup_{f\in\mathcal{H}}\left|\int_{\mathbb{R}}\int_{\mathbb{R}} G_1(\tau)p_E(e-\tau h)d\tau\, p_E(e)de - \int_{\mathbb{R}}(p_E(e))^2 de\right| \end{aligned}$$

But $\int_{\mathbb{R}} G_1(\tau)d\tau = 1$, we see from (3.1) that

$$\mathcal{A}_h = \sup_{f\in\mathcal{H}}\left|\int_{\mathbb{R}} p_{_E}(e)\int_{\mathbb{R}} G_1(\tau)\big(p_{_E}(e-\tau h) - p_{_E}(e)\big)d\tau de\right|$$

$$\leq \sup_{f\in\mathcal{H}}\int_{\mathbb{R}} p_{_E}(e)\int_{\mathbb{R}} G_1(\tau)\int_{\mathcal{X}}\big|p_{\epsilon|X}\big(e-\tau h+f(x)-f^*(x)|x\big)$$

$$- p_{\epsilon|X}\big(e+f(x)-f^*(x)|x\big)\big|d\rho_{_X}(x)d\tau de. \tag{3.3}$$

It follows form Lebesgue's Dominated Convergence Theorem that $\lim_{h\to 0}\mathcal{A}_h = 0$.

If $|p'_{\epsilon|X}| \leq M'$ uniformly for an M', we have

$$\big|p_{\epsilon|X}\big(e-\tau h+f(x)-f^*(x)|x\big) - p_{\epsilon|X}\big(e+f(x)-f^*(x)|x\big)\big| \leq M'|\tau|h.$$

Then from (3.3), we find

$$\mathcal{A}_h \leq \sup_{f\in\mathcal{H}}\int_{\mathbb{R}} p_{_E}(e)de\int_{\mathbb{R}} G_1(\tau)|\tau|d\tau M'h = \frac{2M'}{\sqrt{2\pi}}h \leq M'h.$$

This proves Proposition 3.5. □

We are in a position to prove our first main result Theorem 2.3.

Proof of Theorem 2.3. Let $0 < \delta < 1$. By take $\varepsilon_1 > 0$ such that $\exp(-2nh^2\varepsilon_1^2) = \delta$, i.e., $\varepsilon_1 = \sqrt{\frac{\log(1/\delta)}{2nh^2}}$, we know from Proposition 3.4 that with probability at least $1-\delta$,

$$\mathcal{S}_{\mathbf{z}} \leq \varepsilon_1 + \frac{B}{h^2\sqrt{n}} = \frac{1}{h^2\sqrt{n}}\big(B + \sqrt{\log(1/\delta)}h\big).$$

To prove the first statement, we apply assumption (2.1). For any $\varepsilon > 0$, there exists some $N_1 \in \mathbb{N}$ such that $(B+1)\frac{1}{h^2\sqrt{n}} < \frac{\varepsilon}{2}$ and $\sqrt{\log(1/\delta)}h \leq 1$ whenever $n \geq N_1$. It follows that with probability at least $1-\delta$, $\mathcal{S}_{\mathbf{z}} < \frac{\varepsilon}{2}$. By Proposition 3.5 and $\lim_{n\to\infty} h(n) = 0$, there exists some $N_2 \in \mathbb{N}$ such that $\mathcal{A}_h \leq \frac{\varepsilon}{2}$ whenever $n \geq N_2$. Combining the above two parts for $n \geq \max\{N_1, N_2\}$, we have with probability at least $1-\delta$,

$$V(f_{\mathbf{z}}) - V^* \leq 2\mathcal{S}_{\mathbf{z}} + 2\mathcal{A}_h \leq 2\varepsilon,$$

which implies by Lemma 3.3,

$$\mathscr{R}(f_{\mathbf{z}}) - \mathscr{R}^* \leq \frac{2}{B_L}\varepsilon.$$

Hence the probability of the event $\mathscr{R}(f_{\mathbf{z}}) - \mathscr{R}^* \geq \frac{2}{B_L}\varepsilon$ is at most δ. This proves the first statement of Theorem 2.3.

To prove the second statement, we apply the second part of Proposition 3.5. Then with probability at least $1-\delta$, we have

$$\mathscr{R}(f_{\mathbf{z}}) - \mathscr{R}^* \leq \frac{1}{B_L}\big(V(f_{\mathbf{z}}) - V^*\big) \leq \frac{2}{B_L}\left(\frac{1}{h^2\sqrt{n}}\big(B + \sqrt{\log(1/\delta)}h\big) + M'h\right).$$

Thus, if $C_1' n^{-\frac{1}{6}} \leq h(n) \leq C_2' n^{-\frac{1}{6}}$ for some constants $0 < C_1' \leq C_2'$, we have with probability at least $1-\delta$,

$$\mathscr{R}(f_{\mathbf{z}}) - \mathscr{R}^* \le \frac{1}{B_L}\big(V(f_{\mathbf{z}}) - V^*\big) \le \frac{2}{B_L}\left(\frac{1}{(C_1')^2}\big(B + C_2'\sqrt{\log(1/\delta)}\big) + M'C_2'\right)n^{-\frac{1}{6}}.$$

Then the desired convergence rate is obtained. The proof of Theorem 2.3 is complete. □

4. Regression consistency for homoskedastic models

In this section we prove the regression consistency for homoskedastic models stated in Theorem 2.5. Under the homoskedasticity assumption, the noise ϵ is independent of x, so throughout this section we will simply use p_ϵ to denote the density function for the noise. Also, we use the notations $E = E(f) = Y - f(X)$ and $E^* = Y - f^*(X)$.

The probability density function of the random variable $E = Y - f(X)$ is given by

$$p_E(e) = \int_{\mathcal{X}} p_\epsilon\big(e + f(x) - f^*(x)\big) d\rho_X(x).$$

Then

$$\int_{\mathbb{R}} \big(p_E(e)\big)^2 de = \int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathbb{R}} p_\epsilon\big(e + f(x) - f^*(x)\big)p_\epsilon\big(e + f(u) - f^*(u)\big) de d\rho_X(x) d\rho_X(u).$$

We apply the Planchel formula and find

$$\int_{\mathbb{R}} p_\epsilon\big(e + f(x) - f^*(x)\big)p_\epsilon\big(e + f(u) - f^*(u)\big) de = \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{p_\epsilon}(\xi)\mathbf{e}^{\mathrm{i}\xi(f(x)-f^*(x))}\overline{\widehat{p_\epsilon}(\xi)\mathbf{e}^{\mathrm{i}\xi(f(u)-f^*(u))}} d\xi.$$

It follows that

$$\int_{\mathbb{R}} \big(p_E(e)\big)^2 de = \frac{1}{2\pi}\int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathbb{R}} |\widehat{p_\epsilon}(\xi)|^2 \mathbf{e}^{\mathrm{i}\xi(f(x)-f^*(x)-f(u)+f^*(u))} d\xi d\rho_X(x) d\rho_X(u).$$

This is obviously maximized when $f = f^*$ since $|\mathbf{e}^{\mathrm{i}\xi t}| \le 1$. This proves that f^* is a minimizer of $V(f)$ and $\mathscr{R}(f)$. Since $V(f)$ and $\mathscr{R}(f)$ are invariant with respect to constant translates, we have proved part (i) of Theorem 2.5.

To prove part (ii), we study the excess quantity $V(f) - V(f^*)$ and express it as

$$\begin{aligned} V(f) - V\big(f^*\big) &= \int_{\mathbb{R}} \big(p_{E^*}(e)\big)^2 de - \int_{\mathbb{R}} \big(p_E(e)\big)^2 de \\ &= \frac{1}{2\pi}\int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathbb{R}} \big|\widehat{p_\epsilon}(\xi)\big|^2\big(1 - \mathbf{e}^{\mathrm{i}\xi(f(x)-f^*(x)-f(u)+f^*(u))}\big) d\xi d\rho_X(x) d\rho_X(u) \\ &= \frac{1}{2\pi}\int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathbb{R}} \big|\widehat{p_\epsilon}(\xi)\big|^2 2\sin^2\frac{\xi(f(x)-f^*(x)-f(u)+f^*(u))}{2} d\xi d\rho_X(x) d\rho_X(u) \end{aligned}$$

where the last equality follows from the fact that $V(f) - V(f^*)$ is real and hence equals to its real part.

As both f and f^* take values on $[-M, M]$, we know that $|f(x) - f^*(x) - f(u) + f^*(u)| \le 4M$ for any $x, u \in \mathcal{X}$. So when $|\xi| \le \frac{\pi}{4M}$, we have

Please cite this article in press as: J. Fan et al., Consistency analysis of an empirical minimum error entropy algorithm, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.12.005

$$\left|\frac{\xi(f(x)-f^*(x)-f(u)+f^*(u))}{2}\right| \leq \frac{\pi}{2}$$

and

$$\left|\sin\frac{\xi(f(x)-f^*(x)-f(u)+f^*(u))}{2}\right| \geq \frac{2}{\pi}\left|\frac{\xi(f(x)-f^*(x)-f(u)+f^*(u))}{2}\right|.$$

Observe that the integrand in the expression of $V(f)-V(f^*)$ is nonnegative and

$$\begin{aligned}&\int_{\mathbb{R}} |\widehat{p_\epsilon}(\xi)|^2 2\sin^2\frac{\xi(f(x)-f^*(x)-f(u)+f^*(u))}{2}d\xi \\ &\quad\geq \int_{|\xi|\leq\frac{\pi}{4M}} \left|\widehat{p_\epsilon}(\xi)\right|^2 2\sin^2\frac{\xi(f(x)-f^*(x)-f(u)+f^*(u))}{2}d\xi \\ &\quad\geq \int_{|\xi|\leq\frac{\pi}{4M}} \left|\widehat{p_\epsilon}(\xi)\right|^2 \frac{2}{\pi^2}\xi^2\big(f(x)-f^*(x)-f(u)+f^*(u)\big)^2 d\xi.\end{aligned}$$

Therefore,

$$V(f)-V(f^*) \geq \frac{1}{\pi^3}\int_{|\xi|\leq\frac{\pi}{4M}} \xi^2\left|\widehat{p_\epsilon}(\xi)\right|^2 d\xi \int_{\mathcal{X}}\int_{\mathcal{X}} \big(f(x)-f^*(x)-f(u)+f^*(u)\big)^2 d\rho_X(x)d\rho_X(u).$$

It was shown in [8] that

$$\int_{\mathcal{X}}\int_{\mathcal{X}} \big(f(x)-f^*(x)-f(u)+f^*(u)\big)^2 d\rho_X(x)d\rho_X(u) = 2\left\|f-f^*+\mathbf{E}_x\left[f^*(x)-f(x)\right]\right\|^2_{L^2_{\rho_X}}. \quad (4.1)$$

So we have

$$V(f)-V(f^*) \geq \left(\frac{2}{\pi^3}\int_{|\xi|\leq\frac{\pi}{4M}} \xi^2\left|\widehat{p_\epsilon}(\xi)\right|^2 d\xi\right)\left\|f-f^*+\mathbf{E}_x\left[f^*(x)-f(x)\right]\right\|^2_{L^2_{\rho_X}}.$$

Since the probability density function p_ϵ is integrable, its Fourier transform $\widehat{p_\epsilon}$ is continuous. This together with $\widehat{p_\epsilon}(0)=1$ ensures that $\widehat{p_\epsilon}(\xi)$ is nonzero over a small interval around 0. As a result $\xi^2|\widehat{p_\epsilon}(\xi)|^2$ is not identically zero on $[-\frac{\pi}{4M}, \frac{\pi}{4M}]$. Hence the constant

$$c = \int_{|\xi|\leq\frac{\pi}{4M}} \xi^2\left|\widehat{p_\epsilon}(\xi)\right|^2 d\xi$$

is positive and the conclusion in (ii) is proved by taking $C=\frac{\pi^3 B_U}{2c}$ and applying Lemma 3.3.

Parts (iii) and (iv) are easy corollaries of part (ii) and Theorem 2.3. This finishes the proof of Theorem 2.5.

5. Incoincidence between error entropy consistency and regression consistency

In the previous section we proved that for homoskedastic models the error entropy consistency implies the regression consistency. But for heteroskedastic models, this is not necessarily true. Here we present a counter-example to show this incoincidence between the two types of consistency.

Let $\mathbf{1}_{(\cdot)}$ denote the indicator function on a set specified by the subscript.

Example 5.1. Let $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 = [0, \frac{1}{2}] \cup [1, \frac{3}{2}]$ and ρ_X be uniform on $\mathcal{X}$ (so that $d\rho_X = dx$). The conditional distribution of $\epsilon|X$ is uniform on $[-\frac{1}{2}, \frac{1}{2}]$ if $x \in [0, \frac{1}{2}]$ and uniform on $[-\frac{3}{2}, -\frac{1}{2}] \cup [\frac{1}{2}, \frac{3}{2}]$ if $x \in [1, \frac{3}{2}]$. Then

(i) a function $f^*_{\mathscr{R}} : \mathcal{X} \to \mathbb{R}$ is a minimizer of $\mathscr{R}(f)$ if and only if there are two constants f_1, f_2 with $|f_1 - f_2| = 1$ such that $f^*_{\mathscr{R}} = f_1 \mathbf{1}_{\mathcal{X}_1} + f_2 \mathbf{1}_{\mathcal{X}_2}$;
(ii) $\mathscr{R}^* = -\log(\frac{5}{8})$ and $\mathscr{R}(f^*) = -\log(\frac{3}{8})$. So the regression function f^* is not a minimizer of the error entropy functional $\mathscr{R}(f)$;
(iii) let $\mathcal{F}^*_{\mathscr{R}}$ denote the set of all minimizers. There is an a constant C' depending on $\mathcal{H}$ and M such that for any measurable function f bounded by M,

$$\min_{f^*_{\mathscr{R}} \in \mathcal{F}^*_{\mathscr{R}}} \|f - f^*_{\mathscr{R}}\|^2_{L^2_{\rho_X}} \le C'\big(\mathscr{R}(f) - \mathscr{R}^*\big);$$

(iv) if the error entropy consistency is true, then there holds

$$\min_{f^*_{\mathscr{R}} \in \mathcal{F}^*_{\mathscr{R}}} \|f_{\mathbf{z}} - f^*_{\mathscr{R}}\|_{L^2_{\rho_X}} \longrightarrow 0 \quad \text{and} \quad \min_{b \in \mathbb{R}} \|f_{\mathbf{z}} + b - f^*\|_{L^2_{\rho_X}} \longrightarrow \frac{1}{2}$$

in probability. As a result, the regression consistency cannot be true.

Proof. Without loss of generality we may assume $M \ge 1$ in this example.

Denote $p_1(\epsilon) = p_{\epsilon|X}(\epsilon|x)$ for $x \in \mathcal{X}_1$ and $p_2(\epsilon) = p_{\epsilon|X}(\epsilon|x)$ for $x \in \mathcal{X}_2$. By Lemma 3.1, the probability density function of $E = Y - f(X)$ is given by

$$p_{\scriptscriptstyle E}(e) = \int_{\mathcal{X}} p_{\epsilon|X}\big(e + f(x) - f^*(x)|x\big) d\rho_X(x) = \sum_{j=1}^{2} \int_{\mathcal{X}_j} p_j\big(e + f(x) - f^*(x)\big) dx.$$

So we have

$$\int_{\mathbb{R}} \big(p_{\scriptscriptstyle E}(e)\big)^2 de = \sum_{j,k=1}^{2} \int_{\mathcal{X}_j} \int_{\mathcal{X}_k} \int_{\mathbb{R}} p_j\big(e + f(x) - f^*(x)\big) p_k\big(e + f(u) - f^*(u)\big) de d\rho_X(x) d\rho_X(u).$$

By the Planchel formula,

$$\int_{\mathbb{R}} \big(p_{\scriptscriptstyle E}(e)\big)^2 de = \frac{1}{2\pi} \sum_{j,k=1}^{2} \int_{\mathcal{X}_j} \int_{\mathcal{X}_k} \int_{\mathbb{R}} \widehat{p_j}(\xi) \overline{\widehat{p_k}(\xi)} \mathbf{e}^{\mathrm{i}\xi(f(x) - f^*(x) - f(u) + f^*(u))} d\xi d\rho_X(x) d\rho_X(u).$$

Let $p^* = \mathbf{1}_{[-\frac{1}{2}, \frac{1}{2}]}$ be the density function of the uniform distribution on $[-\frac{1}{2}, \frac{1}{2}]$. Then we have $p_1 = p^*$ and $p_2(e) = \frac{p^*(e+1) + p^*(e-1)}{2}$ which yields

$$\widehat{p_2}(\xi) = \frac{\mathbf{e}^{-\mathrm{i}\xi} + \mathbf{e}^{\mathrm{i}\xi}}{2} \widehat{p^*}(\xi) = \widehat{p^*}(\xi) \cos\xi.$$

These together with $f^* \equiv 0$ allow us to write

$$V(f) = -\int_{\mathbb{R}} \big(p_{\scriptscriptstyle E}(e)\big)^2 de = V_{11}(f) + V_{22}(f) + V_{12}(f), \tag{5.1}$$

where

$$V_{11}(f) = -\frac{1}{2\pi}\int_{\mathcal{X}_1}\int_{\mathcal{X}_1}\int_{\mathbb{R}}\left|\widehat{p^*}(\xi)\right|^2 \mathbf{e}^{\mathrm{i}\xi(f(x)-f(u))}d\xi d\rho_X(x)d\rho_X(u),$$

$$V_{22}(f) = -\frac{1}{2\pi}\int_{\mathcal{X}_2}\int_{\mathcal{X}_2}\int_{\mathbb{R}}\cos^2\xi\left|\widehat{p^*}(\xi)\right|^2 \mathbf{e}^{\mathrm{i}\xi(f(x)-f(u))}d\xi d\rho_X(x)d\rho_X(u),$$

$$V_{12}(f) = -\frac{1}{\pi}\int_{\mathcal{X}_1}\int_{\mathcal{X}_2}\int_{\mathbb{R}}\left|\widehat{p^*}(\xi)\right|^2 \cos\xi\cos\big(\xi\big(f(x)-f(u)\big)\big)d\xi d\rho_X(x)d\rho_X(u).$$

Recall the following identity from Fourier analysis (see e.g. [9])

$$\sum_{\ell\in\mathbb{Z}}\widehat{p^*}(\xi+2\ell\pi)\overline{\widehat{p^*(\cdot-b)}(\xi+2\ell\pi)} = \sum_{\ell\in\mathbb{Z}}\big\langle p^*(\cdot-\ell), p^*(\cdot-b)\big\rangle_{L^2(\mathbb{R})}\mathbf{e}^{\mathrm{i}\ell\xi}, \quad \forall\xi, b\in\mathbb{R}. \tag{5.2}$$

In particular, with $b=0$, since the integer translates of p^* are orthogonal, there hold $\sum_{\ell\in\mathbb{Z}}|\widehat{p^*}(\xi+2\ell\pi)|^2\equiv 1$ and

$$\int_{\mathbb{R}}\left|\widehat{p^*}(\xi)\right|^2\cos^j\xi d\xi = \int_{[-\pi,\pi)}\sum_{\ell\in\mathbb{Z}}\left|\widehat{p^*}(\xi+2\ell\pi)\right|^2\cos^j\xi d\xi = \begin{cases} 0, & \text{if } j=1,\\ 2\pi, & \text{if } j=0,\\ \pi, & \text{if } j=2.\end{cases}$$

For $V_{11}(f)$, notice the real analyticity of the function $\widehat{p^*}(\xi) = \frac{2\sin(\xi/2)}{\xi}$ and the identity

$$\int_{\mathbb{R}}\left|\widehat{p^*}(\xi)\right|^2\mathbf{e}^{\mathrm{i}\xi(f(x)-f(u))}d\xi = \int_{\mathbb{R}}\left|\widehat{p^*}(\xi)\right|^2\cos\big(\xi\big(f(x)-f(u)\big)\big)d\xi.$$

We see that $V_{11}(f)$ is minimized if and only if $f(x)=f(u)$ for any $x,u\in\mathcal{X}_1$. In this case, f is a constant on $\mathcal{X}_1$, denoted as f_1, and the minimum value of $V_{11}(f)$ equals

$$V_{11}^* := -\big(\rho_X(\mathcal{X}_1)\big)^2 = -\frac{1}{4}.$$

Moreover, if a measurable function satisfies $f(x)\in[-M,M]$ for every $x\in\mathcal{X}_1$, we have

$$\begin{aligned}
V_{11}(f)-V_{11}^* &= \frac{1}{2\pi}\int_{\mathcal{X}_1}\int_{\mathcal{X}_1}\int_{\mathbb{R}}\left|\widehat{p^*}(\xi)\right|^2\big(1-\cos\big(\xi\big(f(x)-f(u)\big)\big)\big)d\xi d\rho_X(x)d\rho_X(u)\\
&= \frac{1}{2\pi}\int_{\mathcal{X}_1}\int_{\mathcal{X}_1}\int_{\mathbb{R}}\left|\widehat{p^*}(\xi)\right|^2 2\sin^2\left(\frac{\xi(f(x)-f(u)}{2}\right)d\xi d\rho_X(x)d\rho_X(u)\\
&\geq \frac{1}{2\pi}\int_{\mathcal{X}_1}\int_{\mathcal{X}_1}\int_{|\xi|\leq\frac{\pi}{4M}}\left|\widehat{p^*}(\xi)\right|^2 2\left(\frac{2}{\pi}\frac{\xi(f(x)-f(u))}{2}\right)^2 d\xi d\rho_X(x)d\rho_X(u)\\
&\geq \frac{1}{24\pi^2M^3}\int_{\mathcal{X}_1}\int_{\mathcal{X}_1}\big(f(x)-f(u)\big)^2 d\rho_X(x)d\rho_X(u)\\
&= \frac{1}{12\pi^2M^3}\|f-m_{f,\mathcal{X}_1}\|^2_{L^2_{\rho_X}(\mathcal{X}_1)}
\end{aligned} \tag{5.3}$$

where

$$m_{f,\mathcal{X}_j} = \frac{\mathbf{E}[f\mathbf{1}_{\mathcal{X}_j}]}{\rho_X(\mathcal{X}_j)} = \frac{1}{\rho_X(\mathcal{X}_j)} \int_{\mathcal{X}_j} f(x) d\rho_X(x)$$

denotes the mean of f on $\mathcal{X}_j$.

Similarly, $V_{22}(f)$ is minimized if and only if f is constant on $\mathcal{X}_2$, which will be denoted as f_2, and the corresponding minimum value equals

$$V_{22}^* := -\frac{1}{2}\big(\rho_X(\mathcal{X}_2)\big)^2 = -\frac{1}{8}.$$

Again, if a measurable function satisfies $f(x) \in [-M, M]$ for every $x \in \mathcal{X}_2$, we have

$$V_{22}(f) - V_{22}^* \geq \frac{1}{24\pi^2 M^3} \|f - m_{f,\mathcal{X}_2}\|^2_{L^2_{\rho_X}(\mathcal{X}_2)}. \tag{5.4}$$

For $V_{12}(f)$, we express it as

$$V_{12}(f) = -\frac{1}{4\pi} \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \int_{\mathbb{R}} \big|\widehat{p^*}(\xi)\big|^2 \big(\mathbf{e}^{\mathrm{i}\xi} + \mathbf{e}^{-\mathrm{i}\xi}\big)\big(\mathbf{e}^{\mathrm{i}\xi(f(x)-f(u))} + \mathbf{e}^{-\mathrm{i}\xi(f(x)-f(u))}\big) d\xi d\rho_X(x) d\rho_X(u).$$

Write $f(x) - f(u)$ as $k_{f,x,u} + b_{f,x,u}$ with $k_{f,x,u} \in \mathbb{Z}$ being the integer part of the real number of $f(x) - f(u)$ and $b_{f,x,u} \in [0, 1)$. We have

$$\begin{aligned}
&\int_{\mathbb{R}} \big|\widehat{p^*}(\xi)\big|^2 \big(\mathbf{e}^{\mathrm{i}\xi} + \mathbf{e}^{-\mathrm{i}\xi}\big)\mathbf{e}^{\mathrm{i}\xi(f(x)-f(u))} d\xi \\
&= \int_{\mathbb{R}} \widehat{p^*}(\xi)\overline{\widehat{p^*}(\xi)\mathbf{e}^{-\mathrm{i}\xi b_{f,x,u}}}\big(\mathbf{e}^{\mathrm{i}\xi(k_{f,x,u}+1)} + \mathbf{e}^{\mathrm{i}\xi(k_{f,x,u}-1)}\big) d\xi \\
&= \int_{\mathbb{R}} \widehat{p^*}(\xi)\overline{p^*(\widehat{\cdot - b_{f,x,u}})(\xi)}\big(\mathbf{e}^{\mathrm{i}\xi(k_{f,x,u}+1)} + \mathbf{e}^{\mathrm{i}\xi(k_{f,x,u}-1)}\big) d\xi \\
&= \int_{[-\pi,\pi)} \Big\{\sum_{\ell\in\mathbb{Z}} \widehat{p^*}(\xi + 2\ell\pi)\overline{p^*(\widehat{\cdot - b_{f,x,u}})(\xi + 2\ell\pi)}\Big\}\big(\mathbf{e}^{\mathrm{i}\xi(k_{f,x,u}+1)} + \mathbf{e}^{\mathrm{i}\xi(k_{f,x,u}-1)}\big) d\xi \\
&= \int_{[-\pi,\pi)} \Big\{\sum_{\ell\in\mathbb{Z}} \big\langle p^*(\cdot - \ell), p^*(\cdot - b_{f,x,u})\big\rangle_{L^2(\mathbb{R})} e^{\mathrm{i}\ell\xi}\Big\}\big(\mathbf{e}^{\mathrm{i}\xi(k_{f,x,u}+1)} + \mathbf{e}^{\mathrm{i}\xi(k_{f,x,u}-1)}\big) d\xi,
\end{aligned}$$

where we have used (5.2) in the last step. Since $b_{f,x,u} \in [0, 1)$, we see easily that

$$\big\langle p^*(\cdot - \ell), p^*(\cdot - b_{f,x,u})\big\rangle_{L^2(\mathbb{R})} = \begin{cases} 1 - b_{f,x,u}, & \text{if } \ell = 0, \\ b_{f,x,u}, & \text{if } \ell = 1, \\ 0, & \text{if } \ell \in \mathbb{Z} \setminus \{0, 1\}. \end{cases}$$

Hence

$$\begin{aligned}
&\int_{\mathbb{R}} \big|\widehat{p^*}(\xi)\big|^2 \big(\mathbf{e}^{\mathrm{i}\xi} + \mathbf{e}^{-\mathrm{i}\xi}\big)\mathbf{e}^{\mathrm{i}\xi(f(x)-f(u))} d\xi \\
&= \int_{[-\pi,\pi)} \big(1 - b_{f,x,u} + b_{f,x,u}\mathbf{e}^{\mathrm{i}\xi}\big)\big(\mathbf{e}^{\mathrm{i}\xi(k_{f,x,u}+1)} + \mathbf{e}^{\mathrm{i}\xi(k_{f,x,u}-1)}\big) d\xi
\end{aligned}$$

$$= \begin{cases} 2\pi(1 - b_{f,x,u}), & \text{if } k_{f,x,u} = 1, -1, \\ 2\pi b_{f,x,u}, & \text{if } k_{f,x,u} = 0, -2, \\ 0, & \text{if } k_{f,x,u} \in \mathbb{Z} \setminus \{1, 0, -1, -2\}. \end{cases}$$

Using the same procedure, we see that $\int_{\mathbb{R}} |\widehat{p^*}(\xi)|^2(\mathbf{e}^{\mathrm{i}\xi} + \mathbf{e}^{-\mathrm{i}\xi})\mathbf{e}^{-\mathrm{i}\xi(f(x)-f(u))}d\xi$ has exactly the same value. Thus

$$-\frac{1}{4\pi}\int_{\mathbb{R}} \left|\widehat{p^*}(\xi)\right|^2 \left(\mathbf{e}^{\mathrm{i}\xi} + \mathbf{e}^{-\mathrm{i}\xi}\right)\left(\mathbf{e}^{\mathrm{i}\xi(f(x)-f(u))} + \mathbf{e}^{-\mathrm{i}\xi(f(x)-f(u))}\right)d\xi$$
$$= \begin{cases} b_{f,x,u} - 1, & \text{if } k_{f,x,u} = 1, -1, \\ -b_{f,x,u}, & \text{if } k_{f,x,u} = 0, -2, \\ 0, & \text{if } k_{f,x,u} \in \mathbb{Z} \setminus \{1, 0, -1, -2\}. \end{cases}$$

Denote

$$\Delta_1 = \big\{(x,u) \in \mathcal{X}_1 \times \mathcal{X}_2 : 1 \le f(x) - f(u) < 2\big\} \cup \big\{(x,u) \in \mathcal{X}_1 \times \mathcal{X}_2 : -1 \le f(x) - f(u) < 0\big\},$$
$$\Delta_2 = \big\{(x,u) \in \mathcal{X}_1 \times \mathcal{X}_2 : 0 \le f(x) - f(u) < 1\big\} \cup \big\{(x,u) \in \mathcal{X}_1 \times \mathcal{X}_2 : -2 \le f(x) - f(u) < -1\big\},$$
$$\Delta_3 = \big\{(x,u) \in \mathcal{X}_1 \times \mathcal{X}_2 : f(x) - f(u) < -2\big\} \cup \big\{(x,u) \in \mathcal{X}_1 \times \mathcal{X}_2 : f(x) - f(u) \ge 2\big\}.$$

Note that $k_{f,x,u}$ is the integer part of $f(x) - f(u)$. We have

$$V_{12}(f) = \int_{\mathcal{X}_1}\int_{\mathcal{X}_2} \big\{(b_{f,x,u} - 1)\mathbf{1}_{\Delta_1}(x,u) - b_{f,x,u}\mathbf{1}_{\Delta_2}(x,u)\big\} d\rho_X(x)d\rho_X(u).$$

Since $0 \le b_{f,x,u} < 1$, we see that $V_{12}(f)$ is minimized if and only if $b_{f,x,u} = 0$, $\Delta_1 = \mathcal{X}_1 \times \mathcal{X}_2$ and $\Delta_2 = \emptyset$. These conditions are equivalent to $f(x) - f(u) = k_{f,x,u} = \pm 1$ for almost all $(x,u) \in \mathcal{X}_1 \times \mathcal{X}_2$. Therefore, $V_{12}(f)$ is minimized if and only if $|f(x) - f(u)| = 1$ for almost every $(x,u) \in \mathcal{X}_1 \times \mathcal{X}_2$. In this case, the minimum value of $V_{12}(f)$ equals

$$V_{12}^* := -\rho_X(\mathcal{X}_1)\rho_X(\mathcal{X}_2) = -\frac{1}{4}.$$

Moreover, for any measurable function f, we have

$$V_{12}(f) - V_{12}^* = \int_{\mathcal{X}_1}\int_{\mathcal{X}_2} b_{f,x,u}\mathbf{1}_{\Delta_1}(x,u) + (1 - b_{f,x,u})\mathbf{1}_{\Delta_2}(x,u) + \mathbf{1}_{\Delta_3}(x,u)d\rho_X(x)d\rho_X(u).$$

On Δ_1, we have $b_{f,x,u} = ||f(x)-f(u)|-1|$ and as a number on $[0,1)$, it satisfies $b_{f,x,u} = ||f(x)-f(u)|-1| \ge (|f(x) - f(u)| - 1)^2$. Similarly on Δ_2 we have $1 - b_{f,x,u} = ||f(x) - f(u)| - 1| \ge (|f(x) - f(u)| - 1)^2$. On Δ_3, since the function f takes values on $[-M, M]$, we have $2 \le |f(x) - f(u)| \le 2M$. Therefore $1 \ge \frac{1}{4M^2}(|f(x) - f(u)| - 1)^2$. Thus,

$$V_{12}(f) - V_{12}^* \ge \frac{1}{4M^2}\int_{\mathcal{X}_1}\int_{\mathcal{X}_2} \left(\left|f(x) - f(u)\right| - 1\right)^2 d\rho_X(x)d\rho_X(u)$$
$$\ge \frac{1}{48\pi^2M^3}\int_{\mathcal{X}_1}\int_{\mathcal{X}_2} \left(\left|f(x) - f(u)\right| - 1\right)^2 d\rho_X(x)d\rho_X(u),$$

where we impose a lower bound in the last step in order to use (5.3) and (5.4) later.

Please cite this article in press as: J. Fan et al., Consistency analysis of an empirical minimum error entropy algorithm, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.12.005

To bound $V_{12}(f) - V_{12}^*$ further, we need the following elementary inequality: for $A, a \in \mathbb{R}$,

$$A^2 = a^2 + (A-a)^2 + 2\frac{a}{\sqrt{2}}\sqrt{2}(A-a) \geq a^2 + (A-a)^2 - \frac{a^2}{2} - 2(A-a)^2 = \frac{a^2}{2} - (A-a)^2.$$

Applying it with $A = |f(x) - f(u)| - 1$ and $a = |m_{f,\mathcal{X}_1} - m_{f,\mathcal{X}_2}| - 1$ and using the fact

$$\begin{aligned}\big(\big|f(x) - f(u)\big| - |m_{f,\mathcal{X}_1} - m_{f,\mathcal{X}_2}|\big)^2 &\leq \big(\big(f(x) - m_{f,\mathcal{X}_1}\big) - \big(f(u) - m_{f,\mathcal{X}_2}\big)\big)^2 \\ &\leq 2\big(f(x) - m_{f,\mathcal{X}_1}\big)^2 + 2\big(f(u) - m_{f,\mathcal{X}_2}\big)^2,\end{aligned}$$

we obtain

$$\big(\big|f(x) - f(u)\big| - 1\big)^2 \geq \frac{1}{2}\big(|m_{f,\mathcal{X}_1} - m_{f,\mathcal{X}_2}| - 1\big)^2 - 2\big(f(x) - m_{f,\mathcal{X}_1}\big)^2 - 2\big(f(u) - m_{f,\mathcal{X}_2}\big)^2.$$

It follows to

$$V_{12}(f) - V_{12}^* \geq \frac{1}{48\pi^2 M^3}\left\{\frac{1}{8}\big(|m_{f,\mathcal{X}_1} - m_{f,\mathcal{X}_2}| - 1\big)^2 - \|f - m_{f,\mathcal{X}_1}\|^2_{L^2_{\rho_X}(\mathcal{X}_1)} - \|f - m_{f,\mathcal{X}_2}\|^2_{L^2_{\rho_X}(\mathcal{X}_2)}\right\}. \quad (5.5)$$

Combining (5.3), (5.4), and (5.5), we have with $c = \frac{1}{400\pi^2 M^3}$,

$$V(f) - V^* \geq c\big\{\big(|m_{f,\mathcal{X}_1} - m_{f,\mathcal{X}_2}| - 1\big)^2 + \|f - m_{f,\mathcal{X}_1}\|^2_{L^2_{\rho_X}(\mathcal{X}_1)} + \|f - m_{f,\mathcal{X}_2}\|^2_{L^2_{\rho_X}(\mathcal{X}_2)}\big\}. \quad (5.6)$$

With above preparations we can now prove our conclusions. Firstly, combining the conditions for minimizing V_{11}, V_{22} and V_{12} we see easily the result in part (i).

By $V^* = V_{11}^* + V_{22}^* + V_{12}^* = -\frac{5}{8}$ we get $\mathscr{R}^* = -\log(\frac{5}{8})$. For f^*, a direct computation gives $p_E = \frac{1}{4}\mathbf{1}_{[-\frac{3}{2},-\frac{1}{2}]} + \frac{1}{2}\mathbf{1}_{[-\frac{1}{2},\frac{1}{2}]} + \frac{1}{4}\mathbf{1}_{[\frac{1}{2},\frac{3}{2}]}$. So $\mathscr{R}(f^*) = -\log(\frac{3}{8})$ and we prove part (ii).

For any measurable function f, we take a function $f^*_{\mathscr{R}} = f_1\mathbf{1}_{\mathcal{X}_1} + f_2\mathbf{1}_{\mathcal{X}_2}$ with $f_1 = m_{f,\mathcal{X}_1}$ and $f_2 = f_1 + f_{12}$, where f_{12} is a constant defined to be 1 if $m_{f,\mathcal{X}_2} \geq m_{f,\mathcal{X}_1}$ and -1 otherwise. Then $f^*_{\mathscr{R}} \in \mathcal{F}^*_{\mathscr{R}}$ is a minimizer of the error entropy function $\mathscr{R}(f)$. Moreover, it is easy to check that

$$\|f - f^*_{\mathscr{R}}\|^2_{L^2_{\rho_X}} = \|f - m_{f,\mathcal{X}_1}\|^2_{L^2_{\rho_X}(\mathcal{X}_1)} + \|f - f_2\|^2_{L^2_{\rho_X}(\mathcal{X}_2)}.$$

Since $\int_{\mathcal{X}_2}(f - m_{f,\mathcal{X}_2})d\rho_X = 0$, we have

$$\|f - f_2\|^2_{L^2_{\rho_X}(\mathcal{X}_2)} = \int_{\mathcal{X}_2}(f - m_{f,\mathcal{X}_2})^2 d\rho_X + \int_{\mathcal{X}_2}(m_{f,\mathcal{X}_2} - f_2)^2 d\rho_X.$$

Observe that $m_{f,\mathcal{X}_2} - f_2 = m_{f,\mathcal{X}_2} - m_{f,\mathcal{X}_1} - f_{12}$ and by the choice of the constant f_{12}, we see that

$$|m_{f,\mathcal{X}_2} - f_2| = \big||m_{f,\mathcal{X}_2} - m_{f,\mathcal{X}_1}| - 1\big|.$$

Hence

$$\|f - f^*_{\mathscr{R}}\|^2_{L^2_{\rho_X}} = \|f - m_{f,\mathcal{X}_1}\|^2_{L^2_{\rho_X}(\mathcal{X}_1)} + \|f - m_{f,\mathcal{X}_2}\|^2_{L^2_{\rho_X}(\mathcal{X}_2)} + \frac{1}{2}\big(|m_{f,\mathcal{X}_1} - m_{f,\mathcal{X}_2}| - 1\big)^2.$$

This in combination with (5.6) leads to the conclusion in part (iii) with the constant $C' = 400\pi^2 M^3 B_U$.

For part (iv), the first convergence is a direct consequence of the error entropy consistency. To see the second one, it is suffices to notice

$$\min_{b\in\mathbb{R}}\|f_{\mathbf{z}}+b-f^*\|_{L^2_{\rho_X}}=\min_{b\in\mathbb{R}}\min_{f^*_{\mathscr{R}}\in\mathcal{F}^*_{\mathscr{R}}}\|f_{\mathbf{z}}-f^*_{\mathscr{R}}+f^*_{\mathscr{R}}+b\|_{L^2_{\rho_X}}\longrightarrow\min_{b\in\mathbb{R}}\min_{f^*_{\mathscr{R}}\in\mathcal{F}^*_{\mathscr{R}}}\|f^*_{\mathscr{R}}+b\|_{L^2_{\rho_X}},$$

which has the minimum value of $\frac{1}{2}$ achieved at $b=-\frac{f_1+f_2}{2}$. □

6. Regression consistency

In this section we prove that the regression consistency is true for both homoskedastic models and heteroskedastic models when the bandwidth parameter h is chosen to tend to infinity in a suitable rate. We need the following result proved in [8].

Proposition 6.1. *There exists a constant C'' depending only on $\mathcal{H},\rho$ and M such that*

$$\left\|f+\mathbf{E}_x\left[f^*(x)-f(x)\right]-f^*\right\|^2_{L^2_{\rho_X}}\le C''\left(h^3\left(\mathcal{E}_h(f)-\mathcal{E}^*_h\right)+\frac{1}{h^2}\right),\quad\forall f\in\mathcal{H},\ h>0,$$

where $\mathcal{E}^*_h=\min_{f\in\mathcal{H}}\mathcal{E}_h(f)$.

Theorem 2.7 is an easy consequence of Propositions 6.1 and 3.4. To see this, it suffices to notice that $\mathcal{E}_h(f_{\mathbf{z}})-\mathcal{E}^*_h\le 2\mathcal{S}_{\mathbf{z}}$.

7. Regression consistency for two special models

In previous sections we see the information error $\mathcal{E}_h(f)$ plays a very important role in analyzing the empirical MEE algorithm. Actually, it is of independent interest as a loss function to the regression problem. As we discussed, as h tends to 0, $\mathcal{E}_h(f)$ tends to $V(f)$ which is the loss function used in the MEE algorithm. As h tends to ∞, it behaves like a least square ranking loss [8]. In this section we use it to study the regression consistency of MEE for the two classes of special models $\mathcal{P}_1$ and $\mathcal{P}_2$.

7.1. Symmetric unimodal noise model

In this subsection we prove the regression consistency for the symmetric unimodal noise case stated in Theorem 2.12. To this end, We need the following two lemmas of which the first is from [11]. Let $f*g$ denote the convolution of two integrable functions f and g.

Lemma 7.1. *The convolution of two symmetric unimodal distribution functions is symmetric unimodal.*

Lemma 7.2. *Let $\epsilon_x=y-f^*(x)$ be the noise random variable at x and denote $g_{x,u}$ as the probability density function of $\epsilon_x-\epsilon_u$ for $x,u\in\mathcal{X}$ and $\widehat{g_{x,u}}$ as the Fourier transform of $g_{x,u}$. If ρ belongs to $\mathcal{P}_1$, we have*

(i) *$g_{x,u}$ is symmetric and unimodal for $x,u\in\mathcal{X}$;*
(ii) *$\widehat{g_{x,u}}(\xi)$ is nonnegative for $\xi\in\mathbb{R}$;*
(iii) *$\widehat{g_{x,u}}(\xi)\ge C_0$ for $\xi\in[-c_0,c_0]$, where c_0,C_0 are two positive constants.*

Proof. Since both $p_{\epsilon|X}(\cdot|x)$ and $p_{\epsilon|X}(\cdot|u)$ are symmetric and unimodal, (i) is an easy consequence of Lemma 7.1. With the symmetry property, $-\epsilon_u$ has the same density function as ϵ_u, so we have $g_{x,u}=p_{\epsilon|X}(\cdot|x)*p_{\epsilon|X}(\cdot|u)$, which implies

$$\widehat{g_{x,u}}(\xi)=\widehat{p_{\epsilon|X=x}}(\xi)\widehat{p_{\epsilon|X=u}}(\xi).$$

Since ρ is in $\mathcal{P}_1$, we easily see that $\widehat{g_{x,u}}(\xi)$ is nonnegative for $\xi \in \mathbb{R}$ and that for some positive constants c_0, C_0, there holds $\widehat{g_{x,u}}(\xi) \geq C_0$ for $\xi \in [-c_0, c_0]$. □

The following result gives some regression consistency analysis for the MEE algorithm where the bandwidth parameter h is fixed. It immediately implies Theorem 2.12 stated in the second section.

Proposition 7.3. *Assume ρ belongs to $\mathcal{P}_1$. Then for any* fixed *h*

(i) *$f^* + b$ is a minimizer of $\mathcal{E}_h(f)$ for any constant b;*
(ii) *there exists a constant $C_h > 0$ such that*

$$\left\| f + \mathbf{E}_x\left[f^*(x) - f(x)\right] - f^*\right\|^2_{L^2_{\rho_X}} \leq C_h\left(\mathcal{E}_h(f) - \mathcal{E}_h\left(f^*\right)\right), \quad \forall f \in \mathcal{H}; \tag{7.1}$$

(iii) *with probability at least $1 - \delta$, there holds*

$$\left\| f_{\mathbf{z}} + \mathbf{E}_x\left[f^*(x) - f_{\mathbf{z}}(x)\right] - f^*\right\|^2_{L^2_{\rho_X}} \leq \frac{2BC_h}{h^2\sqrt{n}} + \frac{\sqrt{2}C_h}{h\sqrt{n}}\sqrt{\log(1/\delta)}, \tag{7.2}$$

where B is given explicitly in Appendix A.

Proof. Recall that $\epsilon_x = y - f^*(x)$, $\epsilon_u = v - f^*(u)$ and $g_{x,u}$ is the probability density function of $\epsilon_x - \epsilon_u$. We have for any measurable function f,

$$\begin{aligned}
\mathcal{E}_h(f) &= -\int_{\mathcal{Z}}\int_{\mathcal{Z}} G_h\left(\left(y - f(x)\right) - \left(v - f(u)\right)\right) d\rho(x,y)d\rho(u,v) \\
&= \frac{1}{\sqrt{2\pi}h}\int_{\mathcal{X}}\int_{\mathcal{X}}\left[-\int_{-\infty}^{\infty} \exp\left(-\frac{(w-t)^2}{2h^2}\right) g_{x,u}(w)dw\right] d\rho_X(x)d\rho_X(u)
\end{aligned}$$

where $t = f(x) - f^*(x) - f(u) + f^*(u)$.

Now we apply the Planchel formula and find

$$\begin{aligned}
&\mathcal{E}_h(f) - \mathcal{E}_h\left(f^*\right) \\
&\quad = \frac{1}{\sqrt{2\pi}h}\int_{\mathcal{X}}\int_{\mathcal{X}}\left[\int_{\mathbb{R}} \exp\left(-\frac{w^2}{2h^2}\right) g_{x,u}(w)dw - \int_{\mathbb{R}} \exp\left(-\frac{w^2}{2h^2}\right) g_{x,u}(w+t)dw\right] d\rho_X(x)d\rho_X(u) \\
&\quad = \frac{1}{2\pi}\int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathbb{R}} \exp\left(-\frac{h^2\xi^2}{2}\right)\widehat{g_{x,u}}(\xi)\left(1 - \mathbf{e}^{\mathrm{i}\xi(f(x)-f^*(x)-f(u)+f^*(u))}\right) d\xi d\rho_X(x)d\rho_X(u) \\
&\quad = \frac{1}{2\pi}\int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathbb{R}} \exp\left(-\frac{h^2\xi^2}{2}\right)\widehat{g_{x,u}}(\xi) 2\sin^2\frac{\xi(f(x) - f^*(x) - f(u) + f^*(u))}{2} d\xi d\rho_X(x)d\rho_X(u).
\end{aligned}$$

By Lemma 7.2, $\widehat{g_{x,u}}(\xi) \geq 0$ for $\xi \in \mathbb{R}$. So $\mathcal{E}_h(f) - \mathcal{E}_h(f^*) \geq 0$ for any measurable function f. This tells us that f^* and $f^* + b$ for any $b \in \mathbb{R}$ are minimizers of $\mathcal{E}_h(f)$.

To prove (7.1) we notice that both f and f^* take values on $[-M, M]$. Hence $|f(x) - f^*(x) - f(u) + f^*(u)| \leq 4M$ for any $x, u \in \mathcal{X}$. So when $|\xi| \leq \frac{\pi}{4M}$, we have

$$\left|\frac{\xi(f(x) - f^*(x) - f(u) + f^*(u))}{2}\right| \leq \frac{\pi}{2},$$

and

$$\left|\sin\frac{\xi(f(x)-f^*(x)-f(u)+f^*(u))}{2}\right| \geq \frac{2}{\pi}\left|\frac{\xi(f(x)-f^*(x)-f(u)+f^*(u))}{2}\right|.$$

Then we have

$$\begin{aligned}
&\int_{\mathbb{R}} \exp\left(-\frac{h^2\xi^2}{2}\right)\widehat{g_{x,u}}(\xi)2\sin^2\frac{\xi(f(x)-f^*(x)-f(u)+f^*(u))}{2}d\xi \\
&\geq \int_{|\xi|\leq\frac{\pi}{4M}} \exp\left(-\frac{h^2\xi^2}{2}\right)\widehat{g_{x,u}}(\xi)2\sin^2\frac{\xi(f(x)-f^*(x)-f(u)+f^*(u))}{2}d\xi \\
&\geq \int_{|\xi|\leq\frac{\pi}{4M}} \exp\left(-\frac{h^2\xi^2}{2}\right)\widehat{g_{x,u}}(\xi)\frac{2}{\pi^2}\xi^2\big(f(x)-f^*(x)-f(u)+f^*(u)\big)^2 d\xi \\
&\geq \int_{|\xi|\leq\min\{\frac{\pi}{4M},c_0\}} \exp\left(-\frac{h^2\xi^2}{2}\right)\widehat{g_{x,u}}(\xi)\frac{2}{\pi^2}\xi^2\big(f(x)-f^*(x)-f(u)+f^*(u)\big)^2 d\xi \\
&\geq \frac{2C_0}{\pi^2}\int_{|\xi|\leq\min\{\frac{\pi}{4M},c_0\}} \exp\left(-\frac{h^2\xi^2}{2}\right)\xi^2\big(f(x)-f^*(x)-f(u)+f^*(u)\big)^2 d\xi.
\end{aligned}$$

Therefore, using (4.1)

$$\mathcal{E}_h(f)-\mathcal{E}_h\big(f^*\big) \geq \left(\frac{2C_0}{\pi^3}\int_{|\xi|\leq\min\{\frac{\pi}{4M},c_0\}} \xi^2\exp\left(-\frac{h^2\xi^2}{2}\right)d\xi\right)\big\|f+\mathbf{E}_x\big[f^*(x)-f(x)\big]-f^*\big\|^2_{L^2_{\rho_X}}.$$

Since $c_h = \int_{|\xi|\leq\min\{\frac{\pi}{4M},c_0\}}\xi^2\exp(-\frac{h^2\xi^2}{2})d\xi$ is positive, (7.1) follows by taking $C_h = \frac{\pi^3}{2c_hC_0}$.

With (7.1) valid, (iii) is an easy consequence of Proposition 3.4. □

7.2. Symmetric bounded noise models

In this subsection we prove the regression consistency for the symmetric bounded noise models stated in Theorem 2.13.

Proposition 7.4. *We assume ρ belongs to $\mathcal{P}_2$. Then there exists a constant $h_{\rho,\mathcal{H}} > 0$ such that for any* fixed *$h > h_{\rho,\mathcal{H}}$ the following holds:*

(i) *$f^* + b$ is the minimizer of $\mathcal{E}_h(f)$ for any constant b;*
(ii) *there exists a constant $C_2 > 0$ depending only on $\rho, \mathcal{H}, \widetilde{M}$, M and h such that*

$$\big\|f+\mathbf{E}_x\big[f^*(x)-f(x)\big]-f^*\big\|^2_{L^2_{\rho_X}} \leq C_2\big(\mathcal{E}_h(f)-\mathcal{E}_h\big(f^*\big)\big), \quad \forall f\in\mathcal{H}; \tag{7.3}$$

(iii) *with probability at least $1-\delta$, there holds*

$$\big\|f_{\mathbf{z}}+\mathbf{E}_x\big[f^*(x)-f_{\mathbf{z}}(x)\big]-f^*\big\|^2_{L^2_{\rho_X}} \leq \frac{2BC_2}{h^2\sqrt{n}}+\frac{\sqrt{2}C_2}{h\sqrt{n}}\sqrt{\log(1/\delta)}. \tag{7.4}$$

Proof. Since ρ belongs to $\mathcal{P}_2$, we know that ϵ_x is supported on $[-\widetilde{M}, \widetilde{M}]$ and $g_{x,u}$ on $[-2\widetilde{M}, 2\widetilde{M}]$. So for any measurable function $f : \mathcal{X} \to \mathbb{R}$,

$$\mathcal{E}_h(f) = \frac{1}{\sqrt{2\pi}h} \int_{\mathcal{X}} \int_{\mathcal{X}} T_{x,u}\big(f(x) - f^*(x) - f(u) + f^*(u)\big) d\rho_X(x) d\rho_X(u),$$

where $T_{x,u}$ is a univariate function given by

$$T_{x,u}(t) = -\int_{-2\widetilde{M}}^{2\widetilde{M}} \exp\Big(-\frac{(w-t)^2}{2h^2}\Big) g_{x,u}(w) dw.$$

Observe that

$$\begin{aligned} T'_{x,u}(t) &= -\int_{-2\widetilde{M}}^{2\widetilde{M}} \exp\Big(-\frac{(w-t)^2}{2h^2}\Big)\Big(\frac{w-t}{h^2}\Big) g_{x,u}(w) dw \\ &= -\frac{1}{h^2}\int_0^{2\widetilde{M}} w \exp\Big(-\frac{w^2}{2h^2}\Big)\big[g_{x,u}(w+t) - g_{x,u}(w-t)\big] dw, \end{aligned}$$

and

$$T''_{x,u}(t) = -\frac{1}{h^2}\int_{-2\widetilde{M}}^{2\widetilde{M}} \exp\Big(-\frac{(w-t)^2}{2h^2}\Big)\Big[\frac{(w-t)^2}{h^2} - 1\Big] g_{x,u}(w) dw.$$

So $T'_{x,u}(0) = 0$. Moreover, if we choose $h_{\rho,\mathcal{H}} := 4M + 2\widetilde{M}$, then for $h > h_{\rho,\mathcal{H}}$ and $|t| \le 4M$,

$$\begin{aligned} T''_{x,u}(t) &\ge \frac{1}{h^2}\Big(1 - \frac{(4M+2\widetilde{M})^2}{h^2}\Big) \exp\Big(-\frac{2(2M+\widetilde{M})^2}{h^2}\Big) \int_{-2\widetilde{M}}^{2\widetilde{M}} g_{x,u}(w) dw \\ &= \frac{1}{h^2}\Big(1 - \frac{(4M+2\widetilde{M})^2}{h^2}\Big) \exp\Big(-\frac{2(2M+\widetilde{M})^2}{h^2}\Big) > 0. \end{aligned}$$

So $T_{x,u}$ is convex on $[-4M, 4M]$ and $t = 0$ is its unique minimizer. By the fact $t = f(x) - f^*(x) - f(u) + f^*(u) \in [-4M, 4M]$ for all $x, u \in \mathcal{X}$, we conclude that, for any constant b, $f^* + b$ is the minimizer of $\mathcal{E}_h(f)$.

By Taylor expansion, we obtain

$$T_{x,u}(t) - T_{x,u}(0) = T'_{x,u}(0)t + \frac{T''_{x,u}(\xi)}{2} t^2 = \frac{T''_{x,u}(\xi)}{2} t^2, \quad t \in [-4M, 4M],$$

where ξ is between 0 and t. So $|\xi| \le |t| \le 4M$. It follows that with the constant $C_2 = h^2 \exp(\frac{2(2M+\widetilde{M})^2}{h^2}) / (1 - \frac{(4M+2\widetilde{M})^2}{h^2})$ independent of x and u we have

$$t^2 \le 2C_2\big[T_{x,u}(t) - T_{x,u}(0)\big].$$

By virtue of the equality (4.1),

$$\left\| f + \mathbf{E}_x\left[f^*(x) - f(x)\right] - f^* \right\|^2_{L^2_{\rho_X}} \leq C_2\left(\mathcal{E}_h(f) - \mathcal{E}_h(f^*)\right).$$

Together with Proposition 3.4, (7.3) leads to (7.4). Theorem 2.13 has been proved by taking $h_{\rho,\mathcal{H}} = 4M + 2\widetilde{M}$. □

Appendix A. Proof of Proposition 3.4

In this appendix we prove Proposition 3.4. Let us first give the definition of the empirical covering number which is used to characterize the capacity of the hypothesis space and prove the sample error bound.

The ℓ_2-norm empirical covering number is defined by means of the normalized ℓ_2-metric d_2 on the Euclidean space $\mathbb{R}^n$ given by

$$d_2(\mathbf{a}, \mathbf{b}) = \left(\frac{1}{n}\sum_{i=1}^{n} |a_i - b_i|^2\right)^{1/2}$$

for $\mathbf{a} = (a_i)_{i=1}^n, \mathbf{b} = (b_i)_{i=1}^n \in \mathbb{R}^n$.

Definition A.1. For a subset S of a pseudo-metric space $(\mathcal{M}, d)$ and $\varepsilon > 0$, the covering number $\mathcal{N}(S, \varepsilon, d)$ is defined to be the minimal number of balls of radius ε whose union covers S. For a set $\mathcal{H}$ of bounded functions on $\mathcal{X}$ and $\varepsilon > 0$, the ℓ_2-norm empirical covering number of $\mathcal{H}$ is given by

$$\mathcal{N}_2(\mathcal{H}, \varepsilon) = \sup_{n\in\mathbb{N}} \sup_{\mathbf{x}\in\mathcal{X}^n} \mathcal{N}(\mathcal{H}|_{\mathbf{x}}, \varepsilon, d_2). \tag{A.1}$$

where for $n \in \mathbb{N}$ and $\mathbf{x} = (x_i)_{i=1}^n \in \mathcal{X}^n$, we denote the covering number of the subset $\mathcal{H}|_{\mathbf{x}} = \{(f(x_i))_{i=1}^n : f \in \mathcal{H}\}$ of the metric space $(\mathbb{R}^n, d_2)$ as $\mathcal{N}(\mathcal{H}|_{\mathbf{x}}, \varepsilon, d_2)$.

Definition A.2. Let ρ be a probability measure on a set $\mathcal{X}$ and suppose that $X_1, ..., X_n$ are independent samples selected according to ρ. Let $\mathcal{H}$ be a class of functions mapping from $\mathcal{X}$ to $\mathbb{R}$. Define the random variable

$$\hat{\mathcal{R}}_n(\mathcal{H}) = \mathbf{E}_\sigma\left[\sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\right| \middle| X_1, ..., X_n\right], \tag{A.2}$$

where $\sigma_1, ..., \sigma_n$ are independent uniform $\{\pm 1\}$-valued random variables. Then the Rademacher average [2] of $\mathcal{H}$ is $\mathcal{R}_n(\mathcal{H}) = \mathbf{E}\hat{\mathcal{R}}_n(\mathcal{H})$.

The following lemma from [1] shows that these two complexity measures we just defined are closely related.

Lemma A.3. *For a bounded function class $\mathcal{H}$ on $\mathcal{X}$ with bound M, and $\mathcal{N}_2(\mathcal{H}, \varepsilon)$ is ℓ_2-norm empirical covering number of $\mathcal{H}$, then there exists a constant C_1 such that for every positive integer n the following holds:*

$$\hat{\mathcal{R}}_n(\mathcal{H}) \leq C_1 \int_0^M \left(\frac{\log \mathcal{N}_2(\mathcal{H}, \varepsilon)}{n}\right)^{1/2} d\varepsilon. \tag{A.3}$$

Moreover, we need the following lemma for Rademacher average.

Lemma A.4. *(1) For any uniformly bounded function f,*

$$\mathcal{R}_n(\mathcal{H}+f) \le \mathcal{R}_n(\mathcal{H}) + \|f\|_\infty/\sqrt{n}.$$

(2) Let $\{\phi_i\}_{i=1}^n$ be functions with Lipschitz constants γ_i, then [13] gives

$$\mathbf{E}_\sigma\left\{\sup_{f\in\mathcal{H}}\sum_{i=1}^n \sigma_i\phi_i\big(f(x_i)\big)\right\} \le \mathbf{E}_\sigma\left\{\sup_{f\in\mathcal{H}}\sum_{i=1}^n \sigma_i\gamma_i f(x_i)\right\}.$$

By applying McDiarmid's inequality we have the following proposition.

Proposition A.5. *For every $\varepsilon_1 > 0$, we have*

$$\mathbb{P}\{\mathcal{S}_\mathbf{z} - \mathbf{E}\mathcal{S}_\mathbf{z} > \varepsilon_1\} \le \exp\big(-2nh^2\varepsilon_1^2\big).$$

Proof. Recall

$$\mathcal{S}_\mathbf{z} = \sup_{f\in\mathcal{H}}\big|\mathcal{E}_{h,\mathbf{z}}(f) - \mathcal{E}_h(f)\big|.$$

Let $i \in \{1,\cdots,n\}$ and $\tilde{\mathbf{z}} = \{z_1,\cdots,z_{i-1},\tilde{z}_i,z_{i+1},\cdots,z_n\}$ be identical to $\mathbf{z}$ except the i-th sample. Then

$$\begin{aligned}
|\mathcal{S}_\mathbf{z} - \mathcal{S}_{\tilde{\mathbf{z}}}| &\le \sup_{(x_i,y_i)_{i=1}^n,(\tilde{x}_i,\tilde{y}_i)}\Big|\sup_{f\in\mathcal{H}}\big|\mathcal{E}_{h,\mathbf{z}}(f) - \mathcal{E}_h(f)\big| - \sup_{f\in\mathcal{H}}\big|\mathcal{E}_{h,\tilde{\mathbf{z}}}(f) - \mathcal{E}_h(f)\big|\Big| \\
&\le \sup_{(x_i,y_i)_{i=1}^n,(\tilde{x}_i,\tilde{y}_i)}\sup_{f\in\mathcal{H}}\big|\mathcal{E}_{h,\mathbf{z}}(f) - \mathcal{E}_{h,\tilde{\mathbf{z}}}(f)\big| \\
&\le \frac{1}{n^2}\sum_{j=1}^n \sup_{(x_i,y_i)_{i=1}^n,(\tilde{x}_i,\tilde{y}_i)}\sup_{f\in\mathcal{H}}\big|G_h(e_i,e_j) - G_h(\tilde{e}_i,e_j)\big| \\
&\le \frac{1}{nh}.
\end{aligned}$$

Then the proposition follows immediately from McDiarmid's inequality. □

Now we need to bound $\mathbf{E}\mathcal{S}_\mathbf{z}$.

Proposition A.6.

$$\mathbf{E}\mathcal{S}_\mathbf{z} \le \frac{2}{\sqrt{\pi}h^2}\left(\frac{M}{\sqrt{n}} + \mathcal{R}_n(\mathcal{H})\right) + \frac{2}{\sqrt{2\pi}hn}.$$

Proof. Let $\eta(x,y,u,v) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{[(y-f(x))-(v-f(u))]^2}{2h^2})$ for simplicity. Then

$$\mathcal{E}_{h,\mathbf{z}}(f) = -\frac{1}{n^2h}\sum_{i=1}^n\sum_{j=1}^n \eta(x_i,y_i,x_j,y_j)$$

and

$$\mathcal{E}_h(f) = -\frac{1}{h}\mathbf{E}_{(x,y)}\mathbf{E}_{(u,v)}\eta(x,y,u,v).$$

Then

$$\begin{aligned}
h\mathcal{S}_{\mathbf{z}} &= h\sup_{f\in\mathcal{H}}\left|\mathcal{E}_{h,\mathbf{z}}(f)-\mathcal{E}_h(f)\right| \\
&\le \sup_{f\in\mathcal{H}}\left|\mathbf{E}_{(x,y)}\mathbf{E}_{(u,v)}\eta(x,y,u,v)-\frac{1}{n}\sum_{j=1}^n \mathbf{E}_{(x,y)}\eta(x,y,x_j,y_j)\right| \\
&\quad + \sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{j=1}^n \mathbf{E}_{(x,y)}\eta(x,y,x_j,y_j)-\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \eta(x_i,y_i,x_j,y_j)\right| \\
&\le \mathbf{E}_{(x,y)}\sup_{f\in\mathcal{H}}\left|\mathbf{E}_{(u,v)}\eta(x,y,u,v)-\frac{1}{n}\sum_{j=1}^n \eta(x,y,x_j,y_j)\right| \\
&\quad + \frac{1}{n}\sum_{j=1}^n \sup_{(u,v)\in\mathbf{z}}\sup_{f\in\mathcal{H}}\left|\mathbf{E}_{(x,y)}\eta(x,y,u,v)-\frac{1}{n-1}\sum_{\substack{i=1\\ i\neq j}}^n \eta(x_i,y_i,u,v)\right| \\
&\quad + \frac{1}{n}\sum_{j=1}^n \sup_{f\in\mathcal{H}}\left[\frac{1}{n}\eta(x_j,y_j,x_j,y_j)+\frac{1}{n(n-1)}\sum_{\substack{i=1\\ i\neq j}}^n \eta(x_i,y_i,x_j,y_j)\right] \\
&:= S_1+S_2+S_3.
\end{aligned}$$

Noting that

$$\left|\exp\left(-\left(y_i-f(x_i)\right)^2\right)-\exp\left(-\left(y_i-g(x_i)\right)^2\right)\right| \le \left|f(x_i)-g(x_i)\right|,$$

we have

$$\begin{aligned}
\mathbf{E}S_1 &= \mathbf{E}_{(x,y)}\mathbb{E}\sup_{f\in\mathcal{H}}\left|\mathbb{E}_{(u,v)}\eta(x,y,u,v)-\frac{1}{n}\sum_{j=1}^n \eta(x,y,x_j,y_j)\right| \\
&\le \frac{2}{\sqrt{2\pi}}\sup_{(x,y)\in\mathbf{z}}\mathbf{E}\mathbf{E}_\sigma \sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{j=1}^n \sigma_j \exp\left(-\frac{[(y-f(x))-(y_j-f(x_j))]^2}{2h^2}\right)\right| \\
&\le \frac{1}{h\sqrt{\pi}}\sup_{x\in\mathcal{X}}\mathbf{E}\mathbf{E}_\sigma \sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{j=1}^n \sigma_j\left(f(x)-f(x_j)\right)\right| \\
&\le \frac{1}{h\sqrt{\pi}}\left[\sup_{x\in X}\mathbf{E}_\sigma \sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{j=1}^n \sigma_j f(x)\right| + \mathbf{E}\mathbf{E}_\sigma \sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{j=1}^n \sigma_j f(x_j)\right|\right] \\
&\le \frac{1}{h\sqrt{\pi}}\left(\frac{M}{\sqrt{n}}+\mathcal{R}_n(\mathcal{H})\right),
\end{aligned}$$

where the second inequality is from Lemma A.4. Similarly,

$$\begin{aligned}
\mathbf{E}S_2 &= \frac{1}{n}\sum_{j=1}^n \sup_{(u,v)\in\mathbf{z}}\mathbf{E}\sup_{f\in\mathcal{H}}\left|\mathbf{E}_{(x,y)}\eta(x,y,u,v)-\frac{1}{n-1}\sum_{\substack{i=1\\ i\neq j}}^n \eta(x_i,y_i,u,v)\right| \\
&\le \frac{2}{n\sqrt{2\pi}}\sum_{j=1}^n \sup_{(u,v)\in\mathbf{z}}\mathbf{E}\mathbf{E}_\sigma \sup_{f\in\mathcal{H}}\left|\frac{1}{n-1}\sum_{\substack{i=1\\ i\neq j}}^n \sigma_i \exp\left(-\frac{[(y_i-f(x_i))-(v-f(u))]^2}{2h^2}\right)\right|
\end{aligned}$$

$$\leq \frac{1}{nh\sqrt{\pi}} \sum_{j=1}^{n} \sup_{u \in \mathcal{X}} \mathbf{E}\mathbf{E}_\sigma \sup_{f \in \mathcal{H}} \left| \frac{1}{n-1} \sum_{\substack{i=1 \\ i \neq j}}^{n} \sigma_i \big(f(x_i) - f(u)\big) \right|$$

$$\leq \frac{1}{nh\sqrt{\pi}} \sum_{j=1}^{n} \left[\sup_{u \in X} \mathbf{E}_\sigma \sup_{f \in \mathcal{H}} \left| \frac{1}{n-1} \sum_{\substack{i=1 \\ i \neq j}}^{n} \sigma_i f(u) \right| + \mathbf{E}\mathbf{E}_\sigma \sup_{f \in \mathcal{H}} \left| \frac{1}{n-1} \sum_{\substack{i=1 \\ i \neq j}}^{n} \sigma_i f(x_i) \right| \right]$$

$$= \frac{1}{h\sqrt{\pi}} \left(\frac{M}{\sqrt{n}} + \mathcal{R}_n(\mathcal{H}) \right).$$

It's easy to obtain $\mathbf{E}S_3 \leq \frac{2}{n\sqrt{2\pi}}$. Combining the estimates for S_1, S_2, S_3 completes the proof. □

Now we can prove Proposition 3.4.

If $\mathcal{H}$ is MEE admissible, (A.3) leads to

$$\mathcal{R}_n(\mathcal{H}) = \mathbf{E}\hat{\mathcal{R}}_n(\mathcal{H}) \leq \frac{C_1}{\sqrt{n}} \int_0^M \mathbf{E}\sqrt{\log \mathcal{N}_2(\mathcal{H}, \varepsilon)}\, d\varepsilon$$

$$\leq \frac{C_1}{\sqrt{n}} \int_0^M \sqrt{\mathbf{E} \log \mathcal{N}_2(\mathcal{H}, \varepsilon)}\, d\varepsilon$$

$$\leq \frac{C_1\sqrt{c}}{\sqrt{n}} \int_0^M \varepsilon^{-s/2} d\varepsilon$$

$$= \left(\frac{2C_1\sqrt{c}}{2-s} M^{1-s/2} \right) \frac{1}{\sqrt{n}}.$$

Let $B = \frac{4C_1\sqrt{c}}{(2-s)\sqrt{\pi}} M^{1-s/2} + \frac{2M+\sqrt{2}}{\sqrt{\pi}}$, combining Proposition A.5 and Proposition A.6 yields the desired result.

References

[1] P.L. Bartlett, O. Bousquet, S. Mendelson, Local Rademacher complexities, Ann. Statist. 33 (2005) 1497–1537.
[2] P.L. Bartlett, S. Mendelson, Rademacher and Gaussian complexities: risk bounds and structural results, J. Mach. Learn. Res. 3 (2002) 463–482.
[3] D. Erdogmus, K. Hild II, J.C. Principe, Blind source separation using Rényi's α-marginal entropies, Neurocomputing 49 (2002) 25–38.
[4] D. Erdogmus, J.C. Principe, Comparison of entropy and mean square error criteria in adaptive system training using higher order statistics, in: Proceedings of the Intl. Conf. on ICA and Signal Separation, Springer-Verlag, Berlin, 2000, pp. 75–90.
[5] D. Erdogmus, 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, 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] E. Gokcay, J.C. Principe, Information theoretic clustering, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2) (2002) 158–171.
[8] T. Hu, J. Fan, Q. Wu, D.-X. Zhou, Learning theory approach to a minimum error entropy criterion, J. Mach. Learn. Res. 14 (2013) 377–397.
[9] R.Q. Jia, C.A. Micchelli, Using the refinement equation for the construction of pre-wavelets II: Power of two, in: P.J. Laurent, A. LeMéhauté, L.L. Schumaker (Eds.), Curves and Surfaces, Academic Press, New York, 1991, pp. 209–246.
[10] T.J. Kozubowski, K. Podgórski, G. Samorodnitsky, Tails of Lévy measure of geometric stable random variables, Extremes 1 (3) (1999) 367–378.
[11] R. Laha, On a class of unimodal distributions, Proc. Amer. Math. Soc. 12 (1961) 181–184.
[12] R. Linsker, Self-organization in a perceptual network, Computer 21 (1988) 105–117.
[13] R. Meir, T. Zhang, Generalization error bounds for Bayesian mixture algorithms, J. Mach. Learn. Res. 4 (2003) 149–192.
[14] J.P. Nolan, Stable Distributions – Models for Heavy Tailed Data, Birkhauser, Boston, 2012, forthcoming, Chapter 1 online at academic2.american.edu/~jpnolan.

Please cite this article in press as: J. Fan et al., Consistency analysis of an empirical minimum error entropy algorithm, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.12.005

[15] J.C. Principe, Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives, Springer-Verlag, New York, 2010.
[16] L.M. Silva, J. Marques de Sá, L.A. Alexandre, Neural network classification using Shannon's entropy, in: Proceedings of the European Symposium on Artificial Neural Networks, Bruges, 2005, pp. 217–222, d-side.
[17] L.M. Silva, J. Marques de Sá, L.A. Alexandre, The MEE principle in data classification: a perceptron-based analysis, Neural Comput. 22 (2010) 2698–2728.
[18] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, 1998.
[19] Q. Wu, Classification and Regularization in Learning Theory, VDM Verlag, 2009.

Please cite this article in press as: J. Fan et al., Consistency analysis of an empirical minimum error entropy algorithm, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.12.005