Consistency Analysis of Empirical MEE Algorithm

Anonymous Author(s)

Affiliation

Address

email

Abstract

In this paper we study the consistency of the empirical minimum error entropy (MEE) algorithm for regression learning. Two types of consistency are studied. The error entropy consistency, which requires the error entropy of the learned function approximates 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 approximates 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 counter-example is used to show 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 certain rates. This result, however, contradicts the motivation of MEE principle because the minimum error entropy is believed to be not approximated well with this choice of bandwidth parameter.

1 Introduction

Information theoretical learning (ITL) is an important research area in signal processing and machine learning. It uses the concepts of entropies from information theory to substitute the conventional statistical descriptors of variances and covariances. The idea dates back at least to [7] while its blossom was inspired by a series works of Principe and coworkers. In [2] the minimum error entropy (MEE) principle was introduced to regression problems. Later on its theoretical properties were studied and its applications in feature extraction, clustering, and blind source separation were developed [3, 5, 1, 4]. More recently the MEE principle was applied to classification problems [9, 10]. For a comprehensive survey and more recent advances on ITL and MEE principle, see [8] and references therein.

The main purpose of this paper is rigorous consistency analysis of an empirical MEE algorithm. Although the empirical MEE has been developed and successfully applied in various areas for more than a decade, it is surprising that consistency analysis is still its lack. There are some theoretical studies in the literature which provide some useful guidance on the understanding of the empirical MEE and its parameter strategy. But they are not from asymptotic perspective and cannot explain the effectiveness of the empirical MEE algorithm as the sample size gets large. In this paper we will analyze the algorithm from a statistical learning theory perspective. The asymptotic analysis will help to establish the consistency of the empirical MEE in several different situations. It turns out the consistency of the empirical MEE is a very complicated issue which explains its difficulty.

We will focus on a regression setting in learning theory. In statistics a regression problem is usually modelled as the estimation of a target function f^* from the input data space $\mathcal{X}$ to the output data 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 statistical learning context [11], the regression setting is usually described as the learning of 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 $\mathbf{z} = \{(x_i, y_i)\}$ as the approximation of f^*. The goodness of this approximation can be measured by certain distance between $f_{\mathbf{z}}$ and f^*, for instance, $\|f_{\mathbf{z}} - f^*\|_{L^2_{\rho_{\mathcal{X}}}}$, the L^2 distance with respect to the marginal $\rho_{\mathcal{X}}$.

MEE algorithms for regression are motivated by minimizing some entropies of the error function $E = E(f) = Y - f(X)$. In this paper we focus on the Rényi's entropy of order 2 defined as

$$\mathscr{R}(f) = -\log\left(\mathbf{E}[p_{_E}]\right) = -\log\left(\int_{\mathbb{R}} (p_{_E}(e))^2 \, de\right).$$

Here and in the sequel, $p_{_E}$ is the probability density function of E. Denote $e_i = y_i - f(x_i)$. Then $p_{_E}$ can be estimated from the samples by a kernel density estimator by using a Gaussian kernel $G_h(t) = \frac{1}{\sqrt{2\pi}h}\mathbf{e}^{-\frac{t^2}{2h^2}}$ with bandwidth parameter h:

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

The MEE algorithm learns $f_{\mathbf{z}}$ from a set of hypothesis space $\mathcal{H}$ by minimizing the empirical version of the Renyi's entropy

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

That is, $f_{\mathbf{z}} = arg \min\limits_{f \in \mathcal{H}} \mathscr{R}_{\mathbf{z}}(f)$. It is obvious that the 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 empirical approximation of the regression function f^*.

To study the asymptotical properties of the MEE algorithm we define two types of consistency.

Definition 1.1. The MEE algorithm is consistent with respect to the Rényi's error entropy if $\mathscr{R}(f_{\mathbf{z}})$ converges to $\mathscr{R}^* = \inf\limits_{f:\mathcal{X}\to\mathbb{R}} \mathscr{R}(f)$ in probability as $n \to \infty$, i.e., for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}\Big(\mathscr{R}(f_{\mathbf{z}}) - \mathscr{R}^* > \varepsilon\Big) = 0$$

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 $L^2_{\rho_{\mathcal{X}}}$ 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.,

$$\lim_{n\to\infty} \mathbf{P}\Big(\|f_{\mathbf{z}} + b_{\mathbf{z}} - f^*\|^2_{L^2_{\rho_{\mathcal{X}}}} > \varepsilon\Big) = 0.$$

Note that the error entropy consistency ensures the learnability of minimum error entropy, as is expected from the motivation of empirical MEE algorithms, while the regression function consistency enables good approximations of the regression target function f^*. These two types of consistency, however, is not necessarily coincident. Instead, they may contradict each other.

Our main contributions are to show the incoincidence of the two types of consistency and illustrate complication of the regression function consistency. A couple of main results will be proved: Firstly we will prove that the error entropy consistency is always true by choosing the bandwidth parameter h to tend to 0 slowly enough. This is somewhat an expected result. However 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, we present a counter-example for which the error entropy consistency and regression function consistency do not coincide. Lastly, we prove a quite surprising result which states that the empirical MEE is always consistent with respect to the regression target function if the bandwidth parameter tends to infinity at certain rate. This was observed in some earlier empirical work but clearly contradicts the motivation of MEE algorithms because Parzen windowing for the minimum error entropy does not lead to convergence without $h \to 0$. These results show that the consistency of the empirical MEE is a very complicated issue and needs further investigations.

2 Main results

We state our main results in this section while leaving their proofs in later sections. 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 whole paper.

For the regression model, we assume some regularity but still natural conditions to simplify our analysis.

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.

(ii) the regression function f^* is bounded by a constant $M > 0$;

(iii) the minimum of $\mathscr{R}(f)$ is achievable by a measurable function $f^*_{\mathscr{R}}$.

Note 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 exists between f^* and $f^*_{\mathscr{R}}$. To figure out this relationship is one of our tasks below.

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

Definition 2.2. We say a function class $\mathcal{H}$ 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 (see e.g. [12] 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 of $\mathscr{R}(f)$ and the regression function f^* are in $\mathcal{H}$.

The first condition is usual and natural since we do not expect to learning unbounded functions. The second condition ensures $\mathcal{H}$ is a learnable class so that overfitting will not happen. This is a very common assumption in learning theory. It is also easily fulfilled by many usually 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 $\mathcal{H}$.

Our first main result is the error entropy consistency.

Theorem 2.3. *Let the regression model and the hypothesis class $\mathcal{H}$ be MEE admissible. 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 $p'_{\epsilon|X}$ of the density function exists and is uniformly bounded by a constant M independent of X, a convergence rate of order $O(n^{-\frac{1}{6}})$ can be obtained by choosing $h(n) \sim n^{-\frac{1}{6}}$.

In the literature of MEE study, the optimal choice of h is suggested to be $h(n) \sim n^{-\frac{1}{5}}$ (see e.g. [8]). We see this choice satisfies our condition for the error entropy consistency. But the optimal rate analysis is 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. We show this by results for 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.

Theorem 2.5. *If the regression model is homoskedastic, we have*

(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 an absolute constant* C *such that, for any measurable function* f *bounded by* M,
$$\|f + \mathbf{E}(f^* - f) - f^*\|_{L^2_{\rho_\mathcal{X}}} \leq C\left(\mathscr{R}(f) - \mathscr{R}^*\right);$$

(iii) *If* (2.1) *is true, then* $f_\mathbf{z} + \mathbf{E}_x(f^* - f_\mathbf{z}(x))$ *converges to* f^* *in probability.*

(iv) *If, in addition,* $p'_{\epsilon|X}$ *exists and is uniformly bounded by a constant* M *independent of* X, *the convergence rate of order* $O(n^{-\frac{1}{6}})$ *can be obtained by choosing* $h \sim n^{-\frac{1}{6}}$.

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

Things are much more complicated for heteroskedastic models. The first result we want to show is the incoincidence of the minimizer $f^*_{\mathscr{R}}$ and the regression function f^*.

Theorem 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. Surprisingly we can show that the regression consistency is also always true provided that the bandwidth parameter h is chosen appropriately.

Theorem 2.7. *Let the regression model and the hypothesis class be MEE admissible. Choosing the bandwidth parameter* $h = h(n)$ *such that*
$$\lim_{n\to\infty} h(n) = +\infty, \qquad \lim_{n\to\infty} \frac{h}{\sqrt{n}} = 0, \tag{2.2}$$
we have $f_\mathbf{z} + \mathbf{E}_x(f^*(x) - f_\mathbf{z}(x))$ *converges to* f^* *in probability. A convergence rate of order* $O(n^{-\frac{1}{4}})$ *can be obtained by taking* $h \sim n^{\frac{1}{8}}$.

This 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 empirical MEE algorithm does not justify its motivation.

Another interesting observation is that the regression consistency in Theorem 2.5 and Theorem 2.7 suggest 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.

3 Error entropy consistency

In this section we will prove that $\mathscr{R}(f_{\mathbf{z}})$ converges to $\mathscr{R}^*$ in probability. We need several useful lemmas.

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

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

As a result, we have $|p_E(e)| \le M$.

Proof. The equation (3.1) follows from the fact that

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

and $|p_E(e)| \le M$ 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 $\mathcal{H}$,

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

Lemma 3.2. *We have $0 < B_L$ and $B_U \le M$.*

Define

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

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

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

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

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

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

and its sample limit form

$$\begin{aligned}\mathcal{E}_h(f) &= -\int_{\mathbb{R}}\int_{\mathbb{R}} G_h(e-\tau) p_E(e) p_E(\tau) de d\tau \\ &= -\int_{\mathcal{Z}}\int_{\mathcal{Z}} G_h\Big((y - f(x)) - (v - f(u))\Big) d\rho(x,y) d\rho(u,v).\end{aligned}$$

Again we see the equivalence of minimizing $\mathscr{R}_{\mathbf{z}}(f)$ to 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 $V(f_{\mathbf{z}}) - V^*$ as follows:

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

Next we will estimate $\mathcal{S}_{\mathbf{z}}$ and $\mathcal{A}_h$ respectively.

The first term $\mathcal{S}_{\mathbf{z}}$ depends on the sample. Its estimation requires the use of uniform central limit theorems.

Proposition 3.4. *With $B = \frac{4C_1\sqrt{c}}{(2-s)\sqrt{\pi}}M^{1-s/2} + \frac{2M}{\sqrt{\pi}} + \frac{2}{\sqrt{2\pi}}$, we have for any $\varepsilon_1 > 0$,*

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

This proposition implies that $\mathcal{S}_{\mathbf{z}}$ is bounded by $O\left(\frac{1}{h^2\sqrt{n}} + \frac{1}{h\sqrt{n}}\right)$ with large probability. The proof of this proposition is complicated but rather standard in the context of learning theory. So we omit the proof.

Proposition 3.5. $\lim\limits_{h\to 0}\mathcal{A}_h = 0$. *If in addition $|p'_{\epsilon|X}| \le M$, the convergence rate is of order $O(h)$.*

Proof. We have

$$\begin{aligned}\mathcal{A}_h &= \sup_{f\in\mathcal{H}}\left|\int_{\mathbb{R}}(p_E(e))^2de - \int_{\mathbb{R}}\int_{\mathbb{R}}G_h(e-\tau)p_E(e)p_E(\tau)ded\tau\right| \\ &= \sup_{f\in\mathcal{H}}\left|\int_{\mathbb{R}}(p_E(e))^2de - \int_{\mathbb{R}}\int_{\mathbb{R}}G_1(\tilde\tau)p_E(e-\tilde\tau h)d\tilde\tau p_E(e)de\right| \\ &\le \sup_{f\in\mathcal{H}}\int_{\mathbb{R}}p_E(e)\int_{\mathbb{R}}G_1(\tilde\tau)\Big|p_E(e)-p_E(e-\tilde\tau h)\Big|d\tilde\tau de\end{aligned}$$

By the dominated convergence theorem, we have $\lim\limits_{h\to 0}\mathcal{A}_h = 0$ when $h \to 0$.

If $|p'_{\epsilon|X}| \le M$, we have $\left|p_{\epsilon|X}(e+f(x)-f^*(x)|x) - p_{\epsilon|X}(e-\tilde\tau h+f(x)-f^*(x)|x)\right| \le M\tilde\tau h$ which implies $|p_E(e)-p_E(e-\tilde\tau h)| \le M\tilde\tau h$. This proves the convergence rate of $O(h)$. □

We see that Theorem 2.3 is an easy corollary of Propositions 3.4 and 3.5.

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 Fourier transform of a function $f \in L^2(\mathbb{R})$ is denoted by $\hat{f}$. For complex numbers, we use i to denote the imaginary unit and $\overline{a}$ the conjugate of a complex number a.

Proof of Theorem 2.5. By Lemma 3.1 we have

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

We apply the Planchel formula and find

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

It follows that

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

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 shifts, we prove part (i) of Theorem 2.5.

To prove part (ii), we study the excess quantity $V(f) - V(f^*)$. We have

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

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 $|f(x) - f^*(x) - f(u) + f^*(u)| \leq 4M$ for any $x, u \in \mathcal{X}$. Using the preliminary inequality $|\sin(t)| < \frac{2}{\pi}|t|$ for $|t| < \frac{\pi}{2}$, we obtain

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

Therefore,

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

It is easy to check [6] that

$$\int_{\mathcal{X}} \int_{\mathcal{X}} \left(f(x) - f^*(x) - f(u) + f^*(u)\right)^2 d\rho_{\mathcal{X}}(x) d\rho_{\mathcal{X}}(u) = 2\|f - f^* + \mathbf{E}(f^* - f)\|^2_{L^2_{\rho_{\mathcal{X}}}}.$$

So we have

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

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 $c = \int_{|\xi| \leq \frac{\pi}{4M}} \xi^2 |\widehat{p_\epsilon}(\xi)|^2 d\xi$ 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 two types of consistency.

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

Example 5.1. *Let $\mathcal{X} = \mathcal{X}_1 \bigcup \mathcal{X}_2 = [0, \frac{1}{2}] \bigcup [1, \frac{3}{2}]$ and $\rho_{\mathcal{X}}$ be uniform on $\mathcal{X}$ (so that $d\rho_{\mathcal{X}} = dx$). 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 $[-\frac{3}{2}, -\frac{1}{2}] \bigcup [\frac{1}{2}, \frac{3}{2}]$ 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 constant $f_1,\ f_2$ 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. For any measurable function f bounded by M, there is an absolute constant C independent of f such that*

$$\min_{f^*_{\mathscr{R}} \in \mathcal{F}^*_{\mathscr{R}}} \|f - f^*_{\mathscr{R}}\|_{L^2_{\rho_{\mathcal{X}}}} \leq C\Big(\mathscr{R}(f) - \mathscr{R}^*\Big).$$

(iv) *If the error entropy consistency is true, there holds*

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

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

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 [6].

Proposition 6.1. *There exists an absolute constant C such that*

$$\|f - f^* - \mathbf{E}(f-f^*)\|_{L^2_{\rho_{\mathcal{X}}}} \le C\left(h^3\left(\mathcal{E}_h(f) - \mathcal{E}_h^*\right) + \frac{1}{h^2}\right)$$

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 $\mathcal{E}_h(f) - \mathcal{E}_h^* \le 2\mathcal{S}_{\mathbf{z}}$.

References

[1] D. Erdogmus, K. Hild II, and J. C. Príncipe. Blind source separation using rényi's α-marginal entropies. *Neurocomputing*, 49:25–38, 2002.

[2] D. Erdogmus and J. C. Príncipe. 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*, pages 75–80, 2000.

[3] D. Erdogmus and J. C. Príncipe. An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems. *IEEE Trans. Signal Process*, 50:1780–1786, 2002.

[4] D. Erdogmus and J. C. Príncipe. Convergence properties and data efficiency of the minimum error entropy criterion in adaline training. *IEEE Trans. Signal Process*, 51:1966–1978, 2003.

[5] E. Gokcay and J. C. Príncipe. Information theoretic clustering. *IEEE Trans. on Pattern Analysis and Machine Learning*, 24:2:158–171, 2002.

[6] T. Hu, J. Fan, Q. Wu, and D.-X. Zhou. Learning theory approach to minimum error entropy criterion. *Journal of Machine Learning Research*, preprint, 2012.

[7] R. Linsker. Self-organization in a perceptual network. *IEEE Computer*, 21:105–117, 1988.

[8] J. C. Príncipe. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. New York: Springer-Verlag, 2010.

[9] L. M. Silva, J. Marques de Sá, and L. A. Alexandre. Neural network classification using shannon's entropy. In *Proceedings of the European Symposium on Artificial Neural Networks*, pages 217–222. Bruges: d-side, 2005.

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

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

[12] Q. Wu. *Classification and Regularization in Learning Theory*. VDM Verlag, 2009.