

Article

On the Smoothed Minimum Error Entropy Criterion

Badong Chen ^{1,2,*} and Jose C. Principe ¹

¹ Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA; E-Mail: principe@cnel.ufl.edu

² Department of Precision Instruments and Mechanology, Tsinghua University, Beijing, 100084, China

* Author to whom correspondence should be addressed; E-Mail: chenbd04@mails.tsinghua.edu.cn.

Received: 9 July 2012; in revised form: 1 November 2012 / Accepted: 1 November 2012 /

Published: 12 November 2012

Abstract: Recent studies suggest that the minimum error entropy (MEE) criterion can outperform the traditional mean square error criterion in supervised machine learning, especially in nonlinear and non-Gaussian situations. In practice, however, one has to estimate the error entropy from the samples since in general the analytical evaluation of error entropy is not possible. By the Parzen windowing approach, the estimated error entropy converges asymptotically to the entropy of the error plus an independent random variable whose probability density function (PDF) corresponds to the kernel function in the Parzen method. This quantity of entropy is called the smoothed error entropy, and the corresponding optimality criterion is named the smoothed MEE (SMEE) criterion. In this paper, we study theoretically the SMEE criterion in supervised machine learning where the learning machine is assumed to be nonparametric and universal. Some basic properties are presented. In particular, we show that when the smoothing factor is very small, the smoothed error entropy equals approximately the true error entropy plus a scaled version of the Fisher information of error. We also investigate how the smoothing factor affects the optimal solution. In some special situations, the optimal solution under the SMEE criterion does not change with increasing smoothing factor. In general cases, when the smoothing factor tends to infinity, minimizing the smoothed error entropy will be approximately equivalent to minimizing error variance, regardless of the conditional PDF and the kernel.

Keywords: entropy; supervised machine learning; minimum error entropy criterion

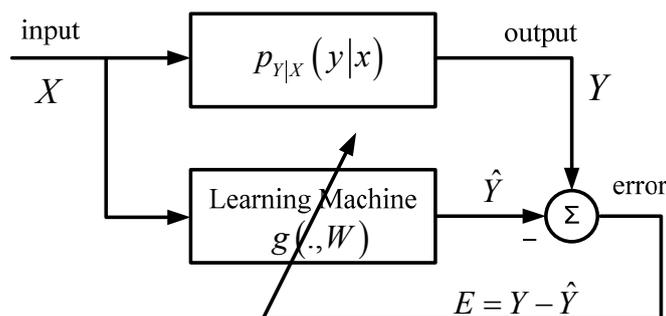
MSC Codes: 62B10

1. Introduction

The principles and methods in Shannon's information theory have been widely applied in statistical estimation, filtering, and learning problems [1–18]. The learning process is essentially a procedure of information processing with the goal of decreasing data redundancy in the presence of uncertainty and encoding the data into a model, and hence it is intrinsically related to information theory. An information theoretic description of learning processes was given in [10], where learning is defined as a process in which the system's subjective entropy or, equivalently, its missing information decreases in time. The mathematical concept of information was also brought to biologically plausible information processing [11]. In addition, a unifying framework for information theoretic learning (ITL) has been presented in [18].

From a statistical viewpoint, learning can be thought of as approximating the *a posteriori* distribution of the targets given a set of examples (training data). Figure 1 shows a general scheme of supervised machine learning, where the desired system output Y (the teacher) is assumed to be related to the input signal X through a conditional probability density function (PDF) $p_{Y|X}(y|x)$, and the learning machine (model) is represented by a parametric mapping $g(\cdot, W)$, where $W \in \mathbb{R}^d$ denotes a parameter vector that needs to be estimated.

Figure 1. A general scheme of supervised machine learning.



The learning goal is then to adapt the parameter W such that the discrepancy between the model output $\hat{Y} = g(X, W)$ and the desired output Y is minimized. How to measure the discrepancy (or model mismatch) is a key aspect in learning. One can use a statistical descriptor of the error ($E = Y - \hat{Y}$) distribution as the measure of discrepancy. The most popular descriptors are the second order moments (variance, correlation, etc.), which combined with the Gaussian assumption, in general leads to mathematically convenient and analytically tractable optimal solutions. A typical example is the mean square error (MSE) criterion in least-squares regression. However, the second order statistics as optimality criteria may perform poorly especially in nonlinear and non-Gaussian (e.g., heavy-tail or

finite range distributions) situations. Recently, the error entropy, as an information theoretic alternative to MSE, has been successfully applied in supervised adaptive system training [12–17]. The minimum error entropy (MEE) criterion usually outperforms MSE criterion in many realistic scenarios, since it captures higher-order statistics and information content of signals rather than simply their energy. Under the MEE criterion, the optimal weight in Figure 1 will be:

$$\begin{aligned} W^* &= \arg \min_{W \in \mathbb{R}^d} H(E) \\ &= \arg \min_{W \in \mathbb{R}^d} - \int p_E(x) \log p_E(x) dx \\ &= \arg \min_{W \in \mathbb{R}^d} \mathbf{E}[-\log p_E(E)] \end{aligned} \quad (1)$$

where $H(E)$ denotes the Shannon entropy of the error E , $p_E(\cdot)$ denotes the error PDF, and \mathbf{E} denotes the expectation operator. Throughout this paper, “log” denotes the natural logarithm. The formulation (1) can be generalized to other entropy definitions, such as the α -order Renyi entropy [18]. Since entropy is shift invariant, the error PDF $p_E(\cdot)$ is in general restricted to zero mean in practice.

The learning machine in Figure 1 can also be a nonparametric and universal mapping $g(\cdot)$. Familiar examples include the support vector machine (SVM) [19,20] and kernel adaptive filtering [21]. In this case, the hypothesis space for learning is in general a high (possibly infinite) dimensional reproducing kernel Hilbert space (RKHS) \mathcal{H} , and the optimal mapping under MEE criterion is:

$$g^* = \arg \min_{g \in \mathcal{H}} - \int p_E(x) \log p_E(x) dx \quad (2)$$

To implement the MEE learning, we should evaluate the error entropy. In practice, however, the error distribution is usually unknown, and the analytical evaluation of error entropy is not possible. Thus we have to estimate the error entropy from the samples. One simple way is to estimate the error PDF based on available samples, and plug the estimated PDF directly into the entropy definition to obtain the entropy estimator. In the literature there are many techniques for estimating the PDF of a random variable based on its sample data. In ITL, the most widely used approach is the Parzen windowing (or kernel density estimation) [22]. By Parzen windowing, the estimated error PDF is:

$$\hat{p}_E(x) = \frac{1}{N} \sum_{i=1}^N \kappa_\lambda(x - e_i) \quad (3)$$

where $\kappa_\lambda(x) = (1/\lambda)\kappa(x/\lambda)$ is the kernel function with smoothing factor (or kernel size) $\lambda > 0$, κ is a continuous density, and $\{e_i\}_{i=1}^N$ are error samples which are assumed to be independent and identically distributed (i.i.d.). As sample number $N \rightarrow \infty$, the estimated PDF will uniformly converge (with probability 1) to the true PDF convolved with the kernel function, that is

$$\hat{p}_E(x) \xrightarrow{N \rightarrow \infty} (p_E * \kappa_\lambda)(x) \quad (4)$$

where $*$ denotes the convolution operator. Then, by the Parzen windowing approach (with a fixed kernel function κ_λ), the estimated error entropy will converge almost surely (a.s.) to the entropy of the convolved density (see [22,23]). Thus, the actual entropy to be minimized is:

$$H(E + \lambda Z) = -\int (p_E * \kappa_\lambda)(x) \log(p_E * \kappa_\lambda)(x) dx \quad (5)$$

where Z denotes a random variable that is independent of X , Y , E , and has PDF $p_Z(x) = \kappa(x)$. Note that the PDF of the sum of two independent random variables equals the convolution of their individual PDFs. Here, we call the entropy $H(E + \lambda Z)$ the smoothed error entropy, and Z the smoothing variable. Under the smoothed MEE (SMEE) criterion, the optimal mapping in (2) becomes:

$$\begin{aligned} g^* &= \arg \min_{g \in \mathcal{H}} H(E + \lambda Z) \\ &= \arg \min_{g \in \mathcal{H}} -\int (p_E * \kappa_\lambda)(x) \log(p_E * \kappa_\lambda)(x) dx \end{aligned} \quad (6)$$

Although SMEE is an actual learning criterion (as sample number $N \rightarrow \infty$) in ITL, up to now its theoretical properties have been little studied. In this work, we study theoretically the SMEE criterion. The rest of the paper is organized as follows: in Section 2, we present some basic properties of the SMEE criterion. In Section 3, we investigate how the smoothing factor λ affects the optimal solution. Finally in Section 4, we give the conclusion.

2. Some Basic Properties of SMEE Criterion

In this section, some basic properties of SMEE criterion are presented. We assume from now on, without explicit mention, that the learning machine is a nonparametric and universal mapping $g(\cdot)$. In addition, the input vector X belongs to an m -dimensional Euclidean space, $X \in \mathbb{R}^m$, and for simplicity, the output Y is assumed to be a scalar signal, $Y \in \mathbb{R}$.

Property 1: Minimizing the smoothed error entropy will minimize an upper bound of the true error entropy $H(E)$.

Proof: According to the entropy power inequality (EPI) [1], we have:

$$\exp(2H(E + \lambda Z)) \geq \exp(2H(E)) + \exp(2H(\lambda Z)) \quad (7)$$

It follows that:

$$H(E) \leq \frac{1}{2} \log(\exp(2H(E + \lambda Z)) - \exp(2H(\lambda Z))) \quad (8)$$

Thus, minimizing the smoothed error entropy $H(E + \lambda Z)$ minimizes an upper bound of $H(E)$.

Remark 1: Although this property does not give a precise result concerning SMEE vs. MEE, it suggests that minimizing the smoothed error entropy will constrain the true error entropy to small values.

Property 2: The smoothed error entropy is upper bounded by $\frac{1}{2} \log(2\pi e(\sigma_E^2 + \lambda^2 \sigma_Z^2))$, where σ_E^2 and σ_Z^2 denote the variances of E and Z , respectively, and this upper bound is achieved if and only if both E and Z are Gaussian distributed.

Proof: The first part of this property is a direct consequence of the maximum entropy property of the Gaussian distribution. The second part comes from Cramer's decomposition theorem [24], which

states that if X_1 and X_2 are independent and their sum $X_1 + X_2$ is Gaussian, then both X_1 and X_2 must also be Gaussian.

Remark 2: According to Property 2, if both E and Z are Gaussian distributed, then minimizing the smoothed error entropy will minimize the error variance.

Property 3: The smoothed error entropy has the following Taylor approximation around $\lambda = 0$:

$$H(E + \lambda Z) = H(E) + \frac{1}{2} \lambda^2 \sigma_Z^2 J(E) + o(\lambda^2) \tag{9}$$

where $o(\lambda^2)$ denotes the higher-order infinitesimal term of Taylor expansion, and $J(E)$ is the Fisher information of E , defined as:

$$J(E) = \mathbf{E} \left[\left(\frac{\partial}{\partial E} \log p_E(E) \right)^2 \right] \tag{10}$$

Proof: This property can be easily proved by using De Bruijn’s identity [25]. For any two independent random variables X_1 and X_2 , $X_1, X_2 \in \mathbb{R}$, De Bruijn’s identity can be expressed as

$$\frac{\partial}{\partial t} H(X_1 + \sqrt{t} X_2) \Big|_{t=0} = \frac{1}{2} \sigma_{X_2}^2 J(X_1) \tag{11}$$

where $\sigma_{X_2}^2$ denotes the variance of X_2 (Classical deBruijn identity assumes that X_2 is Gaussian. Here, we use a generalized deBruijn identity for arbitrary (not necessarily Gaussian) X_2 [25]). So, we have:

$$\frac{\partial}{\partial \lambda^2} H(E + \lambda Z) \Big|_{\lambda^2=0} = \frac{1}{2} \sigma_Z^2 J(E) \tag{12}$$

and hence, we obtain the first-order Taylor expansion:

$$\begin{aligned} H(E + \lambda Z) &= H(E) + \left(\frac{\partial}{\partial \lambda^2} H(E + \lambda Z) \Big|_{\lambda^2=0} \right) \lambda^2 + o(\lambda^2) \\ &= H(E) + \frac{1}{2} \lambda^2 \sigma_Z^2 J(E) + o(\lambda^2) \end{aligned} \tag{13}$$

Remark 3: By Property 3, with small λ , the smoothed error entropy equals approximately the true error entropy plus a scaled version of the Fisher information of error. This result is very interesting since minimizing the smoothed error entropy will minimize a weighted sum of the true error entropy and the Fisher information of error. Intuitively, minimizing the error entropy tends to result in a spikier error distribution, while minimizing the Fisher information makes the error distribution smoother (smaller Fisher information implies a smaller variance of the PDF gradient). Therefore, the SMEE criterion provides a nice balance between the spikiness and smoothness of the error distribution. In [26], the Fisher information of error has been used as a criterion in supervised adaptive filtering.

Property 4: Minimizing the smoothed error entropy $H(E + \lambda Z)$ is equivalent to minimizing the mutual information between $E + \lambda Z$ and the input X , i.e., $\arg \min_{g \in \mathcal{H}} H(E + \lambda Z) = \arg \min_{g \in \mathcal{H}} I(E + \lambda Z; X)$.

Proof: Since mutual information $I(X; Y) = H(X) - H(X|Y)$, where $H(X|Y)$ denotes the conditional entropy of X given Y , we derive:

$$\begin{aligned}
 \arg \min_{g \in \mathcal{H}} I(E + \lambda Z; X) &= \arg \min_{g \in \mathcal{H}} \{H(E + \lambda Z) - H(E + \lambda Z|X)\} \\
 &= \arg \min_{g \in \mathcal{H}} \{H(E + \lambda Z) - H(Y - g(X) + \lambda Z|X)\} \\
 &= \arg \min_{g \in \mathcal{H}} \{H(E + \lambda Z) - H(Y + \lambda Z|X)\} \\
 &\stackrel{(a)}{=} \arg \min_{g \in \mathcal{H}} H(E + \lambda Z)
 \end{aligned}
 \tag{14}$$

where (a) comes from the fact that the conditional entropy $H(Y + \lambda Z|X)$ does not depend on the mapping $g(\cdot)$.

Property 5: The smoothed error entropy is lower bounded by the conditional entropy $H(Y + \lambda Z|X)$, and this lower bound is achieved if and only if the mutual information $I(E + \lambda Z; X) = 0$.

Proof: As $I(E + \lambda Z; X) = H(E + \lambda Z) - H(Y + \lambda Z|X)$, we have:

$$\begin{aligned}
 H(E + \lambda Z) &= H(Y + \lambda Z|X) + I(E + \lambda Z; X) \\
 &\stackrel{(b)}{\geq} H(Y + \lambda Z|X)
 \end{aligned}
 \tag{15}$$

where (b) is because of the non-negativeness of the mutual information $I(E + \lambda Z; X)$.

Remark 4: The lower bound in Property 5 depends only on the conditional PDF of Y given X and the kernel function κ_λ , and is not related to the learning machine. Combining Property 5 and Property 2, the smoothed error entropy satisfies the following inequalities:

$$H(Y + \lambda Z|X) \leq H(E + \lambda Z) \leq \frac{1}{2} \log(2\pi e(\sigma_E^2 + \lambda^2 \sigma_Z^2))
 \tag{16}$$

Property 6: Let $\rho(y|x, \kappa_\lambda) \triangleq p_{Y|X}(y|x) * \kappa_\lambda(y)$ be the smoothed conditional PDF of Y given $X = x$, where the convolution is with respect to y . If for every $x \in \mathbb{R}^m$, $\rho(y|x, \kappa_\lambda)$ is symmetric (not necessarily about zero) and unimodal in $y \in \mathbb{R}$, then the optimal mapping in (6) equals (almost everywhere):

$$g^*(x) = \int_{\mathbb{R}} y \rho(y|x, \kappa_\lambda) dy + c
 \tag{17}$$

where $c \in \mathbb{R}$ is any constant.

Proof: The smoothed error PDF $(p_E * \kappa_\lambda)(\cdot)$ can be expressed as:

$$\begin{aligned}
 (p_E * \kappa_\lambda)(\xi) &= \int_{\mathbb{R}} p_E(\tau) \kappa_\lambda(\xi - \tau) d\tau \\
 &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}^m} p_{Y|X}(\tau + g(x) | x) dF(x) \right) \kappa_\lambda(\xi - \tau) d\tau \\
 &= \int_{\mathbb{R}^m} \left(\int_{\mathbb{R}} p_{Y|X}(\tau + g(x) | x) \kappa_\lambda(\xi - \tau) d\tau \right) dF(x) \\
 &= \int_{\mathbb{R}^m} \left(\rho(\xi + g(x) | x, \kappa_\lambda) \right) dF(x)
 \end{aligned}
 \tag{18}$$

where $F(x)$ denotes the distribution function of X . From (18), we see that the SMEE criterion can be formulated as a problem of shifting the components of a mixture of the smoothed conditional PDFs so as to minimize the entropy of the mixture. Then Property 6 holds since it has already been proved in [27] that, if all components (conditional PDFs) in the mixture are symmetric and unimodal, the conditional mean (or median) will minimize the mixture entropy. The constant c is added since the entropy is shift-invariant (in practice we usually set $c = 0$).

3. How Smoothing Factor Affects the Optimal Solution

The smoothing factor λ is an important parameter in SMEE criterion, which controls the smoothness of the performance surface. In the following, we will investigate how the smoothing factor affects the optimal solution (optimal mapping) $g^*(\cdot)$.

When $\lambda = 0$, the smoothed error entropy becomes the true error entropy, and the SMEE criterion reduces to the original MEE criterion. When $\lambda > 0$, the two criteria are different and, in general have different solutions. However, in some situations, for any λ , the SMEE criterion yields the same solution as the MEE criterion.

Proposition 1: If the desired output Y is related to the input signal X through the nonlinear regression model $Y = f(X) + N$, where f is an unknown nonlinear mapping, and N is an additive noise that is independent of X and Z , then for any λ , the optimal solution under SMEE criterion is:

$$g^*(x) = f(x) + c \tag{19}$$

Proof: For any mapping $g \in \mathcal{H}$, and any λ , we have:

$$\begin{aligned}
 H(E + \lambda Z) &= H(f(X) + N - g(X) + \lambda Z) \\
 &= H([f(X) - g(X)] + [N + \lambda Z]) \\
 &= H(U + V) \\
 &\geq H(U + V | U) \\
 &= H(V | U) \\
 &\stackrel{(c)}{=} H(V)
 \end{aligned}
 \tag{20}$$

where $U = f(X) - g(X)$, $V = N + \lambda Z$, and (c) comes from the fact that U and V are independent. The equality in (20) holds if and only if U is δ distributed, that is, U is a constant. This can be easily proved as follows.

If U is a constant, the equality in (20) will hold. Conversely, if the equality in (20) holds, we can prove that U is a constant. Actually, in this case, U and $U+V$ are independent, and hence, $\mathbf{E}[(U+V)U] = \mathbf{E}[U+V]\mathbf{E}[U]$. It follows that:

$$\begin{aligned} \mathbf{E}[(U+V)U] &= \mathbf{E}[U+V]\mathbf{E}[U] \\ \Leftrightarrow \mathbf{E}[U^2] + \mathbf{E}[UV] &= (\mathbf{E}[U])^2 + \mathbf{E}[U]\mathbf{E}[V] \\ \Leftrightarrow \mathbf{E}[U^2] + \mathbf{E}[U]\mathbf{E}[V] &= (\mathbf{E}[U])^2 + \mathbf{E}[U]\mathbf{E}[V] \quad (21) \\ \Leftrightarrow \mathbf{E}[U^2] &= (\mathbf{E}[U])^2 \\ \Leftrightarrow \mathbf{E}[(U - \mathbf{E}[U])^2] &= 0 \end{aligned}$$

which implies that the variance of U is zero (*i.e.*, U is a constant). Therefore we have $g^*(x) = f(x) + c$.

Remark 5: Proposition 1 implies that for the nonlinear regression problem, the optimal solution under the SMEE criterion does not change with the smoothing factor provided that the additive noise is independent of the input signal.

If the unknown system (the conditional PDF) is not restricted to a nonlinear regression model, under certain conditions the optimal solution of SMEE can still remain unchanged with the smoothing factor λ . Specifically, the following proposition holds.

Proposition 2: If the conditional PDF $p_{y|x}(y|x)$ and the kernel function $\kappa_\lambda(y)$ are both symmetric (not necessarily about zero) and both unimodal in $y \in \mathbb{R}$, then for any λ , the SMEE criterion produces the same solution:

$$g^*(x) = \int_{\mathbb{R}} yp_{y|x}(y|x)dy + c \quad (22)$$

Proof: By Property 6, it suffices to prove that the smoothed conditional PDF $\rho(\cdot|x, \kappa_\lambda)$ is symmetric and unimodal. This is a well-known result and a simple proof can be found in [28].

Remark 6: Note that the optimal solution under the minimum error variance criterion is also given by (22). In particular, if setting $c = 0$, the solution (22) becomes the conditional mean, which corresponds to the optimal solution under the MSE criterion.

Proposition 3: Assume that the conditional PDF $p_{y|x}(\cdot|x)$ is symmetric (not necessarily unimodal) with uniformly bounded support, and the kernel function κ_λ is a zero-mean Gaussian PDF with variance λ^2 . Then, if smoothing factor λ is larger than a certain value, the optimal solution under the SMEE is still given by (22).

Proof: By Property 6, it is sufficient to prove that if smoothing factor λ is larger than a certain value, the smoothed conditional PDF $\rho(\cdot|x, \kappa_\lambda)$ is symmetric and unimodal. Suppose without loss of generality that the conditional PDF $p_{y|x}(\cdot|x)$ is symmetric about zero with bounded support $[-B, B]$, $B > 0$. Since kernel function $\kappa_\lambda(\cdot)$ is a zero-mean Gaussian PDF with variance λ^2 , the smoothed PDF $\rho(\cdot|x, \kappa_\lambda)$ can be expressed as:

$$\begin{aligned} \rho(y|x, \kappa_\lambda) &= \frac{1}{\sqrt{2\pi\lambda}} \int_{-\infty}^{\infty} \exp\left(-\frac{(y-\tau)^2}{2\lambda^2}\right) p_{Y|X}(\tau|x) d\tau \\ &= \frac{1}{\sqrt{2\pi\lambda}} \int_{-B}^B \exp\left(-\frac{(y-\tau)^2}{2\lambda^2}\right) p_{Y|X}(\tau|x) d\tau \\ &\stackrel{(d)}{=} \frac{1}{\sqrt{2\pi\lambda}} \int_0^B \left\{ \exp\left(-\frac{(y-\tau)^2}{2\lambda^2}\right) + \exp\left(-\frac{(y+\tau)^2}{2\lambda^2}\right) \right\} p_{Y|X}(\tau|x) d\tau \end{aligned} \tag{23}$$

where (d) follows from $p_{Y|X}(\tau|x) = p_{Y|X}(-\tau|x)$. Clearly, $\rho(y|x, \kappa_\lambda)$ is symmetric in y . Further, we derive:

$$\frac{\partial}{\partial y} \rho(y|x, \kappa_\lambda) = \frac{-1}{\sqrt{2\pi\lambda^3}} \int_0^B \left\{ \exp\left(-\frac{(y-\tau)^2}{2\lambda^2}\right)(y-\tau) + \exp\left(-\frac{(y+\tau)^2}{2\lambda^2}\right)(y+\tau) \right\} p_{Y|X}(\tau|x) d\tau \tag{24}$$

If $\lambda \geq 2B$, we have:

$$\begin{cases} \frac{\partial}{\partial y} \rho(y|x, \kappa_\lambda) \leq 0 & \text{for } y \geq 0 \\ \frac{\partial}{\partial y} \rho(y|x, \kappa_\lambda) \geq 0 & \text{for } y < 0 \end{cases} \tag{25}$$

We give below a simple proof of (25). It suffices to consider only the case $y \geq 0$. In this case, we consider two subcases:

(1) $y \geq B$: In this case, we have $\forall \tau \in [0, B]$, $(y-\tau) \geq 0$, and hence:

$$\exp\left(-\frac{(y-\tau)^2}{2\lambda^2}\right)(y-\tau) + \exp\left(-\frac{(y+\tau)^2}{2\lambda^2}\right)(y+\tau) \geq 0, \quad y \in [B, \infty), \tau \in [0, B] \tag{26}$$

(2) $0 \leq y < B$: In this case, we have $\forall \tau \in [0, B]$, $0 \leq |y-\tau| \leq y+\tau \leq 2B \leq \lambda$. Since $\forall |x| \leq \lambda$:

$$\frac{\partial}{\partial x} \left\{ \exp\left(-\frac{x^2}{2\lambda^2}\right)x \right\} = \exp\left(-\frac{x^2}{2\lambda^2}\right) \frac{\lambda^2 - x^2}{\lambda^2} \geq 0 \tag{27}$$

we have $\exp\left(-\frac{(y-\tau)^2}{2\lambda^2}\right)|y-\tau| \leq \exp\left(-\frac{(y+\tau)^2}{2\lambda^2}\right)(y+\tau)$, and it follows easily that:

$$\exp\left(-\frac{(y-\tau)^2}{2\lambda^2}\right)(y-\tau) + \exp\left(-\frac{(y+\tau)^2}{2\lambda^2}\right)(y+\tau) \geq 0, \quad y \in [0, B), \tau \in [0, B] \tag{28}$$

Combining (24), (26) and (28), we get $\forall y \geq 0$, $\frac{\partial}{\partial y} \rho(y|x, \kappa_\lambda) \leq 0$. Therefore (25) holds, which implies that if $\lambda \geq 2B$ the smoothed PDF $\rho(\cdot|x, \kappa_\lambda)$ is symmetric and unimodal, and this completes the proof of the proposition.

Remark 7: Proposition 3 suggests that under certain conditions, when λ is larger than a certain value, the SMEE criterion yields the same solution as the minimum error variance criterion. In the next proposition, a similar but more interesting result is presented for general cases where no assumptions on the conditional PDF and on the kernel function are made.

Proposition 4: When the smoothing factor $\lambda \rightarrow \infty$, minimizing the smoothed error entropy will be equivalent to minimizing the error variance plus an infinitesimal term.

Proof: The smoothed error entropy can be rewritten as:

$$\begin{aligned} H(E + \lambda Z) &= H\left(\lambda \left[\frac{1}{\lambda} E + Z\right]\right) = H\left(\frac{1}{\lambda} E + Z\right) + \log \lambda \\ &= H(Z + \sqrt{t} E) - \frac{1}{2} \log t \end{aligned} \quad (29)$$

where $t = 1/\lambda^2$. Since the term $-\frac{1}{2} \log t$ does not depend on learning machine, minimizing $H(E + \lambda Z)$ is equivalent to minimizing $H(Z + \sqrt{t} E)$, that is:

$$\min_{g \in \mathcal{H}} H(E + \lambda Z) \Leftrightarrow \min_{g \in \mathcal{H}} H(Z + \sqrt{t} E) \quad (30)$$

By De Bruijn's identity (11):

$$\left. \frac{\partial}{\partial t} H(Z + \sqrt{t} E) \right|_{t=0} = \frac{1}{2} \sigma_E^2 J(Z) \quad (31)$$

When λ is very large (hence t is very small):

$$H(Z + \sqrt{t} E) = H(Z) + \frac{t}{2} \sigma_E^2 J(Z) + o(t) \quad (32)$$

Combining (30) and (32) yields:

$$\min_{g \in \mathcal{H}} H(E + \lambda Z) \Leftrightarrow \min_{g \in \mathcal{H}} \left(\sigma_E^2 + \frac{2}{J(Z)t} o(t) \right) \quad (33)$$

which completes the proof.

Remark 8: The above result is very interesting: when the smoothing factor λ is very large, minimizing the smoothed error entropy will be approximately equivalent to minimizing the error variance (or the mean square error if the error PDF is restricted to zero-mean). This result holds for any conditional PDF and any kernel function. A similar result can be obtained for the nonparametric entropy estimator based on Parzen windows. It was proved in [14] that in the limit, as the kernel size (the smoothing factor) tends to infinity, the entropy estimator approaches a nonlinearly scaled version of the sample variance.

4. Conclusions

Traditional machine learning methods mainly exploit second order statistics (covariance, mean square error, correlation, etc.). The optimality criteria based on second order statistics are

computationally simple, and optimal under linear and Gaussian assumptions. Although second order statistics are still prevalent today in the machine learning community and provide successful engineering solutions to most practical problems, it has become evident that this approach can be improved, especially when data possess non-Gaussian distributions (multi-modes, fat tails, finite range, *etc.*). In most situations, a more appropriate approach should be applied to capture higher order statistics or information content of signals rather than simple their energy. Recent studies suggest that the supervised machine learning can benefit greatly from the use of the minimum error entropy (MEE) criterion. To implement the MEE learning, however, one has to estimate the error entropy from the samples. In the limit (when sample size tends to infinity), the estimated error entropy by Parzen windowing converges to the smoothed error entropy, *i.e.*, the entropy of the error plus an independent random variable with PDF equal to the kernel function used in Parzen windowing, so the smoothed error entropy is the actual entropy that is minimized in the MEE learning.

In this paper, we study theoretically the properties of the smoothed MEE (SMEE) criterion in supervised machine learning and, in particular, we investigate how the smoothing factor affects the optimal solution. Some interesting results are obtained. It is shown that when the smoothing factor is small, the smoothed error entropy equals approximately the true error entropy plus a scaled version of the Fisher information of error. In some special situations, the SMEE solution remains unchanged with increasing smoothing factor. In general cases, however, when the smoothing factor is very large, minimizing the smoothed error entropy will be approximately equivalent to minimizing the error variance (or the mean square error if the error distribution is restricted to zero-mean), regardless of the conditional PDF and the kernel function.

This work does not address the learning issues when the number of samples is limited. In this case, the problem becomes much more complex since there is an extra bias in the entropy estimation. We leave this problem open for future research. The results obtained in this paper, however, are still very useful since they provide theoretical solutions to which the empirical solution (with finite data) must approximate.

Acknowledgments

This work was supported by NSF grant ECCS 0856441, ONR N00014-10-1-0375, and National Natural Science Foundation of China (No. 60904054).

References

1. Cover, T.M.; Thomas, J.A. *Element of Information Theory*; John Wiley & Sons, Inc.: New York, NY, USA, 1991.
2. Weidemann, H.L.; Stear, E.B. Entropy analysis of estimating systems. *IEEE Trans. Inform. Theor.* **1970**, *16*, 264–270.
3. Tomita, Y.; Ohmatsu, S.; Soeda, T. An application of the information theory to estimation problems. *Inf. Control* **1976**, *32*, 101–111.
4. Janzura, M.; Koski, T.; Otahal, A. Minimum entropy of error principle in estimation. *Inf. Sci.* **1994**, *79*, 123–144.

5. Wolsztynski, E.; Thierry, E.; Pronzato, L. Minimum-entropy estimation in semi-parametric models. *Signal Process.* **2005**, *85*, 937–949.
6. Chen, B.; Zhu, Y.; Hu, J.; Zhang, M. On optimal estimations with minimum error entropy criterion. *J. Frankl. Inst. Eng. Appl. Math.* **2010**, *347*, 545–558.
7. Kalata, P.; Priemer, R. Linear prediction, filtering and smoothing: An information theoretic approach. *Inf. Sci.* **1979**, *17*, 1–14.
8. Feng, X.; Loparo, K.A.; Fang, Y. Optimal state estimation for stochastic systems: An information theoretic approach. *IEEE Trans. Automat. Contr.* **1997**, *42*, 771–785.
9. Guo, L.; Wang, H. Minimum entropy filtering for multivariate stochastic systems with non-Gaussian Noises. *IEEE Trans. Autom. Control* **2006**, *51*, 695–700.
10. Pfaffelhuber, E. Learning and information theory. *Int. J. Neurosci.* **1972**, *3*, 83–88.
11. Barlow, H. Unsupervised learning. *Neural Comput.* **1989**, *1*, 295–311.
12. Erdogmus, D.; Principe, J.C. An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems. *IEEE Trans. Signal Process.* **2002**, *50*, 1780–1786.
13. Erdogmus, D.; Principe, J.C. Generalized information potential criterion for adaptive system training. *IEEE Trans. Neural Netw.* **2002**, *13*, 1035–1044.
14. Erdogmus, D.; Principe, J.C. Convergence properties and data efficiency of the minimum error entropy criterion in Adaline training. *IEEE Trans. Signal Process.* **2003**, *51*, 1966–1978.
15. Erdogmus, D.; Principe, J.C. From linear adaptive filtering to nonlinear information processing—The design and analysis of information processing systems. *IEEE Signal Process. Mag.* **2006**, *23*, 14–33.
16. Santamaria, I.; Erdogmus, D.; Principe, J.C. Entropy minimization for supervised digital communications channel equalization. *IEEE Trans. Signal Process.* **2002**, *50*, 1184–1192.
17. Chen, B.; Hu, J.; Pu, L.; Sun, Z. Stochastic gradient algorithm under (h, ϕ) -entropy criterion, *Circuits Syst. Signal Process.* **2007**, *26*, 941–960.
18. Principe, J.C. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*; Springer: New York, NY, USA, 2010.
19. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995.
20. Scholkopf, B.; Smola, A.J. *Learning with Kernels, Support Vector Machines, Regularization, Optimization and Beyond*; MIT Press: Cambridge, MA, USA, 2002.
21. Liu, W.; Principe, J.C. *Kernel Adaptive Filtering: A Comprehensive Introduction*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2010.
22. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Chapman & Hall: New York, NY, USA, 1986.
23. Beirlant, J.; Dudewicz, E.J.; Györfi, L.; van der Meulen, E.C. Nonparametric entropy estimation: An overview. *Int. J. Math. Statist. Sci.* **1997**, *6*, 17–39.
24. Linnik, Ju.V.; Ostrovskii, I.V. *Decompositions of Random Variables and Vectors*; American Mathematical Society: Providence, RI, USA, 1977.
25. Rioul, O. Information theoretic proofs of entropy power inequalities. *IEEE Trans. Inform. Theor.* **2011**, *57*, 33–55.

26. Xu, J.-W.; Erdogmus, D.; Principe, J.C. Minimizing Fisher information of the error in supervised adaptive filter training. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Quebec, Canada, 17–21 May 2004.
27. Chen, T.-L.; Geman, S. On the minimum entropy of a mixture of unimodal and symmetric distributions. *IEEE Trans. Inf. Theor.* **2008**, *54*, 3166–3174.
28. Purkayastha S. Simple proofs of two results on convolutions of unimodal distributions. *Statist. Prob. Lett.* **1998**, *39*, 97–100.

© 2012 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).