*LETTER* Communicated by José C. Príncipe

# The MEE Principle in Data Classification: A Perceptron-Based Analysis

**Luís M. Silva**
*lmsilva@fe.up.pt*
**J. Marques de Sá**
*jpmdesa@gmail.com*
*Instituto de Engenharia Biomédica, Divisão de Sinal e Imagem, Porto 4200-465, Portugal*

**Luís A. Alexandre**
*lfbaa@di.ubi.pt*
*Departamento de Informática, Universidade da Beira Interior and Instituto de Telecomunicações, 6201-001 Covilhã, Portugal*

**This letter focuses on the issue of whether risk functionals derived from information-theoretic principles, such as Shannon or Rényi's entropies, are able to cope with the data classification problem in both the sense of attaining the risk functional minimum and implying the minimum probability of error allowed by the family of functions implemented by the classifier, here denoted by min *Pe*. The analysis of this so-called minimization of error entropy (MEE) principle is carried out in a single perceptron with continuous activation functions, yielding continuous error distributions. In spite of the fact that the analysis is restricted to single perceptrons, it reveals a large spectrum of behaviors that MEE can be expected to exhibit in both theory and practice. In what concerns the theoretical MEE, our study clarifies the role of the parameters controlling the perceptron activation function (of the squashing type) in often reaching the minimum probability of error. Our study also clarifies the role of the kernel density estimator of the error density in achieving the minimum probability of error in practice.**

## 1 Introduction

Information-theoretic learning (ITL) is an area of research enjoying growing interest and promising new and important breakthroughs in many applications. Its introduction can be traced back at least to Linsker (1988), who introduced the maximization of mutual information between the input and output of a neural network applied, for example, to feature extraction. But the real blossoming of ITL dates back more recently when the minimization of Rényi's quadratic entropy of the difference between the system

output and its desired target for solving regression problems was proposed (Erdogmus & Príncipe, 2000). This was followed by a large set of ITL theoretical results and applications developed by Príncipe and coworkers in time-series prediction (Erdogmus & Príncipe, 2002), feature extraction, clustering (Gokcay & Príncipe, 2002; Jenssen, Hild, Erdogmus, Príncipe, & Eltoft, 2003), and blind source separation (Hild, Erdogmus, & Príncipe, 2001; Erdogmus, Hild, & Príncipe, 2002). The rationale is as follows. Having an adaptive system with output variable $Y$ and target variable $T$, the minimization of the error entropy (MEE), that is, the entropy of $E = T - Y$, implies a reduction of the expected information contained in the error ("error" meaning here, as in adaptive systems terminology, the target-output deviation), leading to the maximization of the mutual information between the desired target and the system output (Erdogmus & Príncipe, 2000, 2002). This means that the system is learning the target variable. Entropy-based cost functions, since they depend on the full probability distribution of $E$, reflect the global behavior of the error distribution; therefore, learning systems with entropic cost functions may often outperform those using the classic and popular mean square error (MSE) cost, which reflects only the second-order statistics of the error. An objection that could be raised to using the MEE principle in the case of continuous error distributions is the need to estimate the probability density function (pdf) of $E$, since it is well known that accurate pdf estimation may be a tougher problem than having to solve a related regression or classification problem. However, it turns out that when the MEE principle using Rényi's entropy is applied, pdf estimation is short-circuited altogether (Príncipe, Xu, & Fisher, 2000). Even if one uses Shannon's entropy, usually a simple and coarse pdf estimate is all that is needed (Silva, Marques de Sá, & Alexandre, 2005).

The application of the MEE principle to solving data classification problems has been carried out by our team and divulged in several papers (the principle is coined EEM in these references), using either MLPs (Santos, Marques de Sá, Alexandre, & Sereno, 2004; Santos, Marques de Sá, & Alexandre, 2005; Silva et al., 2005; Santos, 2007) or recurrent networks (Alexandre & Marques de Sá, 2006). It has been applied with success in classifiers using a kernel-based approach (Han, 2007). We have also applied entropic cost functions with excellent results in a new data clustering algorithm (Santos, Marques de Sá, & Alexandre, 2007). A careful comparison study on the performance of classifiers applied to real-world data sets was carried out showing the high competitiveness of the MEE method (Silva, Embrechts, Santos, & Marques de Sá, 2008). Despite the evidence of good performance provided by the experimental results presented in these references, very little is known about the properties of MEE when applied to data classification.

For that purpose, we consider a classification problem with a set of classes $T = \{t\}$ and a parametric machine (parameter set $W = \{w\}$) performing a mapping $Y = \varphi_w(X)$, where $X$ and $Y$ are the input and output spaces,

respectively (we use capital letters denoting variables and their supports). The machine is trained by some algorithm to minimize a risk functional on the parameter set $W$ of the function class $\Phi = \{\varphi_w\}$, implemented by the classifier, which is often written for continuous data distributions as

$$R_\Phi \equiv R_\Phi(w) = \sum_T P(t) \int_X L(t, y) f(x \mid t)\, dx \quad \text{with } y = \varphi_w(x), \tag{1.1}$$

where $T$ is the target space, $f(x \mid t) \equiv f_X(x \mid t)$ is the conditional density function of the inputs, and $P(t)$ are the class priors. The target-output distance, that is, the cost function $L(\cdot)$, can be chosen in various ways. For instance, for MSE, $L(t, y) = (t - y)^2$, and for cross-entropy (CE) (Bishop, 1995) and two-class problems with $Y \in [-1, 1]$ and $T \in \{-1, 1\}$, $L(t, y) = \ln(1 + ty)$. Minkowski and exponentially weighted distances have also been proposed. The risk functional for MEE is written not as a distance functional but as a functional of the error pdf $f(e) \equiv f_E(t - \varphi_w(x))$ (assuming it exists), namely, as

$$R_\Phi \equiv H_S(E) = -\int_E f(e) \ln f(e) de \tag{1.2}$$

for the Shannon entropy of the error, $H_S$, or as

$$R_\Phi \equiv H_{R_2}(E) = -\ln \int_E f^2(e) de \tag{1.3}$$

for the quadratic Rényi entropy, $H_{R_2}$.

The main problem in data classification, which we refer to as the *classifier problem*, is whether it is possible to attain the minimum probability of error afforded by the machine architecture, that is, by the family of functions $\Phi$, for some $w^*$—the so-called optimal solution. Let us denote by $\min_W Pe_\Phi$ the minimum probability of error, achievable in $\Phi$.[1] From now on, whenever we refer to the optimal solution, $w^*$, we always mean optimal in the $\min_W Pe_\Phi$ sense. The classifier problem corresponds to the following question: Does $\min_W R_\Phi$ imply $\min_W Pe_\Phi$? (Note that $\min_W Pe_\Phi$ corresponds in the distance functional to setting $L(t, y) = \{0, \text{ if } t = y\ ; 1, \text{ otherwise}\}$; however, we are interested only in risk functionals with continuous integrands, for which efficient optimization algorithms exist.) For instance, if hypothetically $\min_W R_\Phi$ does not lead to $\min_W Pe_\Phi$, one has to conclude that a risk functional is being used that fails to adequately take into account the whole $\Phi$ set complexity. One should then turn to another risk functional. This essential problem has been somewhat overlooked. Concerning MSE, the main

[1] For some architectures, $\min_W Pe_\Phi$ may correspond to the optimal Bayes' error. However, this issue will not occupy us here.

and often mentioned results are that for gaussian distributions, MSE yields the optimal regression solution and the outputs of a neural network (NN) trained with MSE correspond to Bayesian posterior probabilities (Bishop 1995; Richard & Lippmann, 1991), which allow some confidence that MSE will also perform well in classification problems. However, MSE may dramatically fail in classification problems where MEE performs in the optimal way, as shown in appendix C. Since MEE is a more sophisticated approach than the often used MSE or CE, because it takes into account the whole distribution of the errors, and given the large amount of good experimental results obtained with MEE, it seems worthwhile to investigate the classifier problem with MEE. In this investigation, many interesting aspects and new insights come to light. In previous work (Silva, Felgueiras, Alexandre, & Marques de Sá, 2006), we showed that for univariate data and the Stoller split setting (Stoller 1954), a popular setting in decision trees, the MEE principle does not always lead to $\min_W Pe_\Phi$ (or simply $\min Pe$), and we were able to rigorously state the very general conditions when it does. In this work, we go a step further and investigate the behavior of perceptrons trained with MEE using continuous activation functions (a.f.).

The organization of the letter is as follows. Section 2 introduces notation and presents the error entropy expressions for continuous density distributions of the error. We also show in section 2 that for data classification, the MEE principle is harder to apply than for regression. Section 3 starts by comparing MSE and CE in terms of $f(e)$ functionals, elucidates how the practical implementation of MEE (empirical MEE) differs from its theoretical formulation, and analyzes how classifier problems are solved by perceptrons using theoretical and empirical MEE. In section 4 we analyze in detail the influence of the kernel density estimator in achieving an error entropy minimum. Finally, in section 5, we discuss the results and draw the main conclusions.

## 2 MEE Is Harder for Classification Than for Regression

**2.1 The Error Entropy for Data Classification.** We consider two-class problems where a given instance $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$ from $X$ is to be classified in one of two classes, $\mathcal{C}_{-1}$ or $\mathcal{C}_1$, the target set is $T \in \{-1, 1\}$, and a machine (e.g., NN) implements a parameterized function family $\Phi = \{\varphi_w\}$, $w \in W$ and issues a single output $y \in [-1, 1]$. Any other supports for $T$ and $Y$ could be used. The ones indicated are used for ease of computation only. The output random variable (r.v.) $Y$ is assumed to be continuous with a conditional pdf $f_{Y|t}(y)$. With this setting, the density function of the error r.v. $E = T - Y$ can be derived as

$$f_E(e) = P(1) f_{Y|1}(1 - e) + P(-1) f_{Y|-1}(-1 - e), \tag{2.1}$$

where $P(t) \equiv P(T = t)$ for $t \in \{-1, 1\}$ are the priors, often denoted $q$, $p$, respectively. These definitions imply necessarily that $E \in [-2, 2]$ and that

each $f_{Y|t}(t-e)$ lies in separate intervals $[t-1, t+1]$. As a consequence of equation 2.1, the (differential) Shannon's entropy of the error $H_S(E)$ (or simply $H_S$) can be decomposed as

$$H_S = pH_{S|1} + qH_{S|-1} + H_S(T), \tag{2.2}$$

where $H_{S|t}$ is the Shannon's entropy of the error for class $\mathcal{C}_t$ and $H_S(T) = \sum_{t\in T} P(t)\ln P(t)$ is the Shannon's entropy of the priors. Rényi's entropy also satisfies a similar additivity property in the exponential scale (see appendix A for both derivations). The distributions and entropies are functions of $w$, the machine parameter vector, although we omit this dependency for the sake of simpler notation.

**2.2 The Minimum of the Error Entropy.** Looking at equation 2.2 and since $H_S(T)$ is a constant, $\min H_S$ implies $\min\{pH_{S|1} + qH_{S|-1}\}$. Thus, in general, one can say nothing about the minimum (location and value) since it will depend on the particular shapes of $H_{S|t}$ as functions of $w$, and the particular value of $p$. Nevertheless, the minimum entropy distribution (with a value of $-\infty$) is the Dirac distribution $\delta_X(x; a)$ (centered at $a$). Thus, although $\min\{pH_{S|1} + qH_{S|-1}\}$ achieves its lowest value only when both $f_{Y|1}(1-e)$ and $f_{Y|-1}(-1-e)$ are Dirac pdf's, a very low value can be achieved if at least one of $f_{Y|t}$ is nearly Dirac.

**2.3 The Minimum of the KL Divergence.** An important result concerning the (Shannon's) entropy of the error minimum was presented by Erdogmus and Príncipe (2002). These authors demonstrated that the MEE principle corresponds to the minimum of the Kullback-Leibler (KL) divergence between the joint distributions $f_{X,Y}$ and $f_{X,T}$. This probability density matching result was demonstrated for the regression setting. However, for the data classification setting, two difficulties arise. First, for the regression setting, one may write $f_E(e) = f_{Y|x}(d-e \mid x)$ as in the cited paper, since there is only one distribution of $y$ values and $d$ can be seen as the mean of the $y$ values. However, for the classification setting, one has to write $f_{E|t}(e \mid t) = f_{E|t,x}(d-e \mid t, x)$. That is, one has to study what happens to each class-conditional distribution individually and therefore to individually study the KL divergence relative to each class-conditional distribution, that is,

$$KL_t = \int_X \int_Y f_{XY|t}(x, y) \ln \frac{f_{XY|t}(x, y)}{d_{XY|t}(x, y)}, \tag{2.3}$$

where $d_{XY|t}(x, y)$ is the desired input-output probability density function. Second, the KL divergence is undefined whenever $d_{XY|t}(x, y)$ has zeros in the supports of $X$ and $Y$. This problem, which may or may not be present in the regression setting, is almost always present in the classification setting,

since the desired input-output probability densities' functions are continuous functions with zeros in their supports (Dirac functions in the separable case).

Even if we relax the conditions on the desired input-output probability density, for instance, by choosing functions with no zeros on the $Y$ support but sufficiently close to Dirac functions, we may not yet reach the MEE condition for classification because of section 2.2. Attaining the KL minimum for a class-conditional distribution says nothing about the other class-conditional distribution and about $H_S$.

## 3 MEE with Continuous Errors

**3.1 MSE, CE, and MEE.** Although the risk functionals are usually presented and studied as in equation 1.1, that is, relative to the $X \times T$ space, we may as well analyze them in other spaces. In fact, in order to appreciate how the various risk functionals cope with the classifier problem, it turns out to be worth expressing them in terms of the error r.v. We now proceed to do exactly this for the MSE, CE, and MEE functionals. In the following derivations, we often use the well-known theorem of r.v. transformation (see Rényi, 1970):

**Theorem 1.** *Let $f(x)$ be the pdf of the r.v. $X$. Assume $\varphi(x)$ to be monotonic and differentiable, and suppose $\varphi'(x) \neq 0\ \forall x$. If $g(y)$ is the density of $Y = \varphi(X)$, then*

$$g(y) = \begin{cases} \dfrac{f(\varphi^{-1}(y))}{|\varphi'(\varphi^{-1}(y))|}, & \inf \varphi(x) < y < \sup \varphi(x) \\ 0, & otherwise \end{cases}, \tag{3.1}$$

*where $x = \varphi^{-1}(y)$ is the inverse function of $y = \varphi(x)$.*

We also assume monotonic increasing $\varphi(x)$ functions (e.g., NN squashing activation functions) with $\varphi'(x) > 0$. Simple mathematical manipulations lead us to

$$R_\Phi = \sum_{t \in \{-1,1\}} P(t) \int_{t-1}^{t+1} L(t,e) f(e \mid t)\, de. \tag{3.2}$$

For MSE, $L(t,e) = (t-y)^2 = e^2$ depends only on $e$. Moreover, $R_\Phi = \mathbb{E}_{T,E}\{e^2\}$ (where $\mathbb{E}\{\cdot\}$ means expected value), which is empirically estimated as

$$MSE = \frac{1}{n} \sum_{t_i \in \{-1,1\}} \sum_i (t_i - y_i)^2. \tag{3.3}$$

Let us now consider the other classic risk functional, the so-called cross-entropy (Bishop, 1995), which in its popularized empirical form (based on

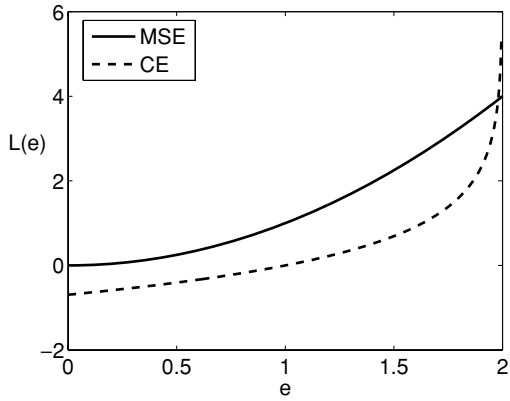



Figure 1: $L_{MSE}$ and $L_{CE}$ ($t = 1$) as functions of $e$.

sample outputs $y_i$ and targets $t_i$) is expressed for the same two-class setting as

$$CE = -\sum_{t_i=-1} \ln\left(\frac{1-y_i}{2}\right) - \sum_{t_i=1} \ln\left(\frac{1+y_i}{2}\right).$$

CE, in terms of the error r.v., is $n$ times the empirical estimate of the following risk functional:

$$R_\Phi = \sum_{t\in\{-1,1\}} P(t) \int_{t-1}^{t+1} \ln\left(\frac{1}{2-te}\right) f_E(e \mid t)\, de + \ln 2. \tag{3.4}$$

Thus, for cross-entropy $L(t,e) = \ln\left(\frac{1}{2-te}\right)$. Figure 1 shows (for $t = 1$) the distance functions $L_{MSE}(e) = e^2$ and $L_{CE}(e) = \ln\left(\frac{1}{2-te}\right)$. A machine minimizing the MSE risk functional is minimizing the second-order moment of the errors, favoring input-output mappings with low error spread and deviation from zero. A machine minimizing the CE risk functional is minimizing an average logarithmic distance of the error from its worst value (2 for $t = 1$ and $-2$ for $t = -1$), as shown in Figure 1. As a consequence of this logarithmic behavior $L_{CE}(e)$ tends to focus mainly on large errors, as opposed to $L_{MSE}(e)$.

We now rewrite expression 1.2 in a similar way as in expression 3.2:

$$H_S = \sum_{t\in\{-1,1\}} P(t) \int_{t-1}^{t+1} \ln\left(\frac{1}{f(e \mid t)}\right) f(e \mid t)\, de + H_S(T). \tag{3.5}$$

As for the Shannon entropy, a similar expression can be written for $V_{R_2} = \exp(-H_{R_2})$. We observe in equation 3.5 that $-\ln f(e \mid t)$ plays the role of the cost function. The difference relative to MSE and CE (and other conventional risk functionals) is that when $H_S$ is applied to train a given machine, its properties change from iteration to iteration. Therefore, instead of studying its properties based on a functional description or graph as in Figure 1, we are compelled to study it from the point of view of entropy properties.

**3.2 Theoretical MEE and Empirical MEE.** As with any other risk functional, the practical application of the MEE approach relies on using adequate estimates, in this case, of the error pdf. This can be achieved with the kernel density estimator (kde) (Parzen, 1962), which, given a data set $x_1, x_2, \ldots, x_n$, provides the following estimate of the pdf $f(x)$,

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{3.6}$$

where $K$ is a kernel function and $h$ is the kernel bandwidth. The kde estimate can also be expressed in terms of a convolution:

$$\hat{f}(x) = f_n(x) * K_h(x), \tag{3.7}$$

where $f_n(x)$ is the empirical pdf, $K_h(x) = K(x/h)$, and $*$ is the convolution operator. The standardized gaussian pdf $G(x)$ enjoys desirable properties as a kernel function; is popularly used and is the one we consider. The estimated pdf, is always a smoothed version of the original pdf, and this will show up (for appropriate choices of $h$) as a fundamental feature of the practical MEE implementation.

With $\hat{f}(x)$, one can then estimate the error entropies as follows:

$$H_S = \mathbb{E}\{-\ln f(e)\} : \hat{H}_S = -\frac{1}{n} \sum_{i=1}^{n} \ln f(e_i) \approx -\frac{1}{n} \sum_{i=1}^{n} \ln \hat{f}(e_i), \tag{3.8}$$

$$H_{R_2} = -\ln \mathbb{E}\{f(e)\} : \hat{H}_{R_2} = -\ln \frac{1}{n} \sum_{i=1}^{n} f(e_i) \approx -\ln \frac{1}{n} \sum_{i=1}^{n} \hat{f}(e_i). \tag{3.9}$$

These are precisely the formulas used in the backpropagation error algorithm in MLP training with MEE (Silva et al., 2005; Santos, 2007). These formulas correspond to what we call *empirical MEE*.

In what follows we are also interested in analyzing the theoretical behavior of the MEE risk functionals. However, the theoretical MEE can be analyzed mathematically only in simple situations of error pdf's, such as uniform and gaussian. For more realistic situations with known, albeit more complex, distributions, one has to resort to numerical simulation, which can be carried out as follows. Generate a large number of samples of each input class-conditional pdf; compute the output pdf's with the kde approach;

finally compute $H_S$ using equation 2.2 (see also appendix A for $H_{R_2}$). "Large number of samples" means that the value of $n$ being used in equation 3.6 guarantees a very low integrated mean square error (say, $IMSE < 0.01$) of $\hat{f}(x)$ computed with the optimal $h(n)$. For this purpose (choice of optimal $h(n)$) one can use the formulas given in Thompson and Tapia (1978).

The main differences between empirical and theoretical MEE are as follows:

- Whereas the theoretical MEE implies the separate estimate of $f(e \mid t)$, the empirical MEE relies on the estimate of the whole $f(e)$.
- For this reason, the theoretical MEE cannot be applied in discriminative training (at each training epoch the $f(e \mid t)$ are not easily computable); one may, however, compute the theoretical MEE in a neighborhood of a weight vector, as we do in following sections.
- Whereas the kernel smoothing effect when using the optimal $h(n)$ in the computation of the theoretical MEE can be neglected, the smoothing effect can be made arbitrarily large when applying empirical MEE.

**3.3 The Quest for Minimum Entropy.** Although pattern recognition is a quest for minimum entropy (Watanabe, 1981), the topic of entropy-minimizing distributions has only occasionally been studied. Whereas entropy-maximizing distributions obeying given constraints are well known, minimum entropy distributions on the real line are often difficult to establish (Kapur, 1993). Nevertheless, it is known that the minimum entropy ($-\infty$) of unconstrained continuous distributions occurs for an infinite family of Dirac combs (including the single Dirac density). Entropy magnitude is often associated with the magnitude of the distribution tails, in the sense that larger tails imply higher entropy. However, this fails even in simple cases of constrained densities: the unit variance gaussian, $G(x; 0, 1)$ has a smaller tail (in the sense that for positive $x$, $\exists x_0,\ \forall x > x_0,\ f(x) > g(x)$) than the unit variance bilateral exponential, $e(x; \sqrt{2}) = \lambda \exp(-\sqrt{2}|x|)/2$; however, the former has a larger entropy, $\sqrt{2\pi e} = 2.84$, than the latter, $1 + \ln(\sqrt{2}) = 2.41$. Notwithstanding these difficulties, one can still present three properties of $H_S(E)$ that are useful in justifying experimental findings or guiding the search of experimental settings (as in appendix C):

1. $H_S$ is invariant to translations and partitions of $f_E(e)$ components. This property, a direct consequence of the well-known entropy invariance to translations, and of a result on density partitions presented in appendix A, is illustrated in Figure 2. This is a property that may lead MEE to perform worse than MSE or CE in some cases where an equal probability of error may correspond to distinct configurations.
2. In a large class of $f_E(e)$ families $H_S$, increases with the variance. This property (illustrated in Figure 2) is explained in appendix B, where the meaning of "large class" is also clarified. This property is associated

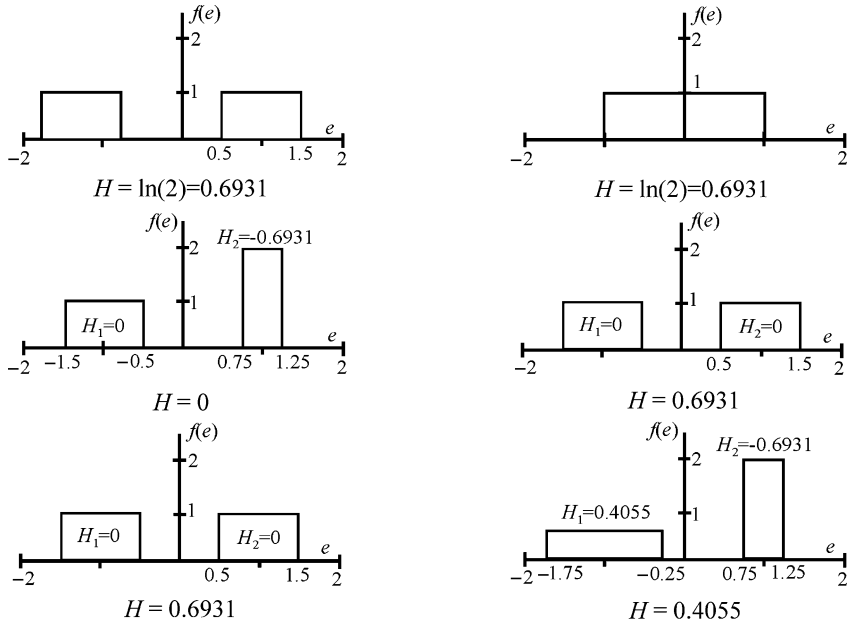



Figure 2: Error $H_S$ properties for $p = q = 1/2$. Top row (property 1): Entropy is invariant to partitions and translations. Middle row (property 2): In the same pdf family, an increase in variance implies an increase in entropy (see the text). Bottom row (property 3): The standard deviation of the right component decreased by $0.5/\sqrt{12}$ while the other increased by the same amount; however, the decrease in entropy of the right component more than compensated for the increase in entropy of the left component.

with the common idea that within the same density family, "longer tails" (in the sense of larger variance) imply larger entropy. Although there are exceptions to this rule, one can quite safely use it in the case of $f_E(e)$ densities. As a consequence, one can say that MEE favors the "order" of the errors.

3. Whenever $f_E(e)$ has two components of equal functional form and equal priors, then for a large class of $f_E(e)$ families, $H_S$ will decrease when the smaller (or equal) standard deviation component decreases while the standard deviation of the other component increases by the same amount (keeping the functional form unchanged). This property (illustrated in Figure 2) is a consequence of the fact that entropy is an up-saturating function of the standard deviation (as well as of the variance) for a large class of $f_E(e)$ families (see appendix B); therefore, although the larger variance component dilates, the corresponding increase in entropy is outweighed by the decrease in entropy of the

other component. As a consequence of this property, MEE is more tolerant than MSE or CE to tenuous tails or outlier errors, as exemplified in appendix C.

The quadratic Rényi entropy also enjoys these three properties.

We may then expect different behaviors between empirical MEE and theoretical MEE. In particular, the minimum value that can be attained by the empirical entropies $\hat{H}_S$ or $\hat{H}_{R_2}$ occurs necessarily for $\delta_E(e;0)$ (Erdogmus & Príncipe 2003; Silva 2008). This constrasts with theoretical entropies where the minimum value is attained by any Dirac comb, as mentioned above. For example, due to the smoothing effect around the origin introduced by the kde, empirical MEE can distinguish between the two situations illustrated in the top of Figure 2 (property 1), favoring the right one. Also, for property 3, empirical MEE will favor smaller tails. An example of this type of situation is illustrated in section 4. Both situations have an important impact on the classifier's performance.

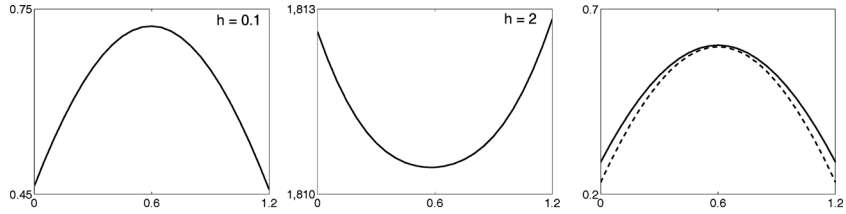
**3.4 The Split-Type Setting.** We start by considering perceptrons applied to two-class problems, with only one output given by $y = \tanh(x - w_0)$, therefore, with a single adjustable parameter. More precisely, the perceptron is trained to find a split point between the classes. Despite the simplicity of this setting, it allows us to establish the connection with the discrete error setting (step function instead of tanh) already studied in Silva et al. (2006) and to elucidate fundamental differences in the behavior of $H_S$ and $\hat{H}_S$ (or $H_{R_2}$ and $\hat{H}_{R_2}$).

We first assume uniform class-conditional inputs ( allowing the analytical derivation of the theoretical entropies):

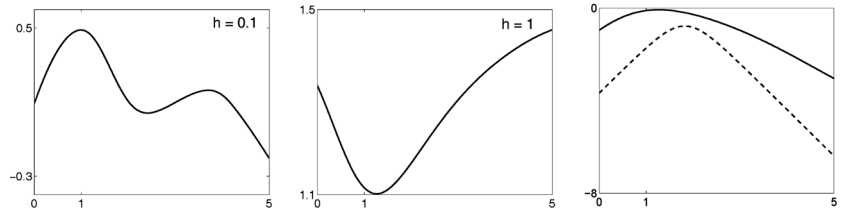
$$f_{X|-1}(x) = \frac{1}{b-a} I_{[a,b]}(x) \qquad f_{X|1}(x) = \frac{1}{d-c} I_{[c,d]}(x), \quad \text{with } a < c \leq b < d, \tag{3.10}$$

where $I$ is the indicator function. The error class-conditional pdf's can be obtained using theorem 1, and both Shannon and Rényi entropies can be then derived. For example, Rényi's formula comes as

$$H_{R_2} = -\ln\left[ -\frac{q^2}{4}\left[\frac{2+e(2+e)\ln\left(\frac{|e|}{2+e}\right)+2e}{(b-a)^2(2+e)e}\right]_{-1-\tanh(b-w_0)}^{-1-\tanh(a-w_0)} + \right.$$
$$\left. + \frac{p^2}{4}\left[\frac{2+\ln\left(\frac{e}{|e-2|}\right)e(e-2)-2e}{(d-c)^2(e-2)e}\right]_{1-\tanh(d-w_0)}^{1-\tanh(c-w_0)}\right]. \tag{3.11}$$




(a) $c = 0.2,\ d = 1.2$




(b) $c = 0.8,\ d = 5$

Figure 3: In the left and middle figures $\hat{H}_S$ is plotted as a function of $w_0$ for different values of $h$ and different $\mathcal{C}_1$ supports $[c, d]$. The right-most figures present the corresponding theoretical curves of Shannon (solid) and Rényi's (dashed) entropies.

It is then possible to show that neither theoretical entropies has a minimum at the optimal split (for class conditionals with equal-length supports, they have a maximum; Silva, 2008). This contrasts with the discrete error setting where for uniform class conditionals, the optimal solution is always at an entropy minimum (Silva et al., 2006).

Figure 3 compares the behaviors of both theoretical and empirical (2000 points per class) entropy curves as functions of the split parameter $w_0$, varying the class-conditional setting. In all cases we fix $[a, b] = [0, 1]$ and set $p = q = 1/2$.

In the top row of Figure 3, the class conditionals have equal-length supports, which means that the optimal solution is any point of the overlapped region, $[c, 1]$. If $h$ is too small, $\hat{H}_S$ reveals a maximum at the optimal split, just as $H_S$; above a sufficiently large $h$, $\hat{H}_S$ shows a minimum at the optimal split (with higher $h$ for increased overlap). Note that entropy identifies the middle point of the overlapped region as its optimal location, corresponding to the situation of equal class error probability. The bottom row of Figure 3 illustrates the unequal class error probabilities case (due to the increased $\mathcal{C}_1$ support), where the optimal split is $w_0^* = 1$. Both the theoretical and empirical curves fail to find $w_0^*$. However, the empirical minimum occurs in a close neighborhood of $w_0^*$.

We also considered gaussian class-conditional inputs (providing absolutely continuous pdf's), reaching essentially the same conclusions.

**3.5 The Perceptron Setting.** We now assume a more realistic perceptron with $y = \tanh(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0)$, that is, we now control the $\varphi(x)$ function shape.

*3.5.1 Gaussian Inputs.* To derive the error pdf for gaussian input distributions, we take into account that gaussianity is preserved under linear transformations: if $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$ has multivariate gaussian distribution, $\mathbf{x} \sim G_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\mathbf{z} = \mathbf{W}\mathbf{x} + \mathbf{w}_0 \sim G_m(\mathbf{W}\boldsymbol{\mu} + \mathbf{w}_0, \mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^{\mathrm{T}}).$$

Making use of theorem 1, we obtain the error distribution for each class as follows:

$$f_{Y|t}(t-e) = \frac{\exp\left(-\frac{1}{2}\frac{\left(\operatorname{arctanh}(t-e)-(\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}_t+w_0)\right)^2}{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}_t\mathbf{w}}\right)}{\sqrt{2\pi\,\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}_t\mathbf{w}}\;e(2t-e)} I_{[t-1,t+1]}(e). \tag{3.12}$$

Numerical integration can now be applied to determine $H_S$ (or $H_{R_2}$).

We first considered the univariate case and the perceptron defined by $y = \tanh(w_1 x + w_0)$, with the parameter $w_1$ controling the shape of the activation function (its steepness). The computed values of $H_S$ (or $H_{R_2}$) show that for appropriate choices of $w_0$ and $w_1$, it is possible to turn the theoretical entropy maximum into a minimum and with a higher value of the shaping parameter $w_1$ as the classes get closer. Moreover, this minima are optimal (correspond to min $Pe$). These results suggest the need of to use function-shaping parameters (as is the case with multilayer perceptrons) in order to reach the theoretical entropy minima.

For the bivariate case, we fix $\boldsymbol{\mu}_{-1} = (0, 0)$, $\boldsymbol{\Sigma}_t = \boldsymbol{I}$ and study two different settings: $\boldsymbol{\mu}_1 = (5, 0)$ (distant classes) and $\boldsymbol{\mu}_1 = (1, 0)$ (close classes) with min $Pe$ of 0.62% and 30.85%, respectively. The optimal solution is given by $\mathbf{w}^* = (w_1^*, 0, w_0^*)$ such that $-w_0^*/w_1^* = 2.5$ and $-w_0^*/w_1^* = 0.5$, respectively. So in theory, an infinite number of optimal solutions exists.

Using the Nelder-Mead minimization algorithm (Lagarias, Reeds, Wright, & Wright, 1998), we found for the distant classes case the solution $(w_1, w_2, w_0) = (4.75, 0.00, -11.87)$, corresponding to the vertical line $x_1 = 2.50$, the optimal solution. Figure 4a shows $H_S$ (represented by dot size) for a grid surrounding the found solution: the central dot with minimum size. The same algorithm was not able to find the optimal solution for the case of close classes. The reason is illustrated in Figures 4b and 4c using a grid around a candidate solution. We encounter a minimum for the $w_1$ and $w_2$ directions and a maximum in the $w_0$ direction. This means that for close classes when only rotations of the optimal line are allowed, the best solution is in fact the vertical line; if shifts are allowed, the previous

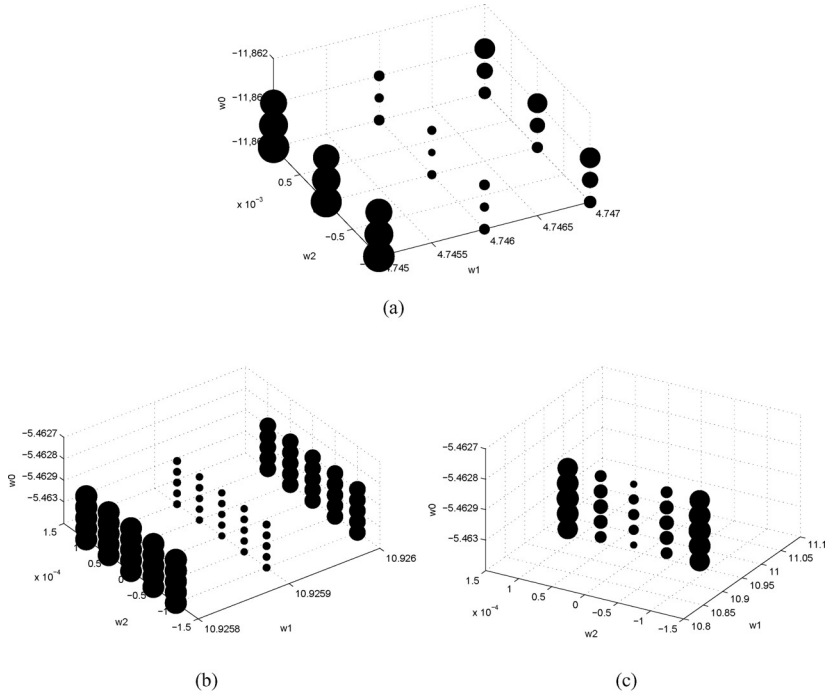
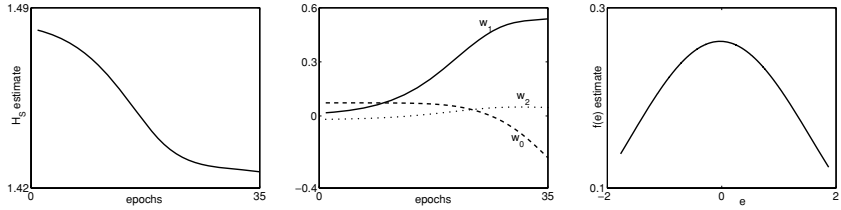



Figure 4: $H_S$ for bivariate gaussian class conditionals computed on a grid of $(w_1, w_2, w_0)$ values around the central dot $\bar{\mathbf{w}}$. The sizes of the dots represent values of $H_S$. (a) Corresponds to the $\boldsymbol{\mu}_1 = (5, 0)$ case with $\bar{\mathbf{w}} = \mathbf{w}^*$. (b, c) (The latter is a zoom of the vertical central layer). Corresponds to $\boldsymbol{\mu}_1 = (1, 0)$. In this case, $\bar{\mathbf{w}}$ is a minimizer in the $w_1$ and $w_2$ directions but a maximizer in the $w_0$ direction.

solution corresponds to an entropy maximum. One can understand this behavior in the following way. For shifts, the degree of disorder in the errors is decreased if we assign all the errors to one class (related to property 1 in section 3.3); by rotating the line, we increase the degree of disorder, which means that entropy will be minimum when the line is vertical (related to property 2 in section 3.3).

If empirical entropy is considered, both classifier problems can be solved with the MEE approach. A training set with 500 points per class was generated, and a test set with 5000 points per class was created in order to obtain an accurate estimate of min $Pe$. For the "distant classes" case we got a final solution (after 40 epochs of training with $h = 0.8$) $\mathbf{w} = (0.87, 0.01, -2.1)$ with a training error of 1% and a test error of 0.66%, very close to the min $Pe$ for this problem (see above). (Note that while with theoretical entropy, only a single solution was found, an infinite number of optimal solutions with the empirical entropy, corresponding to the appropriate $-w_0/w_1$ ratio, is




(a) (b) (c)

Figure 5: The first two plots, from left to right, show the empirical entropy and the weights of the perceptron across the epoch number for the close classes case. The right-most plot shows the estimated error pdf at the last epoch.

found.) The same happens for the "close classes" case, using, as expected, a larger value of $h$ ($h = 1$). The final solution is now $\mathbf{w} = (0.54, 0.05, -0.23)$ corresponding to a training error of 28.8% and a test error of 31.19%, also close to the min $Pe$ value. Figure 5 shows the convergence of a perceptron for the "close classes" case.

This example illustrates that the empirical MEE is able to work more easily in a wide range of configurations, provided the value of $h$ is properly chosen.

*3.5.2 Theoretical and Empirical MEE in Other Data Sets.* With the aim of illustrating how theoretical and empirical MEE behave with more realistic data sets, we now present two sets of experiments differing only in the family of functions implemented by the perceptron. We restrict ourselves to two-dimensional problems to allow graphical representations and have the search algorithms running in reasonable time for these heavy computation problems. With these concerns in mind, we consider classifier problems represented in the plane of the first two principal components $(x_1, x_2)$ of original real-world data sets. Since the true distributions are unknown, the true theoretical MEE solutions cannot in rigor be derived. However, we are still able to study theoretical MEE solutions for very closely resembling problems proceeding as follows. We first model the bivariate real-world data pdf's by appropriate distributions, achieving the same covariance metrics and with the minimum $L_1$ distance of the marginal pdf's. Next, we apply to these modeled pdf's the procedure outlined in section 3.2. For a fair comparison, the empirical MEE solutions for the same generated data sets were computed. min $Pe$ solutions (using the Nelder-Mead optimization algorithm) were also computed.

The following data sets were considered:

WDBC: This data set corresponds here to the first two principal components of the WDBC (Wisconsin Diagnostic Breast Cancer) data (Asuncion & Newman 2007), with 569 cases (212 from the malignant class and

357 from the benign class). The generated data contained 2390 cases, maintaining the original class proportions.

Wine: This data set corresponds here to the first two principal components of the Wine data (Asuncion & Newman, 2007), characterizing three different wine cultivars. There are 178 cases. The generated data contain 5000 cases, maintaining the original class proportions.

Thyroid: This data set corresponds here to the first two principal components of the New-Thyroid data (Asuncion & Newman, 2007), related to the state of the thyroid gland (normal, 150 cases; hyperthyroidism, 35 cases; hypothyroidism, 30 cases). The generated data contain 2509 cases, maintaining the original class proportions.

PB12: This data set is a speaker-independent, four-class, vowel discrimination problem (Jacobs, Jordan, Nowlan, & Hinton, 1991). The data consist of the first and second formants of the vowels *i*, *I*, *a*, and *A* from 75 speakers (men, women, and children). Vowels *i* and *I* form one overlapping pair of classes, and vowels *a* and *A* form the other pair. All the classes have 152 cases. The generated data contain 6000 cases, maintaining the original class proportions.

Ionosphere: This data set corresponds to the first two principal components of the IONOSPHERE data (Asuncion & Newman 2007), with 351 cases (225 from the "bad" class and 126 from the "good" class). The generated data contain 5778 cases, maintaining the original class proportions.

*Set 1: Family of Lines.* In this first set of experiments we consider four of the data sets: WDBC, Wine, Thyroid, and PB12. The perceptron implements a separating line $\varphi_{\mathbf{w}}(\mathbf{x}) = \tanh(w_1 x_1 + w_2 x_2 + w_0)$. Notice that Thyroid, Wine, and PB12 are multiclass problems. To overcome this difficulty, we applied the algorithms in a sequential strategy to two-class subproblems. For example, in Wine, we first discriminate the bottom class from the upper ones and then discriminate between the upper classes.

Table 1 shows the training error obtained with the algorithms. In general, both theoretical and empirical MEE obtain reasonable solutions, most of them very close to the min $Pe$ solution. However, we encounter a worse performance of theoretical MEE in the PB12 and Thyroid problems. It is worth noting that theoretical MEE was also more sensitive to the starting point needed by the optimization algorithm, converging many times to suboptimal or even poor solutions. Moreover, the final theoretical MEE solutions have higher weights when compared to empirical MEE or min $Pe$. The separating lines are shown in Figure 6. In most cases, the lines are almost coincident (although with different weights, the ratios are close to each other), except for the Thyroid and PB12 cases.

*Set 2: Family of Circles.* In this set of experiments, the perceptron implements a circle centered at $(w_1, w_2)$ with radius $w_0$, that is,

Table 1: Training Error (in Percentage) of Empirical MEE, Theoretical MEE, and min $Pe$ Algorithms on Four Data Sets.

| | Empirical MEE | Theoretical MEE | min $Pe$ |
|---|---|---|---|
| WDBC | 8.24 | 8.90 | 8.08 |
| Thyroid | 3.67 | 4.58 | 3.75 |
| Wine | 5.53 | 5.46 | 5.26 |
| PB12 | 10.72 | 14.10 | 10.67 |

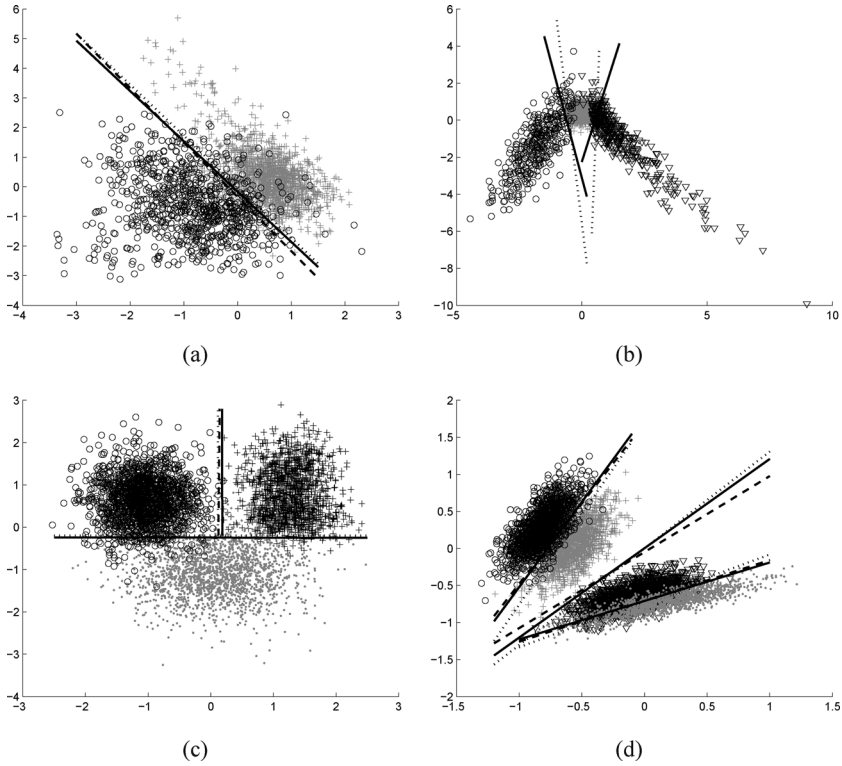



Figure 6: Separating lines obtained with empirical MEE (dashed), theoretical MEE (dotted), and min $Pe$ (solid) for WDBC (top left), Wine (bottom left), Thyroid (top right), and PB12 (bottom right) data sets.

$\varphi_{\mathbf{w}}(\mathbf{x}) = \tanh((x_1 - w_1)^2 + (x_2 - w_2)^2 - w_0^2)$ aiming to discriminate one class from the other ones. By using this richer class of functions, one is able to obtain with a single perceptron a reasonable data classification that would normally require a nontrivial MLP, with the added benefit that the theoretical

Table 2: Training Error (in Percentage) of Empirical MEE, Theoretical MEE, and min $Pe$ Algorithms on Three Data Sets.

| | Empirical MEE | Theoretical MEE | min $Pe$ |
|---|---|---|---|
| Wine | 4.15 | 5.12 | 4.02 |
| Thyroid | 4.94 | 21.52 | 3.59 |
| Ionosphere | 18.88 | - | 18.65 |

MEE solution search is again carried out in 3D space. We consider the Wine, Thyroid, and Ionosphere problems.

Table 2 shows the training error for these data sets. While empirical MEE solved all the problems (and almost optimally), theoretical MEE was not able to obtain reasonable solutions for the Thyroid and Ionosphere problems. In this last case, characterized by a large class overlap, the theoretical MEE solution corresponded to classify completely one of the classes at the cost of the other. It is also interesting to note in Wine (see Figure 7a) that theoretical MEE finds a circle that has a higher radius, but the center is positioned such as to catch only one of the classes.

*Discussion.* Several interesting aspects were observed in the experiments. First, empirical MEE always finds a good solution for the problem at hand, while theoretical MEE may encounter difficulties, as illustrated with the family of circles. We also verified with set 1 of the experiments that theoretical MEE usually converges to a solution with high values for the weights. In fact, if small weights (but with the appropriate ratio to ensure the optimal line) were used, theoretical MEE often diverged. This is explained as follows: $w_1x_1 + w_2x_2 + w_0$ is a projection of $(x_1, x_2)$ onto unidimensional space; when higher weights are used, points on opposite sides of the separating line are projected far away; this implies Dirac-like error pdf's and thus, by formula 2.2, an entropy minimum. For the family of circles in set 2, similar behavior is not possible because "higher weights" cannot be found maintaining the same center and radius of the circle. The only possibility is to increase the radius of the circle (controlling the center), as illustrated with the Wine data set.

We also notice that the simplicity of the classifiers (as those we have been using) may have a negative impact on the theoretical MEE performance. In fact, for quite general classification problems, their optimal solutions for low-complexity-function classes $\Phi$ may produce multimodal error pdf's, implying a decreased ability of theoretical MEE to solve such problems (e.g., many local suboptimal solutions may appear). When more complex classifiers (e.g., MLPs) are used, the effect of multimodality is decreased, which means that theoretical MEE may perform better and similar to empirical MEE. However, for the same problem (and the ones used in the experiments serve as examples), we may obtain a solution with a multimodal

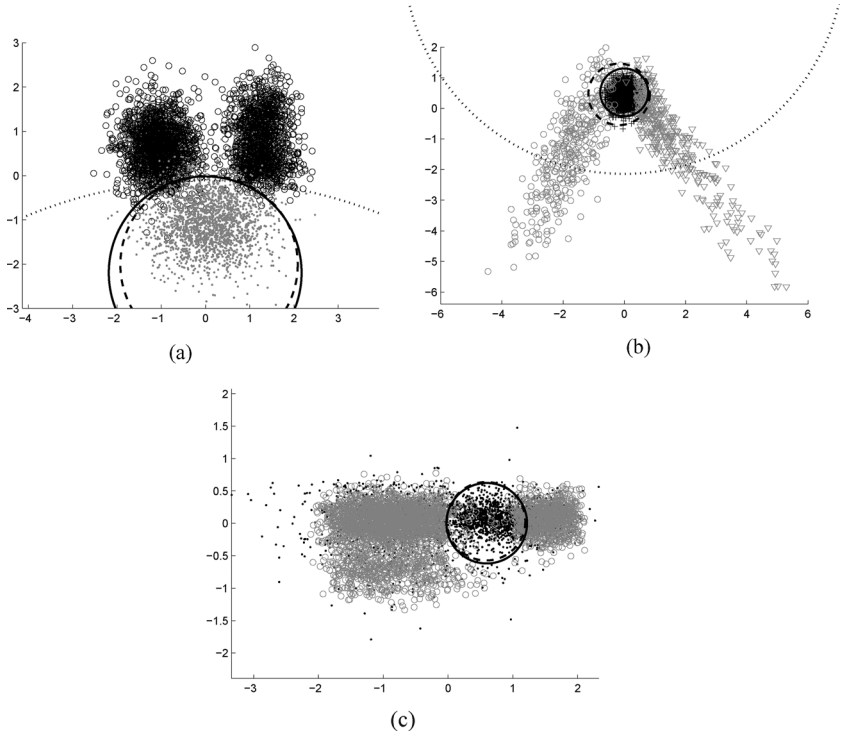



Figure 7: Separating circles obtained with empirical MEE (dashed), theoretical MEE (dotted), and min $Pe$ (solid) for the Wine (top left), Thyroid (top right), and Ionosphere (bottom) data sets.

error pdf or an equivalent solution with higher weights with unimodal error pdf (see the discussion above). The main problem for the theoretical MEE is that reaching this latter solution may be difficult due to the decoupled role of the $H_{S|t}$ terms in formula 2.2, already commented on in section 2.2: there are many ways to reach an entropy minimum, namely, if all the errors can be assigned to one of the classes (property 1). In the next section, we study the influence of the kernel smoothing in providing the needed coupling of the $f_{Y|t}$ pdf's and explaining the good behavior of empirical MEE.

## 4 Influence of the Smoothing Parameter on the Empirical MEE

We mentioned in section 3.2 the essential differences between theoretical and empirical MEE. We now discuss why the empirical MEE usually works (we have not yet encountered a single problem, after processing many and

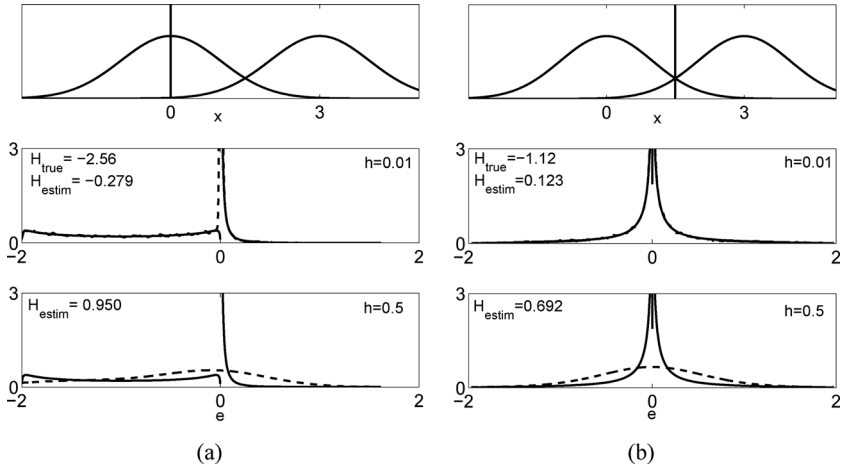



Figure 8: The kde smoothing effect. The top figures show the class-conditional pdf's with the split location (solid vertical line). The middle and bottom figures show the theoretical (solid line) and kde (dashed line) error pdf's for the corresponding split for two different values of $h$.

diversified sets of problems, where it did not work; see Silva et al., 2005, 2008; Santos, 2007) that is, the estimated error entropy (be it Shannon or Rényi) will reach a minimum. Moreover, the corresponding solutions often outperform the solutions obtained by other sophisticated methods.

Consider the split-type setting for the gaussian case (see section 3.4). Figure 8 illustrates the influence of kde in determining the error distribution. It shows the theoretical and empirical error pdf's for two split locations: off-optimal (left figures) and optimal (right figures). Note the smoothing imposed by the kde: an increased value of $h$ implies an oversmoothed estimate with greater impact near the origin. If we look to the bottom figures, we can understand the theoretical entropy maximum previously mentioned and the changes operated after the kde smoothing. In the left figure, the true error pdf for class $\mathcal{C}_{-1}$ is nearly uniform, which implies a high value for $H_{S|-1}$. However, the error pdf for $\mathcal{C}_1$ is highly concentrated at the origin, producing a quite low $H_{S|1}$. Due to property 3 and formula 2.2, the overall value of $H_S$ will be lower than the one in the right figure, where the overall true error pdf is more concentrated around zero. This is why theoretical entropy in this case has a maximum at the optimal split. When density estimation is used with sufficiently high values for $h$, these behaviors are smoothed out (the error pdf is then seen as a "whole," ignoring relation 2.2); now the nonoptimal split estimated pdf has a long left tail, whereas the optimal one is more concentrated around the origin, yielding a minimum. This example, illustrating property 3 in section 3.3, shows that for appropriate (high) values of $h$, the smoothing effect of the kde is

responsible for producing a minimum of entropy. A similar behavior can be found for the perceptron with one weight plus bias and gaussian inputs (Silva, 2008), which means that the maximum-to-minimum flip observed is quite general.

Let us study the theoretical behavior of kernel smoothing on two distinct pdf's assumed as error pdf's. One of them, say $f_1(x)$, corresponds to an off-optimal point pdf characterized by a large tail of errors for one class and a fast-decaying function of the errors for the other class, modeled as

$$f_1 = \frac{1}{2}u(-1, 0) + \frac{1}{2}e_+(p), \tag{4.1}$$

where $e_+(p)$ is the exponential pdf with parameter $p$, decaying for $x \geq 0$. The second one, $f_2(x)$, corresponds to a decision border set at the optimal point, implying a decaying error pdf for both classes, modeled as

$$f_2 = \frac{1}{2}e_+(a) + \frac{1}{2}e_-(a), \tag{4.2}$$

where $e_-(a)$ is the exponential pdf with parameter $a$, decaying for $x \leq 0$. Simple calculations of the Rényi entropy (for Shannon entropy, the problem becomes quickly untractable) show that the respective entropies $H_1$ and $H_2$ are such that $H_1 < H_2$ (i.e., a maximum as in Figure 8) if $p > 2(a - 1)$. We now proceed to convolve these pdf's with a gaussian kernel $K_h$ with bandwidth $h$. The resulting pdf's, after some mathematical manipulation, are

$$(g * f_1)(x) = \frac{1}{2}\left[\Phi\left(\frac{x+1}{h}\right) - \Phi\left(\frac{x}{h}\right)\right] + \frac{p}{2}e^{\frac{p^2h^2}{2} - px}\Phi\left(\frac{x - ph^2}{h}\right) \tag{4.3}$$

$$(g * f_2)(x) = \frac{a}{2}e^{\frac{a^2h^2}{2}}\left[e^{ax}\left[1 - \Phi\left(\frac{x + ah^2}{h}\right)\right] + e^{-ax}\left[1 - \Phi\left(\frac{x - ah^2}{h}\right)\right]\right], \tag{4.4}$$

where $\Phi$ denotes here the probability cumulative function of the standardized gaussian distribution. Using these formulas and setting $p = 2(a - 1) + 1$ in order to have an entropy maximum for the original $f_2$ pdf, one may always find a sufficiently large $h$ such that the produced smearing out of the $f_1$ tail will turn $f_2$ entropy into a minimum. This is exemplified in Figure 9 for $a = 2.8$. This behavior is quite general, in the sense that even for arbitrarily large MLPs, one always gets the error pdf behavior shown in Figure 9 or Figure 8. Also for the multiclass case, we observe the same behavior of the error entropy for the distinct MLP's outputs, justifying the MEE efficacy in this scenario too (see Santos, 2007).

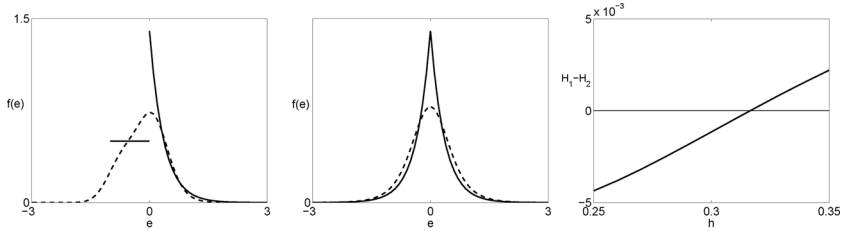



Figure 9: Error pdf models (solid line) for the off optimal (top left) and optimal decision border (top right). The convoluted pdf's are plotted with a broken line using $h = 0.316$ corresponding to the entropy maximum to minimum cross-over point. The bottom figure plots the difference between the entropies of the convoluted pdf's $f_1$ and $f_2$ (as given in equations 4.3 and 4.4) as a function of $h$. Note the zero crossing at $h = 0.316$.

Finally, we show that the change of error entropy behavior as a function of $h$ can also be understood directly from its estimated formulas. Consider, for simplicity, Rényi's expression. The minimization of equation 3.9 is equivalent to the maximization of

$$\hat{V}_{R_2} = \exp(-\hat{H}_{R_2}) = \frac{1}{n^2 h} \sum_{i=1}^{n} \sum_{j=1}^{n} G\left(\frac{e_i - e_j}{h}\right), \tag{4.5}$$

where $G$ is the gaussian kernel. Let $G_{ij} = G(\frac{e_i - e_j}{h})$, $c = \frac{1}{n^2 h}$, and $c_t = \frac{1}{n_t^2 h}$ for $t \in \{-1, 1\}$, where $n_t$ is the number of samples from class $\mathcal{C}_t$. Then, as $G$ is symmetrical about the origin, we may write

$$\begin{aligned}\hat{V}_{R_2} &= \left(\frac{n_{-1}}{n}\right)^2 c_{-1} \sum_{i \in \mathcal{C}_{-1}} \sum_{j \in \mathcal{C}_{-1}} G_{ij} + \left(\frac{n_1}{n}\right)^2 c_1 \sum_{i \in \mathcal{C}_1} \sum_{j \in \mathcal{C}_1} G_{ij} + 2c \sum_{i \in \mathcal{C}_1} \sum_{j \in \mathcal{C}_{-1}} G_{ij} \\ &= \hat{q}^2 \hat{V}_{R_2|-1} + \hat{p}^2 \hat{V}_{R_2|1} + 2c \sum_{i \in \mathcal{C}_1} \sum_{j \in \mathcal{C}_{-1}} G_{ij}. \end{aligned} \tag{4.6}$$

Entropy is therefore decomposed as a weighted sum of positive quantities exclusively related to each class (just as in the theoretic derivation; see appendix A), plus a term that exclusively relates the cross-errors. Let $\hat{V}_{R_2^*}$ be the estimator 4.6 without the cross-errors term. In Figure 10 we compare the behavior of $V_{R_2}$, $\hat{V}_{R_2}$, and $\hat{V}_{R_2^*}$ as a function of the split parameter $w_0$ for the same problem as in Figure 8. First, we notice that instead of a maximum of $V_{R_2}$, we encounter a minimum (in the same sense that in section 3.4, we encountered a maximum of entropy instead of a minimum). In Figure 10a, we see that if $h \to 0$, both $\hat{V}_{R_2}$ and $\hat{V}_{R_2^*}$ will converge as expected to $V_{R_2}$. If $h$ is

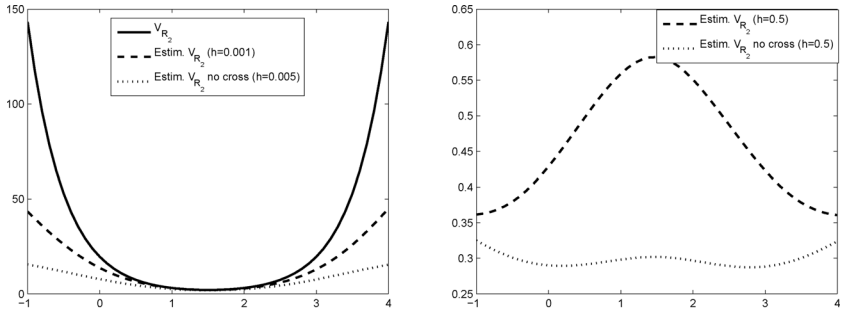



Figure 10: (Left) $V_{R_2}$ (solid), $\hat{V}_{R_2}$ (dashed), and $V_{R_2^*}$ (dotted) plotted for a split-type problem as a function of the split $w_0$ with small values of $h$ on the left and higher values of $h$ on the right.

increased, both $\hat{V}_{R_2}$ and $\hat{V}_{R_2^*}$ will exhibit a maximum at the optimal solution ($w_0 = 1.5$), but with an important difference: while the maximum of $\hat{V}_{R_2^*}$ is not global (for any $h$), the maximum of $\hat{V}_{R_2}$ turns out to be global for a sufficiently large $h$. Thus, to maximize $\hat{V}_{R_2}$, it is important not only to maximize $\hat{q}^2 \hat{V}_{R_2|-1} + \hat{p}^2 \hat{V}_{R_2|1}$ as for the theoretical counterpart (but this can be achieved with very different pdf configurations with consequences to the classifier's performance; see property 7) but also to maximize $2c \sum_{i \in \mathcal{C}_1} \sum_{j \in \mathcal{C}_{-1}} G_{ij}$ (not available in the theoretical framework), which is achieved if the errors are concentrated around the origin. This cross-error term is due to the kde estimator. Also, the value of $h$ needed to produce the flip is necessarily related to the amount of overlap between the classes, in the sense that a higher $h$ is needed for higher overlap.

## 5 Conclusion

All risk functionals (r.f.) used in nonparametric data classification are essentially concentration measures of the error r.v. This topic, somewhat neglected in the literature, was discussed in detail for the MSE, CE, and entropy of error; the corresponding r.f. expressed as functionals of the error r.v. were presented and their properties analyzed. We have particularly discussed and illustrated with examples which properties may work advantageously in the application of the MEE approach. One illustration was a class of classifying problems where MSE and CE totally fail, whereas MEE solves the classifier problem by capitalizing on the up-saturating nature of the $H(\cdot)$ curve, rendering entropy "immune" to long tenuous tails. We analyzed the simple perceptron at work with the MEE approach in both its theoretical and practical (empirical) implementations. The comparison of theoretical and empirical implementations of r.f. may help to detect and

elucidate relevant practical aspects of why r.f. work in practice when they do. This aspect has also been somewhat neglected. For instance, although a lot is known about theoretical MSE, the same cannot be said about CE. In what concerns theoretical MEE, we have shown the essential role of the squashing activation functions and shaping parameters in driving the entropy of error toward a minimum whenever the degree of overlap of the classes is moderate (say, means apart by more than one standard deviation). Moreover, based on theoretical derivations for simple two-class problems, we were able to prove the ability of the theoretical MEE to often reach or come close to the minimum probability of error. We also discussed that the multimodality of the error pdf (due to the simplicity of the classifiers analyzed) may deteriorate the theoretical MEE performance. On the other hand, empirical MEE is immune to multimodality provided an adequate kernel smoothing is employed. In fact, we were able to elucidate why it works in a much larger class of problems, failing only when the classes are so overlapped that the optimum probability of error is close to 0.5 (for two-class data sets with equal priors). We showed the transition from an entropy maximum to an entropy minimum with the increase of the kernel bandwidth and analyzed a hypothetical error pdf model proving the maximum-to-minimum transition. This was also evidenced by direct manipulation of entropy formulas. The presence of a cross-error term revealed itself as an important influencer of the maximum-to-minimum transition.

Section 3.2 presented the theoretical and empirical MEE at work in two-dimensional real-world data sets, with varied distributions and number of classes. The results confirm the preceding findings: the good behavior of the empirical MEE, providing good solutions close to min $Pe$ solutions, as well as of the theoretical MEE solutions when these can be found.

The letter has thus provided theoretical and experimental evidence justifying and clarifying the use of the empirical MEE principle in data classification with perceptron-based machines. Still, the difficulty of evaluating and analyzing the MEE criterion for more general classifier problems paves the way for continuing and future work. Among the interesting aspects to study are the relation between the entropy estimator and some robust risk measure as discussed in Liu, Pokharel, and Príncipe (2007); the behavior of the MEE criterion when using different costs for the false negative and false positive errors; the implementation to decision trees (which is already being pursued) and support vector machines; and implementation of the training of recurrent networks as a way to increase their classification and forecasting capabilities.

## Appendix A: Entropy and Variance of Partitioned pdf's

Consider a pdf $f(x)$ defined by a weighted sum of functions with disjoint supports,

$$f(x) = \sum_i a_i f_i(x),$$

such that

1. Each $f_i(x)$ is a pdf with support $D_i$.
2. $D_i \cap D_j = \varnothing, \forall i \neq j$.
3. $f$ has support $D = \cup_i D_i$.
4. $\sum_i a_i = 1$.

We call such an $f(x)$ a *partitioned pdf*. In this case, the Shannon entropy assumes a particular form given by

$$H_S(f) = \sum_i a_i H_S(f_i) - \sum_i a_i \ln(a_i),$$

that is, the entropy of $f$ is a weighted sum of the entropies of each component plus the entropy of the weighting factors. One can also derive a similar formula for the variance, $V_f$, of a partitioned pdf $f$ as

$$V_f = \sum_i a_i V_{f_i} + \sum_i a_i (\mu_i - \mu)^2,$$

where $\mu$ ($\mu_i$) is the expected value associated with $f$ ($f_i$). Rényi's quadratic entropy of a partitioned pdf comes as

$$H_{R_2}(f) = -\ln\left[\sum_i a_i^2 \int_{D_i} f_i^2\right],$$

that is, $H_{R_2}(f)$ is not decomposable as Shannon's counterpart. Nevertheless, as the minimization of $H_{R_2}$ is equivalent to the maximization of $V_{R_2} = \exp(-H_{R_2})$, we may write

$$V_{R_2} = \sum_i a_i^2 \int_{D_i} f_i^2$$

which is now a weighted sum of positive quantities, each exclusively related to one support.

## Appendix B: Entropy Dependence on the Standard Deviation

**Definition.** *A real function $f(x)$ with support $\mathbb{R}^+$ is an up-saturating function (down-saturating) if it is strictly concave (convex) and increasing (decreasing).*

**Proposition 1.** *If $f(x)$ is a saturating function and $g(x) \geq 0$ is an up-saturating function, with the domain of $g$ contained in the support of $f$, then $f(g(x))$ is a saturating function. (Therefore, a saturating function of $\sigma^2$ is a saturating function of $\sigma$.)*

**Remark.** $\sqrt{x}$ and $\ln(x)$ are up-saturating functions. The digamma, $\psi(x) = \frac{d\Gamma(x)}{dx}$, and trigamma function, $\psi_1(x) = \frac{d\psi(x)}{dx}$, are up and down-saturating functions, respectively.

**Proposition 2.** *An up-saturating (down-saturating) function, $f(x)$, has a strictly decreasing (increasing) derivative.*

We analyze the up-saturating case. Since $f(x)$ is concave, we have, for $h > 0$, $(1-h)f(x) + hf(x+h) < f((1-h)x + h(x+h)) = f(x+h^2)$; therefore, $\frac{f(x+h)-f(x)}{h} < \frac{f(x+h^2)-f(x)}{h^2}$ and since $f(x)$ is a strictly increasing function, the strictly decreasing derivative follows.

**Proposition 3.** *For a large class of continuous distributions (with "large" detailed in the following), the Shannon entropy is an up-saturating function of the standard deviation ($\sigma$).*

We first consider the densities of the Pearson system of distributions (Johnson and Kotz, 1970):

$$f(x) = C(a + bx + cx^2)^{-1/(2c)} \exp\left( \frac{(b-2cm)\tan^{-1}\left(\frac{b+2cx}{\sqrt{4ac-b^2}}\right)}{c\sqrt{4ac-b^2}} \right). \tag{B.1}$$

Provided $m$ is not a root of the denominator, $f$ is finite and $f'$ is 0 when $x = m$. The slope $f'$ is also 0 at $f = 0$. The conditions $\int f = 1$ and $f \geq 0$ imply that $f$ and $f'$ must tend to 0 as $x$ tends to infinity, restricting the range of $x$ values if necessary. The parameters $a$, $b$, and $c$ control the shape of $f$. Let $a_1$ and $a_2$ be the solutions of the denominator. The possible types of curves with explicit entropy formula are:

(a) $b = c = 0,\ a > 0$. This corresponds to the normal distribution with mean $-m$ and standard deviation $\sigma = \sqrt{a}$. The entropy $H(\sigma) = \ln(\sigma\sqrt{2\pi e})$ is an up-saturating function of $\sigma$.

(b) $b^2 - 4ac < 0$ (real roots), $a_1 < 0 < a_2$. This corresponds to the generalized beta family of distributions (Pearson types I and II), usually written as $f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$, $\alpha, \beta > 0$, with $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ and $H(\alpha,\beta) = \ln(B(\alpha,\beta)) - (\alpha-1)\psi(\beta) + (\alpha+\beta-2)\psi(\alpha+\beta)$. Beta densities can be symmetric, asymmetric, convex, or concave. Under certain conditions on $(\alpha, \beta)$, the entropy is not an up-saturating function. Numerical computation shows that it is enough that one of the parameters is smaller than one-half of the other for the result to hold.

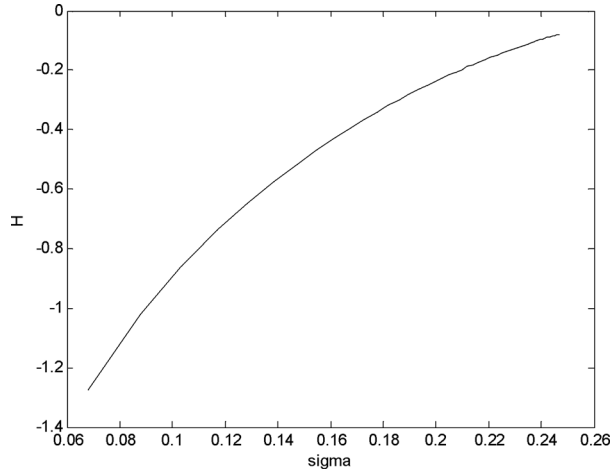



Figure 11: $H(\sigma)$ for the pdf, equation 3.12, with $a = 0.5$ and $b \in [0.015, 0.5]$.

(c) $c = 0$ with $b \neq 0$. This corresponds to the Pearson type III family, the gamma family $f(x) = x^{k-1}e^{-x/\theta}/(\theta^k \Gamma(k))$ for $x > 0$ and $k$ (shape), $\theta$ (scale) $> 0$, often used as pdf model since it is easily adjusted to many types of distributions. The entropy of this family is an up-saturating function of $\sigma$.

(d) $c = 4ab$. This corresponds to the Pearson type V family, the inverse gamma family, $f(x) = \beta^\alpha \frac{1}{x^{\alpha+1}} \exp(-\beta/x)/\Gamma(\alpha)$, where $\alpha$ is the shape parameter and $\beta$ the scale parameter. In order for $f(x)$ to have a variance, $\alpha$ must be larger than 2. Although no theoretical justification is available, the entropy is an up-saturating function of $\sigma$.

(e) $b = 0$ with $a, c > 0$. This corresponds to the Pearson type VII family, $f(x) = K(a + cx^2)^{-1/(2c)}$. There is no explicit formula for the entropy. However, an important subfamily is the Student's-$t$ family with $k$ degrees of freedom, $f(x) = K(a + x^2/k)^{-(k+1)/2}$, where $K$ is the normalizing factor. For $k > 2$, entropy is an up-saturating function of $\sigma$.

Other distribution families with up-saturating densities (easily checked) are uniform, triangular, Rayleigh, Laplace, lognorm, and Weibull.

Of particular interest is the pdf, equation 3.12, corresponding to the MLP error with gaussian inputs. The gaussianity of the inputs is not a stringent condition if the input distributions can be assumed to be independent. No entropy formula is available for this pdf, which we rewrite as $f(x) = K \exp(-(\operatorname{arctanh}(x) - a)^2/b)/((1 - x)(1 + x))$, with $a$ governing the mean and $b$ the variance. By numerical computation one can check that the entropy is indeed an up-saturating function of $\sigma$ (see Figure 11).

## Appendix C: An Example Where MSE and CE Fail

We present an example of a data classification problem where MEE provides the correct solution and MSE and CE do not. Let us consider two-class classification problems in bivariate space $\mathbb{R}^2$, target space $T = \{-1, 1\}$. We denote the input vectors by $[x_1\ x_2]^{\mathrm{T}}$ and consider the following marginal and independent pdf's:

$$f_1(x_1) = \frac{1}{2}[U(a, 1) + U(b, a)],\ f_{-1}(x_1) = f_1(-x_1);\qquad f_t(x_2) = U\left(-\frac{c}{2}, \frac{c}{2}\right), \tag{C.1}$$

where $U(a, b)$ is the uniform distribution in $[a, b]$ and $a, c > 0$.

Assuming $P(1) = P(-1) = 1/2$, the classification problem consists of selecting the straight line passing through the origin ($x_2 = \tan(\alpha)x_1$), yielding min $Pe$. Theoretical MEE ($H$) and MSE ($V$) values are easily computed for two configurations:

- Configuration 1 (the min $Pe$ solution) with $\alpha = -\pi/2$, $\mathbf{w} = [0\ 1]^T$:

$$H = \frac{1}{2}\ln(1 - a) + \frac{1}{2}\ln(a - b) - \ln\frac{1}{4} \tag{C.2}$$

$$V = \frac{(1 - a)^2}{6} + \frac{(a - b)^2}{24} + \frac{(2 - a - b)^2}{8} \tag{C.3}$$

- Configuration 2 with $\alpha = 0$, $\mathbf{w} = [1\ 0]^{\mathrm{T}}$:

$$H = \ln c - \ln\frac{1}{2} \tag{C.4}$$

$$V = \frac{c^2}{12} + 1 \tag{C.5}$$

In this problem, MEE always picks the correct solution ($\alpha = -\pi/2$), but for some values of the parameters, MSE and CE do not. For instance, if $a = 0.95$, $b = 1.7$, and $c = 0.9$, we obtain the $H$, $V$ curves in Figure 12a with min $Pe = 0.321$ and MSE selecting $\alpha = 0.377$ with $Pe = 0.355$. CE will also make the wrong decision for longer tails of configuration 1. For instance, with $a = 0.95$, $b = 2.4$, and $c = 0.9$, the cross-entropy obtained is shown in Figure 12b, selecting $\alpha = 0.346$ with $Pe = 0.411$, whereas min $Pe = 0.362$. Even if instead of the theoretical MEE, MSE, and CE, we use empirical estimates (we performed experiments with 500 points per class), the same conclusions are obtained.

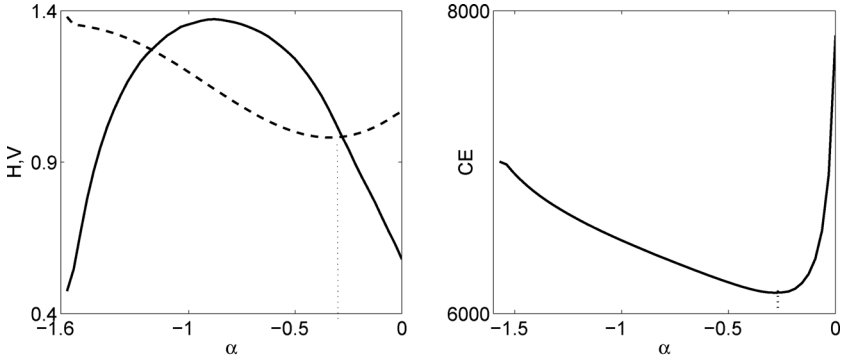



Figure 12: (Left) The $H$ (solid line), $V$ (broken line) curves for $a = 0.95$, $b = 1.7$, and $c = 0.9$. (Right) The $CE$ curve for $a = 0.95$, $b = 2.4$, and $c = 0.9$. The curves were obtained by numerical simulation with 8000 points.

## References

Alexandre, L. A., & Marques de Sá, J. (2006). Error entropy minimization for LSTM training. In *Proceedings of the 16th International Conference on Artificial Neural Networks* (pp. 244–253). Berlin: Springer.

Asuncion, A., & Newman, D. (2007). *UCI machine learning repository*. Irvine University of California, Irvine, School of Information and Computer Sciences. Available online at http://www.ics.uci.edu/∼mlearn/MLRepository.html.

Bishop, C. (1995). *Neural networks for pattern recognition*. New York: Oxford University Press.

Erdogmus, D., Hild II, K., & Príncipe, J. C. (2002). Blind source separation using Rényi's $\alpha$-marginal entropies. *Neurocomputing*, *49*, 25–38.

Erdogmus, D., & Príncipe, J. C. (2000). 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* (pp. 75–80). Berlin: Springer-Verlag.

Erdogmus, D., & Príncipe, J. C. (2002). An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems. *IEEE Transactions on Signal Processing*, *50*(7), 1780–1786.

Erdogmus, D., & Príncipe, J. C. (2003). Convergence properties and data efficiency of the minimum error entropy criterion in adaline training. *IEEE Transactions on Signal Processing*, *51*(7), 1966–1978.

Gokcay, E., & Príncipe, J. C. (2002). Information theoretic clustering. *IEEE Trans. on Pattern Analysis and Machine Learning*, *24*(2), 158–171.

Han, L. (2007). *Kernel partial least squares (K-PLS) for scientific data mining*. Unpublished doctoral dissertation, Rensselaer Polytechnic Institute.

Hild II, K., Erdogmus, D., & Príncipe, J. C. (2001). Blind source separation using Rényi's mutual information. *IEEE Signal Processing Letters*, *8*, 174–176.

Jacobs, R., Jordan, M., Nowlan, S., & Hinton, G. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*, 79–87.

Jenssen, R., Hild, K., Erdogmus, D., Príncipe, J.-C., & Eltoft, T. (2003). Clustering using Rényis entropy. In *Proceedings of the Int. Joint Conference on Neural Networks* (pp. 523–528). Piscataway, NJ: IEEE Press.

Johnson, N., & Kotz, S. (1970). *Continuous univariate distributions*. Hoboken, NJ: Wiley.

Kapur, J. (1993). *Maximum-entropy models in science and engineering* (rev. ed.). Hoboken, NJ: Wiley.

Lagarias, J., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM Journal of Optimization*, *9*(1), 112–147.

Linsker, R. (1988). Self-organization in a perceptual network. *IEEE Computer*, *21*, 105–117.

Liu, W., Pokharel, P., & Príncipe, J. (2007). Correntropy: Properties and applications in non-gaussian signal processing. *IEEE Transactions on Signal Processing*, *55*(11), 5286–5298.

Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, *33*, 1065–1076.

Príncipe, J. C., Xu, D., & Fisher, J. (2000). Information theoretic learning. In S. Haykin (Ed.), *Unsupervised adaptive filtering, Vol. 1: Blind source separation* (pp. 265–319). Hoboken, NJ: Wiley.

Rényi, A. (1970). *Probability theory*. Dordrecht: North-Holland.

Richard, M., & Lippmann, R. (1991). Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Computation*, *3*, 461–483.

Santos, J. (2007). *Data classification with neural networks and entropic criteria*. Unpublished doctoral dissertation, University of Porto.

Santos, J. M., Marques de Sá, J., & Alexandre, L. A. (2005). Batch-sequential algorithm for neural networks trained with entropic criteria. In *Proceedings of the International Conference on Artificial Neural Networks* (pp. 91–96). Berlin: Springer-Verlag.

Santos, J. M., Marques de Sá, J., & Alexandre, L. A. (2007). Legclust—a clustering algorithm based on layered entropic subgraphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(1), 62–75.

Santos, J. M., Marques de Sá, J., Alexandre, L. A., & Sereno, F. (2004). Optimization of the error entropy minimization algorithm for neural network classification. In C. Dagli, A. Buczak, D. Enke, M. Embrechts, & O. Ersoy (Eds.), *Intelligent engineering systems through artificial neural networks* (pp. 81–86). Washington, DC: ASME.

Silva, L. M. (2008). *Neural networks with error-density risk functionals for data classification*. Unpublished doctoral dissertation, University of Porto.

Silva, L. M., Embrechts, M., Santos, J. M., and Marques de Sá, J. (2008). The influence of the risk functional in data classification with MLPs. In *Proceedings of the International Conference on Artificial Neural Networks* (pp. 185–194). Berlin: Springer-Verlag.

Silva, L. M., Felgueiras, C., Alexandre, L. A., & Marques de Sá, J. (2006). Error entropy in classification problems: A univariate data analysis. *Neural Computation*, *18*(9), 2036–2061.

Silva, L. M., Marques de Sá, J., & Alexandre, L. A. (2005). Neural network classification using Shannon's entropy. In *Proceedings of the European Symposium on Artificial Neural Networks* (pp. 217–222). Bruges: d-side.

Stoller, D. (1954). Univariate two-population distribution free discrimination. *Journal of the American Statistical Association*, *49*, 770–777.

Thompson, J., & Tapia, R. (1978). *Nonparametric probability density estimation*. Baltimore, MD: Johns Hopkins University Press.

Watanabe, S. (1981). Pattern recognition as a quest for minimum entropy. *Pattern Recognition*, *13*(5), 381–387.

Received July 27, 2009; accepted March 14, 2010.