Equivalence Proofs for Multi-Layer Perceptron Classifiers and the Bayesian Discriminant Function*

John B. Hampshire II
Dept. of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Barak A. Pearlmutter†
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Abstract

This paper presents a number of proofs that equate the outputs of a Multi-Layer Perceptron (MLP) classifier and the optimal Bayesian discriminant function for asymptotically large sets of statistically independent training samples. Two broad classes of objective functions are shown to yield Bayesian discriminant performance. The first class are "reasonable error measures," which achieve Bayesian discriminant performance by engendering classifier outputs that asymptotically equate to *a posteriori* probabilities. This class includes the mean-squared error (MSE) objective function as well as a number of information theoretic objective functions. The second class are classification figures of merit (CFM_{mono}), which yield a qualified approximation to Bayesian discriminant performance by engendering classifier outputs that asymptotically identify the maximum *a posteriori* probability for a given input. Conditions and relationships for Bayesian discriminant functional equivalence are given for both classes of objective functions. Differences between the two classes are then discussed very briefly in the context of how they might affect MLP classifier generalization, given relatively small training sets.

1 INTRODUCTION

The use of multi-layer perceptron (MLP) classifiers in statistical pattern recognition requires that there be some mathematically defensible link between MLP outputs and the true *a posteriori* probabilities associated with the input random vector (RV) $\mathbf{x}$ being classified. We present a number of proofs that detail the link for an N-output MLP classifier and the N-class RV $\mathbf{x}$, possessing an input feature space dimensionality of M. The number of classes N and the feature space dimensionality M of $\mathbf{x}$ are arbitrary, as is the *specific* parameterization (or connectivity) of the MLP classifier. For our purposes the term "multi-layer perceptron" is used to describe a backpropagation network using any continuous sigmoidal nonlinearity, although the proofs herein can be extended to networks employing other non-linearities.

Proofs of the relationship between both linear and non-linear classifiers trained with the mean-squared-error (MSE) objective function and the Bayesian discriminant function are not new. Duda and Hart formulated the proof for a simple perceptron in [6] (pp. 154-155). More recently, [1, 3, 7, 11] have given variations of the proof for MSE-trained MLPs. We extend these proofs to the N-output MLP classifier trained with *any* objective function belonging to one of two broad classes. The proofs herein give detailed relationships among the MLP outputs, the Bayesian discriminant function, and the class conditional densities of $\mathbf{x}$. In this sense, they have their conceptual basis in the proof of [6].

We show that the MSE proofs of [1, 3, 6, 7, 11] pertain to one specific member of a broad class of error measure objective functions. This class of "reasonable" error measures yields MLP outputs that converge to the Bayesian *a posteriori* probabilities $P(\omega_i \mid \mathbf{x})$ (where ω_i represents the ith class) for networks with sufficient functional capacity (see section 3.1.2) to classify asymptotically large sets of statistically independent training samples. The MSE and Cross Entropy (CE) [10] objective functions are members of this class of functions[1], as are other objective functions stemming from information theoretic learning rules (such as Maximum Mutual Information and Maximum Likelihood), and the Kullback-Liebler distance measure. These reasonable error measures all yield optimal Bayesian discriminant performance[2], given sufficient training data.

* Appearing in *Proceedings of the 1990 Connectionist Models Summer School*, Touretzky, Elman, Sejnowski, and Hinton, eds., San Mateo, CA: Morgan Kaufmann, 1990.

† Hertz Fellow

[1] Strictly speaking, the Cross Entropy objective function does not require that MLP outputs be compared with *binary* target values. Thus, it is fair to categorize the Cross Entropy objective function in this way *only* when binary target values are specified in its form.

[2] See section 2.

Given these results, one is inclined to conclude that all these objective functions yield equivalent classification performance, and that all MLPs are — in effect — no more than exotic estimators of Bayesian *a posteriori* probabilities. In fact, neither conclusion is correct. A broad class of objective functions called "N-monotonic Classification Figures of Merit" (CFM_{mono}) [8] are shown to approximate Bayesian classification performance under the same conditions for which the reasonable error measures yield Bayesian performance. However the CFM_{mono} class of functions *does not* produce MLP output activations that reflect *a posteriori* probabilities $P(\omega_i \mid \mathbf{x})$; instead it asymptotically identifies the maximum *a posteriori* probability for a given input $P(\omega_{max} \mid \mathbf{x}_p)$, as long as $P(\omega_{max} \mid \mathbf{x}_p) > 0.5$ (see section 4). Despite this limitation, [8] indicates that CFM_{mono}-trained MLPs can be more robust approximations to the Bayesian discriminant than their reasonable error measure counterparts, given small training sample sizes.

While the findings of [8] are not broad enough to be considered conclusive, they do argue against the maxim "all objective functions yield equivalent classification performance," when one's training set is limited in size. Section 5 contains some brief comments regarding the following proofs' applicability to real-world classification problems. Particular attention is paid to how the different objective functions might yield (or fail to yield) near-optimal classification boundaries for small training sets. These observations are made with an eye towards further investigation of MLP classifier generalization in the probabilistic context presented by this paper.

By the "asymptotic behavior" of a classifier we mean its behavior for an asymptotically large set of statistically independent training samples.

2 A GENERAL DESCRIPTION OF THE N-CLASS PROBLEM AND THE BAYESIAN DISCRIMINANT FUNCTION

In this section we give a brief description of the general *N*-class problem and the Bayesian discriminant function in the context of the connectionist and pattern recognition literature. The syntax and notation used herein is an expanded version of that used in [6].

The *N*-class classification problem is depicted in Figure 1. A random vector $\mathbf{x}$ is to be classified by a classifier with parameterization specified by the state variable θ. The classifier has *N* outputs, each one of which corresponds to one of *N* possible classes. Table 1 defines the variables used to describe the basic classification process. Simply stated, the objective is to associate a particular sample of the RV $\mathbf{x}$ — denoted $\mathbf{x}_p$ — with the correct class ω_c. The method for deciding the class of $\mathbf{x}_p$ yielding the fewest errors [6] (pp. 16-20) can be stated simply:

> *associate* $\mathbf{x}_p$ *with the class* ω_c *that has the largest a posteriori probability:*
>
> $$P(\omega_c \mid \mathbf{x}_p) = P(\omega_{max} \mid \mathbf{x}_p) > P(\omega_j \mid \mathbf{x}_p) \quad \forall\, j \neq c$$

In simple terms, any function that implements this classification procedure constitutes the Bayesian discriminant function.

Clearly, being able to estimate all *N* $P(\omega_i \mid \mathbf{x}_p)$ accurately for each and every $\mathbf{x}_p$ allows one to implement the Bayesian discriminant function. Indeed, a large number of pattern classifiers do precisely this. The degree to which they succeed in the classification task is directly related to the accuracy with which they estimate the *a posterioris*. Another perhaps less obvious way to implement the Bayesian discriminant function is to consistently identify the largest $P(\omega_i \mid \mathbf{x}_p)$ for each and every $\mathbf{x}_p$ — an approach that does not require accurate estimation of the *a posterioris*. The salient point here is that while accurate estimation of the *a posterioris* is *sufficient* for Bayesian discriminant performance, it is *not necessary*. All that is necessary for Bayesian discrimination is accurate identification of the largest *a posteriori*.

These two approaches to implementing the Bayesian discriminant function lead to two broad classes of objective functions that one can use to train the classifier in Figure 1: the class of "reasonable error measures" achieves Bayesian performance by explicitly estimating the *a posterioris* associated with the input $\mathbf{x}_p$, while the Classification Figures of Merit (CFM_{mono}) achieve Bayesian performance by estimating the identity of the maximum *a posteriori* probability $P(\omega_{max} \mid \mathbf{x}_p)$.

3 REASONABLE ERROR MEASURES: BAYESIAN PERFORMANCE VIA ACCURATE ESTIMATION OF *A POSTERIORI* PROBABILITIES

The first class of objective functions that yield Bayesian discriminant performance comprises those error measures engendering classifier outputs that are true estimates of the *a posteriori* probabilities $P(\omega_i \mid \mathbf{x}_p)$. The necessary and sufficient conditions on the form of these functions are given below, followed by a number of familiar examples of the class and detailed proofs of their asymptotic Bayesian performance.

3.1 THE NECESSARY CONDITIONS FOR REASONABLE ERROR MEASURES

Consider a class of error measures $\mathcal{E}[\mathcal{O}_i(\mathbf{x}_p, \theta), \mathcal{D}_i(\mathbf{x}_p)]$ that give the "loss" of a single output $\mathcal{O}_i(\mathbf{x}_p, \theta)$ when its desired or "target" activation is $\mathcal{D}_i(\mathbf{x}_p)$. Tables 1 and 2 define the symbols used to derive this class of error measures. The concept of a *prototype* of $\mathbf{x}$ introduced in these tabulated definitions warrants explanation.

Table 1: Definitions of symbols used to describe the general N-class classification problem.

Symbol	Definition
$\mathbf{x}$	The RV to be classified.
$\mathcal{O}_i$	The ith output of the N-output classifier.
ω_i	The ith of N classes to which $\mathbf{x}$ can belong.
$\mathbf{x}_p$	The pth unique sample (or *prototype*) of $\mathbf{x}$.
$\boldsymbol{\theta}$	The parameterization of the classifier. In the case of an MLP classifier, $\boldsymbol{\theta}$ would represent the connections of the network.
$\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})$	The ith output of the N-output classifier, given the input $\mathbf{x}_p$ and the classifier parameterization $\boldsymbol{\theta}$.
$\mathrm{P}(\omega_i \mid \mathbf{x}_p)$	The *a posteriori* probability of the ith class (ω_i), given the input $\mathbf{x}_p$.
$\mathrm{P}(\overline{\omega_i} \mid \mathbf{x}_p)$	$1 - \mathrm{P}(\omega_i \mid \mathbf{x}_p)$.
$\rho(\mathbf{x} \mid \omega_i)$	The "class conditional" probability density function (PDF) for the RV $\mathbf{x}$ (given class ω_i).

3.1.1 Prototypes: bounds on the complexity of the class-conditional densities of the RV x

A prototype is a *unique* sample $\mathbf{x}_p$ of the RV $\mathbf{x}$. Thus, if one obtains two identical yet statistically independent samples of the RV $\mathbf{x}$, these samples are two instantiations of the same prototype. The notion of obtaining more than one *statistically independent* sample of $\mathbf{x}$ with the *exact* same value $\mathbf{x}_p$ is difficult to envision — even for large training sets. However, if one considers regions on the domain of $\mathbf{x}$ over which the class-conditional densities $\rho(\mathbf{x} \mid \omega_i)$ are essentially constant for all classes, one can associate each of these regions with a *prototypical* value of $\mathbf{x}$. The prototype for the pth of such regions is given by $\mathbf{x}_p$. For an input feature space of dimensionality M and a sufficiently large number of statistically independent samples of $\mathbf{x}$, one might envision an $(M+1)$-dimensional histogram of the samples as an embodiment of this concept of prototypes. Such a view is consistent with the limited resolution of data acquisition systems used to measure real-world RVs.

Clearly this view of regions on $\mathbf{x}$ with constant class-conditional densities places an implicit restriction on the probabilistic nature of $\mathbf{x}$. A simple yet elegant description of a 2-class problem ($N = 2$) involving a 2-dimensional RV $\mathbf{x}$ ($M = 2$) that does *not* have a bounded number of regions of constant class-conditional density is illustrated by the following: if one envisions a two dimensional fractal coastline forming the boundary between land and sea, one finds that in the vicinity of the boundary (shore line) there is no observation scale large enough to yield a bounded number P of regions $\mathbf{x}_p$ within which $\rho(\mathbf{x} \mid \omega_i)$ is constant on all sub-regions of each $\mathbf{x}_p$ for both classes. The RV $\mathbf{x}$ is therefore not "well behaved". Obviously, if $\mathbf{x}$ comprises a finite number of discrete states, then it will be well behaved.

The necessary conditions for reasonable error measures that follow — and all subsequent proofs in this paper — rely on this notion of prototypes. We assume that the RV $\mathbf{x}$ is well-behaved to the extent that P is bounded. This restriction places some limit on the complexity of the class-conditional densities of $\mathbf{x}$ that one can expect to model accurately using an MLP classifier — an issue that we discuss further in section 5.

3.1.2 The reasonable condition

In general, we assume that the outputs of the classifier are bounded on the closed interval [0,1], that there is minimal loss incurred when the output equals its target value, and that there is a symmetry to the loss function:

$$0 \leq \mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}) \leq 1 \qquad \forall\, \mathbf{x}_p, \boldsymbol{\theta} \tag{1}$$

$$\mathcal{E}[z, z] \;<\; \mathcal{E}[y \neq z, z] \tag{2}$$

$$\begin{aligned} &\mathcal{E}[\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}), \mathcal{D}_i(\mathbf{x}_p)] \\ &\quad = \mathcal{E}[\mathcal{D}_i(\mathbf{x}_p) - \mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}), \overline{\mathcal{D}_i}(\mathbf{x}_p)] \end{aligned} \tag{3}$$

The symmetry constraint of (3) can be taken to mean that the reasonable error measure is a function of the absolute difference between the output and its target:

$$\mathcal{E}[\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}), \{\mathcal{D}\}] = f(|\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}) - \{\mathcal{D}\}|) \tag{4}$$

where

$$\{\mathcal{D}\} = \mathcal{D}_i(\mathbf{x}_p) \quad \text{or} \quad \overline{\mathcal{D}_i}(\mathbf{x}_p)$$

Furthermore, if we choose binary targets for our error measure (which incidentally correspond to the upper and lower bounds on the classifier outputs)

$$\begin{aligned} \mathcal{D}_i(\mathbf{x}_p) &\triangleq 1 \\ \overline{\mathcal{D}_i}(\mathbf{x}_p) &\triangleq 0 \end{aligned} \tag{5}$$

then (4) leads to the following functional description of the reasonable error measure:

Table 2: Definitions of symbols used to derive reasonable error measures.

Symbol	Definition
i	index denoting MLP output i of N, associated with class ω_i.
N	the total number of classes.
p	index denoting the pth prototype of $\mathbf{x}$.
P	the total number of prototypes on the domain of $\mathbf{x}$.
n_{pi}	the number of statistically independent occurrences of prototype $\mathbf{x}_p$ belonging to class ω_i.
$n_{p\bar{i}}$	the number of statistically independent occurrences of prototype $\mathbf{x}_p$ *not* belonging to class ω_i.
n_p	$n_{pi} + n_{p\bar{i}}$: the total number of *statistically independent* occurrences of prototype $\mathbf{x}_p$ in the training set.
n_i	$\sum_P n_{pi}$: the total number of statistically independent samples in the training set belonging to class ω_i.
$n_{\bar{i}}$	$\sum_P n_{p\bar{i}}$: the total number of statistically independent samples in the training set *not* belonging to class ω_i.
n_t	$\sum_P n_p$: the total number of statistically independent samples in the training set.
$\mathcal{D}_i(\mathbf{x}_p)$	The target value for $\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})$ when $\mathbf{x}_p$ belongs to class ω_i.
$\overline{\mathcal{D}_i}(\mathbf{x}_p)$	The target value for $\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})$ when $\mathbf{x}_p$ does *not* belong to class ω_i.
$\mathcal{E}[\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}), \mathcal{D}_i(\mathbf{x}_p)]$	The error measure (or loss) for output $\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})$ when its target value is $\mathcal{D}_i(\mathbf{x}_p)$ (i.e., when $\mathbf{x}_p$ belongs to class ω_i).
$\mathcal{E}[\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}), \overline{\mathcal{D}_i}(\mathbf{x}_p)]$	The error measure (or loss) for output $\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})$ when its target value is $\overline{\mathcal{D}_i}(\mathbf{x}_p)$ (i.e., when $\mathbf{x}_p$ does *not* belong to class ω_i).

$$\begin{aligned} \mathcal{E}[\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}), \mathcal{D}_i(\mathbf{x}_p)] &\equiv f(1 - \mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})) \\ \mathcal{E}[\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}), \overline{\mathcal{D}_i}(\mathbf{x}_p)] &\equiv f(\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})) \end{aligned} \tag{6}$$

Using the definitions in tables 1 and 2 with (6), we can express the average error produced by n_t samples of $\mathbf{x}$. Note that these n_t samples are grouped into P prototypes; there are n_p samples of the pth prototype $\mathbf{x}_p$:

$$\begin{aligned} \overline{\mathcal{E}} = \frac{1}{n_t} \sum_{p=1}^{P} \sum_{i=1}^{N} \Big\{ & n_{pi} \cdot f(1 - \mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})) \\ & + n_{p\bar{i}} \cdot f(\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})) \Big\} \end{aligned} \tag{7}$$

Equation (7) can be restated as

$$\begin{aligned} \overline{\mathcal{E}} = \sum_{i=1}^{N} \sum_{p=1}^{P} \frac{n_p}{n_t} \Big\{ & \frac{n_{pi}}{n_p} \cdot f(1 - \mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})) \\ & + \frac{n_{p\bar{i}}}{n_p} \cdot f(\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})) \Big\} \end{aligned} \tag{8}$$

The law of large numbers leads to the following asymptotic form for the average error:

$$\begin{aligned} \lim_{n_t \to \infty} \overline{\mathcal{E}} = \sum_{i=1}^{N} \sum_{p=1}^{P} \mathrm{P}(\mathbf{x}_p) \{ & \mathrm{P}(\omega_i \,|\, \mathbf{x}_p) \cdot f(1 - \mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})) \\ & + \mathrm{P}(\overline{\omega_i} \,|\, \mathbf{x}_p) \cdot f(\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})) \} \end{aligned} \tag{9}$$

A necessary and sufficient condition for minimizing $\overline{\mathcal{E}}$ in (9) is $|\nabla_{\mathcal{O}} \overline{\mathcal{E}}| = 0$, which requires that

$$\begin{aligned} \frac{d}{d\mathcal{O}_i(\mathbf{x}, \boldsymbol{\theta})} \overline{\mathcal{E}} &= \sum_{p=1}^{P} \mathrm{P}(\mathbf{x}_p) \{ -\mathrm{P}(\omega_i \,|\, \mathbf{x}_p) \cdot f'(1 - \mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})) \\ &\quad + \mathrm{P}(\overline{\omega_i} \,|\, \mathbf{x}_p) \cdot f'(\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})) \} \\ &= 0 \quad \forall\, i \end{aligned} \tag{10}$$

where

$$f'(u) \triangleq \frac{d}{du} f(u)$$

Equation (10), in turn, is satisfied if

$$\begin{aligned} \frac{f'(\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}))}{f'(1 - \mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}))} &= \frac{\mathrm{P}(\omega_i \,|\, \mathbf{x}_p)}{\mathrm{P}(\overline{\omega_i} \,|\, \mathbf{x}_p)} \\ &= \frac{\mathrm{P}(\omega_i \,|\, \mathbf{x}_p)}{1 - \mathrm{P}(\omega_i \,|\, \mathbf{x}_p)} \quad \forall\, \mathbf{x}_p \end{aligned} \tag{11}$$

Note that (11) is both a necessary and sufficient condition for satisfying (10) for *all* possible distributions of $\mathbf{x}_p$ (which are directly related to the class-conditional densities of $\mathbf{x}$ (see section 3.1.1)). While it is possible to satisfy (10) without satisfying (11) for *some* distributions of $\mathbf{x}_p$ (e.g., some sets of class-conditional densities $\{\rho(\mathbf{x} \mid \omega_i)\}$), (11) must hold for (10) to hold independent of $\{\rho(\mathbf{x} \mid \omega_i)\}$. As a trivial example, if $\mathrm{P}(\mathbf{x}_p)$ were zero for all but one prototype, satisfying (10) would require satisfying (11).

Clearly, the Hessian of the average error ($\mathrm{H}_{\mathcal{O}}\,\overline{\mathcal{E}}$) must be positive definite in order for (11) to yield a minimum average error:

$$\left|\mathrm{H}_{\mathcal{O}}\,\overline{\mathcal{E}}\right| \;>\; 0 \tag{12}$$

One can show that if $\overline{\mathcal{E}}$ in (4) is a strictly increasing function of $|\mathcal{O}_i(\mathbf{x}_p,\,\boldsymbol{\theta}) \;-\; \{\mathcal{D}\}|$, (12) will hold.

Equations (11) and (12) ensure a minimum of $\overline{\mathcal{E}}$ but they place no explicit condition on the form of $\mathcal{O}_i(\mathbf{x}_p,\,\boldsymbol{\theta})$. Since we wish the outputs of the classifier to equal the *a posteriori* probabilities, we can assure this equivalence by constraining the reasonable error measure's functional form based on (11):

$$f'(\mathcal{O}_i) \;=\; \frac{\mathcal{O}_i}{1 \;-\; \mathcal{O}_i} f'(1 - \mathcal{O}_i) \quad 0 \le \mathcal{O}_i < 1 \tag{13}$$

Any function satisfying the conditions of (3) – (6) and (10) – (13) is a reasonable error measure. Such a measure will yield classifier outputs that asymptotically equate to the *a posteriori* probabilities $\mathrm{P}(\omega_i \mid \mathbf{x})$, provided the functional capacity of the classifier (i.e., the classifier's ability to model the function that maps the RV $\mathbf{x}$ to the *a posterioris* $\mathrm{P}(\omega_i \mid \mathbf{x})$ for all $\mathbf{x}_p$), embodied in the parameterization variable $\boldsymbol{\theta}$, is at least as great as the complexity of all the class-conditional densities $\rho(\mathbf{x} \mid \omega_i)$. This statement relies on the assumption that these class conditional densities are restricted to those that are well behaved (see section 3.1.1). This, combined with the finding that a MLP with a single hidden layer of adequate connectivity can — under mild constraints consistent with our assumptions — approximate any continuous function mapping $\mathbf{x}$ onto the N-dimensional hypercube [4], assures that there exists a MLP that will accurately model the Bayesian discriminant functions of any well-behaved $\mathbf{x}$, given a sufficiently large set of statistically independent training samples.

Finally, one can show that any positively scaled reasonable error measures is, itself, a reasonable error measure. That is, if $f_1(\mathcal{O})$ is a reasonable error measure, then $a f_1(\mathcal{O})$ will also be reasonable if $a > 0$.

3.2 THE GENERAL REASONABLE ERROR MEASURE APPROXIMATION TO THE BAYESIAN DISCRIMINANT FUNCTION

If one defines the Bayesian discriminant function for the ith of N possible classes as

$$g_i(\mathbf{x}) \;\triangleq\; \mathrm{P}(\omega_i \mid \mathbf{x}) \tag{14}$$

where

$$\mathrm{P}(\mathbf{x}) \;=\; \sum_{j=1}^{N} \mathrm{P}(\mathbf{x} \mid \omega_j) \tag{15}$$

one can define the *reasonable approximation error* for the ith discriminant function as

$$\begin{aligned}\epsilon_i \;\triangleq\; \int_{\mathbf{X}} \big[& f\,(1 - \mathcal{O}_i(\mathbf{x},\,\boldsymbol{\theta})) \cdot g_i(\mathbf{x}) \\ & + f\,(\mathcal{O}_i(\mathbf{x},\,\boldsymbol{\theta})) \cdot (1 \;-\; g_i(\mathbf{x}))\big]\; \rho(\mathbf{x})\, d\mathbf{x}\end{aligned} \tag{16}$$

Additionally, one can define the *aggregate* reasonable approximation error as

$$\epsilon \;=\; \sum_{i=1}^{N} \epsilon_i \tag{17}$$

Given (9), one can express the asymptotic average reasonable error of the training set. One can in turn express the asymptotic average reasonable error in terms of the aggregate reasonable approximation error to the Bayesian discriminant function expressed in (16) and (17). Duda and Hart first showed such a relationship for the simple perceptron trained with the MSE objective function in [6] (pp. 154-155). The symbol "$\asymp$" should be read as, "asymptotically equals."

$$\begin{aligned}\lim_{n_t \to \infty} \overline{\mathcal{E}} \;&=\; \sum_{p=1}^{P} \mathrm{P}(\mathbf{x}_p) \sum_{i=1}^{N} \{\mathrm{P}(\omega_i \mid \mathbf{x}_p) \cdot f\,(1 - \mathcal{O}_i(\mathbf{x}_p,\,\boldsymbol{\theta})) \\ &\quad + \mathrm{P}(\overline{\omega_i} \mid \mathbf{x}_p) \cdot f\,(\mathcal{O}_i(\mathbf{x}_p,\,\boldsymbol{\theta}))\} \\ &=\; \sum_{p=1}^{P} \sum_{i=1}^{N} \Big\{ \mathrm{P}(\omega_i,\,\mathbf{x}_p) \cdot f\,(1 - \mathcal{O}_i(\mathbf{x}_p,\,\boldsymbol{\theta})) \\ &\quad + \mathrm{P}(\overline{\omega_i},\,\mathbf{x}_p) \cdot f\,(\mathcal{O}_i(\mathbf{x}_p,\,\boldsymbol{\theta})) \Big\} \\ &\asymp\; \sum_{i=1}^{N} \left\{ \mathrm{P}(\omega_i)\, \mathrm{E}\left[f\,(1 - \mathcal{O}_i(\mathbf{x},\,\boldsymbol{\theta})) \mid \omega_i \right] \right. \\ &\quad \left. + \mathrm{P}(\overline{\omega_i})\, \mathrm{E}\left[f\,(\mathcal{O}_i(\mathbf{x},\,\boldsymbol{\theta})) \mid \overline{\omega_i} \right] \right\} \\ &=\; \sum_{i=1}^{N} \left\{ \int_{\mathbf{X}} f\,(1 - \mathcal{O}_i(\mathbf{x},\,\boldsymbol{\theta})) \cdot \rho(\mathbf{x} \mid \omega_i) \right. \\ &\quad \left. \cdot\, \mathrm{P}(\omega_i)\, d\mathbf{x} \right.\end{aligned}$$

$$+ \int_{\mathbf{x}} f\left(\mathcal{O}_i(\mathbf{x},\,\theta)\right) \cdot \rho(\mathbf{x} \mid \overline{\omega_i}) \cdot \mathrm{P}(\overline{\omega_i})\, d\mathbf{x} \Bigg\}$$

$$= \sum_{i=1}^{N} \left\{ \int_{\mathbf{x}} f\left(1 - \mathcal{O}_i(\mathbf{x},\,\theta)\right) \rho(\mathbf{x},\,\omega_i)\, d\mathbf{x} + \int_{\mathbf{x}} f\left(\mathcal{O}_i(\mathbf{x},\,\theta)\right) \rho(\mathbf{x},\,\overline{\omega_i})\, d\mathbf{x} \right\} \tag{18}$$

Since

$$\begin{aligned} \rho(\mathbf{x},\,\omega_i) &= \frac{d}{d\mathbf{x}} \mathrm{P}(\mathbf{x},\,\omega_i) \\ &= \frac{d}{d\mathbf{x}} \mathrm{P}(\omega_i,\,\mathbf{x}) \\ &= \frac{d}{d\mathbf{x}} \left[\mathrm{P}(\omega_i \mid \mathbf{x}) \cdot \mathrm{P}(\mathbf{x})\right] \\ &= \mathrm{P}(\omega_i \mid \mathbf{x}) \cdot \rho(\mathbf{x}) \end{aligned} \tag{19}$$

and

$$\rho(\mathbf{x},\,\overline{\omega_i}) = \mathrm{P}(\overline{\omega_i} \mid \mathbf{x}) \cdot \rho(\mathbf{x}) \tag{20}$$

one can re-state the expression of (18) as

$$\begin{aligned} \lim_{n_t \to \infty} \overline{\mathcal{E}} &= \\ &\sum_{i=1}^{N} \left\{ \int_{\mathbf{x}} f\left(1 - \mathcal{O}_i(\mathbf{x},\,\theta)\right) \mathrm{P}(\omega_i \mid \mathbf{x}) \cdot \rho(\mathbf{x})\, d\mathbf{x} + \int_{\mathbf{x}} f\left(\mathcal{O}_i(\mathbf{x},\,\theta)\right) \mathrm{P}(\overline{\omega_i} \mid \mathbf{x}) \cdot \rho(\mathbf{x})\, d\mathbf{x} \right\} \\ &= \sum_{i=1}^{N} \left\{ \underbrace{\int_{\mathbf{x}} \left[f\left(1 - \mathcal{O}_i(\mathbf{x},\,\theta)\right) \cdot g_i(\mathbf{x}) + f\left(\mathcal{O}_i(\mathbf{x},\,\theta)\right) \cdot \left(1 - g_i(\mathbf{x})\right) \right] \rho(\mathbf{x})\, d\mathbf{x}}_{\epsilon_i} \right\} \\ &= \epsilon \end{aligned} \tag{21}$$

Clearly then, minimizing the reasonable error measure of (7) also minimizes the reasonable approximation errors of (16) and (17). In order for ϵ in (17) and (21) to be zero, it is necessary that the MLP's functional capacity exceed the functional complexity of all the class-conditional densities $\rho(\mathbf{x} \mid \omega_i)$ (see section 3.1.2).

3.3 SPECIFIC EXAMPLES OF REASONABLE ERROR MEASURES

One family of reasonable functions, which can be derived by inspection of (13), is

$$f(\mathcal{O}) = \int \mathcal{O}^r (1 - \mathcal{O})^{r-1}\, d\mathcal{O} \tag{22}$$

This family has two special cases of great practical importance.

3.3.1 $r = 0$: Information Theoretic objective functions

One function that satisfies the reasonable conditions is

$$\begin{aligned} f(\mathcal{O}) &= \int (1 - \mathcal{O})^{-1}\, d\mathcal{O} \\ &= -\log(1 - \mathcal{O}) \end{aligned} \tag{23}$$

— the functional expression used to implement the Cross Entropy, Maximum Mutual Information, Kullback-Liebler distance, and Maximum Likelihood objective functions [7, 10].

3.3.2 $r = 1$: Mean Squared Error

The MSE objective function is also a special case of (22):

$$\begin{aligned} f(\mathcal{O}) &= \int \mathcal{O}\, d\mathcal{O} \\ &= \frac{1}{2} \mathcal{O}^2 \end{aligned} \tag{24}$$

3.4 SOME "UNREASONABLE" ERROR MEASURES

Obviously, any objective function which does not satisfy the *necessary* reasonable conditions will be an unreasonable function for estimating *a posteriori* probabilities. Nevertheless, many such unreasonable error functions will still yield asymptotic Bayesian discriminant performance. If its outputs asymptotically reflect the correct *ranking* of the *a posterioris*, an unreasonable error measure will yield Bayesian discriminant performance. We discuss two classes of objective functions that are unreasonable.

3.4.1 Minkowski-R error measures

When the objective function is of the form $f(\mathcal{O}) = \mathcal{O}^R$ — which corresponds to a Minkowski-R (L_R) metric [9] — one finds that the reasonable condition is satisfied only when

$$\begin{aligned} \mathcal{O}^{R-1} &= \frac{\mathcal{O}}{1 - \mathcal{O}} (1 - \mathcal{O})^{R-1} \\ \mathcal{O}^{R-2} &= (1 - \mathcal{O})^{R-2} \end{aligned}$$

or $R = 2$ ($r = 1$, in section 3.3.2). Another perspective is that $\overline{\mathcal{E}}$ is minimized when

$$\mathcal{O}_i(\mathbf{x}_p,\,\theta) = \sqrt[R-1]{\mathrm{P}(\omega_i \mid \mathbf{x}_p)} \cdot \left[\sqrt[R-1]{\mathrm{P}(\overline{\omega_i} \mid \mathbf{x}_p)} + \sqrt[R-1]{\mathrm{P}(\omega_i \mid \mathbf{x}_p)} \right]^{-1}$$

which simplifies to $\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}) = \mathrm{P}(\omega_i \mid \mathbf{x}_p)$ only when $R = 2$ (note that the L_2 metric is the MSE objective function). Since an L_R metric is reasonable only when $R = 2$, this argues against using L_R metrics other than L_2 when the output of the classifier is being interpreted as an *a posteriori* probability.

Figure 2 gives an intuitive feel for how various L_R metrics bias the output $\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})$ towards certainty (for $R \rightarrow 1$), or away from it (for $R \rightarrow \infty$): the minimum error value for $\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})$ is plotted as a function of $\mathrm{P}(\omega_i \mid \mathbf{x}_p)$ for various values of R. It should be noted that while L_R metrics are generally not reasonable error measures, they do in fact yield classifier outputs that asymptotically reflect the correct *ranking* of *a posteriori* probabilities.[3] Strictly speaking, they will yield Bayesian discriminant performance, and one can defend their use in training classifiers if the biases towards or away from certainty depicted in Figure 2 are not excessive for one's application.

3.4.2 Error measures with non-binary targets

Another class of unreasonable error measures is found if one employs otherwise reasonable error measures with non-binary targets $\{\mathcal{D}\}$. In such cases the resulting error measure will not be reasonable. Whether or not the resulting error measure reflects the correct *a posteriori* probability rankings depends on the choice of non-binary targets. We illustrate this point in the following sections as we derive the approximation error to the Bayesian discriminant function for the MSE and information theoretic error measures.

3.5 THE MSE APPROXIMATION TO THE BAYESIAN DISCRIMINANT FUNCTION

Using (16), one can define the *mean-squared approximation error* for the ith discriminant function as

$$\epsilon_i \triangleq \int_{\mathbf{X}} \Big\{ \big[\mathcal{O}_i(\mathbf{x}) - g_i(\mathbf{x})\big]^2 - g_i(\mathbf{x})^2 + g_i(\mathbf{x}) \Big\} \rho(\mathbf{x})\, d\mathbf{x} \tag{25}$$

Additionally, one can define the *aggregate* mean-squared approximation error as

$$\epsilon = \sum_{i=1}^{N} \epsilon_i \tag{26}$$

One can express the average mean-squared error of the training set as

$$MSE \triangleq \frac{1}{n_t} \sum_{p=1}^{P} \sum_{i=1}^{N} \Big\{ n_{pi} \cdot \big[\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}) - \mathcal{D}_i(\mathbf{x}_p)\big]^2 + n_{p\bar{i}} \cdot \big[\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}) - \overline{\mathcal{D}_i}(\mathbf{x}_p)\big]^2 \Big\} \tag{27}$$

Following the litany of section 3.1.2, one can express the asymptotic average mean-squared error as

$$\lim_{n_t \rightarrow \infty} MSE = \sum_{p=1}^{P} \mathrm{P}(\mathbf{x}_p) \sum_{i=1}^{N} \Big\{ \mathrm{P}(\omega_i \mid \mathbf{x}_p) \cdot \big[\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}) - \mathcal{D}_i(\mathbf{x}_p)\big]^2 + \mathrm{P}(\overline{\omega_i} \mid \mathbf{x}_p) \cdot \big[\mathcal{O}_i(\mathbf{x}_p) - \overline{\mathcal{D}_i}(\mathbf{x}_p)\big]^2 \Big\} \tag{28}$$

From this asymptotic form, one can show that the necessary condition for minimum MSE is

$$\lim_{n_t \rightarrow \infty} \mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}) = \mathcal{D}_i(\mathbf{x}_p) \cdot \mathrm{P}(\omega_i \mid \mathbf{x}_p) + \overline{\mathcal{D}_i}(\mathbf{x}_p) \cdot \mathrm{P}(\overline{\omega_i} \mid \mathbf{x}_p) \;\; \forall\, \mathbf{x}_p \tag{29}$$

For the case in which binary targets as specified in (5) are used, the MSE objective function constitutes a reasonable error measure, and the classifier outputs asymptotically equate to the *a posterioris*. If $\mathcal{D}_i$ and $\overline{\mathcal{D}_i}$ are both set equal to the same value on the closed interval $[0, 1]$, this will lead to a most undesirable asymptotic state in which all classifier outputs converge to $\mathcal{D}_i$ — a state of complete uncertainty analogous to that attained by the Minkowski-R error metric L_∞. For the case in which $\mathcal{D}_i > \overline{\mathcal{D}_i}$ and both targets are non-binary on [0,1], one finds that $\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})$ is no longer an accurate estimate of $\mathrm{P}(\omega_i \mid \mathbf{x}_p)$, although it does remain a strictly increasing function of $\mathrm{P}(\omega_i \mid \mathbf{x}_p)$. For the bizarre case in which $\mathcal{D}_i < \overline{\mathcal{D}_i}$, $\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})$ becomes a strictly increasing function of $\mathrm{P}(\overline{\omega_i} \mid \mathbf{x}_p)$ (or $1 - \mathrm{P}(\omega_i \mid \mathbf{x}_p)$). Figure 3 illustrates the effect of different target values on the asymptotic value of $\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})$ plotted as a function of $\mathrm{P}(\omega_i \mid \mathbf{x}_p)$. As we shall see in the next section, this figure is relevant to information theoretic objective functions as well.

Returning to (28) one can derive the asymptotic mean-squared error (binary targets: $\mathcal{D}_i = 1, \overline{\mathcal{D}_i} = 0$) in terms of the aggregate approximation error to the Bayesian discriminant function (expressed in (25) and (26)). Using derivational procedures analogous to those of equations (18) – (21), one finds

$$\lim_{n_t \rightarrow \infty} MSE = \sum_{p=1}^{P} \mathrm{P}(\mathbf{x}_p) \sum_{i=1}^{N} \Big\{ \mathrm{P}(\omega_i \mid \mathbf{x}_p) \cdot \big[\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}) - 1\big]^2 + \mathrm{P}(\overline{\omega_i} \mid \mathbf{x}_p) \cdot \mathcal{O}_i(\mathbf{x}_p)^2 \Big\} \asymp \sum_{i=1}^{N} \Big\{ \mathrm{P}(\omega_i)\, \mathrm{E}\Big[(\mathcal{O}_i(\mathbf{x}, \boldsymbol{\theta}) - 1)^2 \mid \omega_i \Big]$$

[3] This is because $\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})$ is asymptotically a strictly increasing function of $\mathrm{P}(\omega_i \mid \mathbf{x}_p)$.

$$+ \mathrm{P}(\overline{\omega_i})\ \mathrm{E}\left[(\mathcal{O}_i(\mathbf{x},\ \boldsymbol{\theta}))^2\ |\ \overline{\omega_i}\right]\Big\}$$

$$= \sum_{i=1}^{N}\Big\{\int_{\mathbf{X}}(\mathcal{O}_i(\mathbf{x},\ \boldsymbol{\theta})\ -\ 1)^2\ \rho(\mathbf{x},\ \omega_i)\,d\mathbf{x} + \int_{\mathbf{X}}\mathcal{O}_i(\mathbf{x},\ \boldsymbol{\theta})^2\ \rho(\mathbf{x},\ \overline{\omega_i})\,d\mathbf{x}\Big\} \tag{30}$$

$$= \sum_{i=1}^{N}\Big\{\int_{\mathbf{X}}\left[\mathcal{O}_i(\mathbf{x},\ \boldsymbol{\theta})^2\ -\ 2\,\mathcal{O}_i(\mathbf{x},\ \boldsymbol{\theta}) + 1\right] \cdot g_i(\mathbf{x})\,\rho(\mathbf{x})\,d\mathbf{x} + \int_{\mathbf{X}}\mathcal{O}_i(\mathbf{x},\ \boldsymbol{\theta})^2\ \left[1\ -\ g_i(\mathbf{x})\right] \cdot\ \rho(\mathbf{x})\,d\mathbf{x}\Big\} \tag{31}$$

$$= \sum_{i=1}^{N}\Big\{\int_{\mathbf{X}}\mathcal{O}_i(\mathbf{x},\ \boldsymbol{\theta})^2\ \rho(\mathbf{x})\,d\mathbf{x} - 2\int_{\mathbf{X}}\mathcal{O}_i(\mathbf{x},\ \boldsymbol{\theta})\ g_i(\mathbf{x})\ \rho(\mathbf{x})\,d\mathbf{x} + \int_{\mathbf{X}}g_i(\mathbf{x})\ \rho(\mathbf{x})\,d\mathbf{x}\Big\}$$

$$\lim_{n_t\to\infty} MSE = \sum_{i=1}^{N}\left\{\begin{array}{l}\int_{\mathbf{X}}\left[\mathcal{O}_i(\mathbf{x},\ \boldsymbol{\theta})\ -\ g_i(\mathbf{x})\right]^2\ \rho(\mathbf{x})\,d\mathbf{x} \\ \underbrace{-\int_{\mathbf{X}}(g_i(\mathbf{x}))^2\ \rho(\mathbf{x})\,d\mathbf{x}\ +\ \mathrm{P}(\omega_i)}_{\epsilon_i}\end{array}\right\} \tag{32}$$

$$= \epsilon$$

This result is the MLP analog of Duda and Hart's result for the MSE-trained perceptron ([6], pp. 154-155). A comparison of (32) and (25) confirms that each of the N terms in (32) is equivalent to the mean-squared approximation error term of (25). Thus, minimizing the MSE objective function of (27) (binary targets) also minimizes the mean-squared approximation errors of (25) and (26). Note that only the first term in (32) depends upon the output activations $\mathcal{O}_i$ of the MLP. In order for ϵ in (26) to be zero, it is necessary that the MLP's functional capacity exceed the functional complexity of all the class-conditional densities $\rho(\mathbf{x}\,|\,\omega_i)$ (see section 3.1.2).

Equation (32) illustrates the manner in which MSE is minimized during classifier training. The mean-squared approximation error term (ϵ) indicates that MSE is in fact a weighted integral sum of the squared errors between the MLP outputs $\mathcal{O}_i$ and their corresponding discriminant functions. The weighting factor is $\rho(\mathbf{x})$. The form of ϵ_i in (32) indicates that the approximation error minimization process focuses on the mode(s) of $\mathbf{x}$, where $\rho(\mathbf{x})$ is large. This issue is discussed further in section 5.

3.6 INFORMATION THEORETIC APPROXIMATIONS TO THE BAYESIAN DISCRIMINANT FUNCTION

Reference [7] shows that the information theoretic learning paradigms of Maximum Mutual Information, Kullback-Liebler distance, and Maximum Likelihood lead to a reasonable error measure known in the connectionist literature as the Cross Entropy (CE) objective function [10]. This error measure applied to a single input sample $\mathbf{x}_p$ belonging to class ω_i is expressed by

$$CE \triangleq -\sum_{i=1}^{N}\Big\{\mathcal{D}_i(\mathbf{x}_p)\ \log\{\mathcal{O}_i(\mathbf{x}_p,\ \boldsymbol{\theta})\} + (1\ -\ \mathcal{D}_i(\mathbf{x}_p))\ \log\{1\ -\ \mathcal{O}_i(\mathbf{x}_p,\ \boldsymbol{\theta})\}\Big\} \tag{33}$$

Given the Bayesian discriminant functions of (14), one can define the *cross-entropy approximation error* for the ith discriminant function as

$$\epsilon_i \triangleq -\int_{\mathbf{X}}\Big[g_i(\mathbf{x})\ \log\{\mathcal{O}_i(\mathbf{x},\ \boldsymbol{\theta})\} + (1\ -\ g_i(\mathbf{x}))\ \log\{1\ -\ \mathcal{O}_i(\mathbf{x},\ \boldsymbol{\theta})\}\Big]\ \rho(\mathbf{x})\,d\mathbf{x} \tag{34}$$

The *aggregate* cross-entropy approximation error is then given by

$$\epsilon = \sum_{i=1}^{N}\epsilon_i \tag{35}$$

Given the definitions of tables 1 and 2, one can express the total cross entropy of the training set as

$$CE \triangleq -\frac{1}{n_t}\sum_{p=1}^{P}\sum_{i=1}^{N}\Big\{n_{pi}\ \cdot\ \left[\mathcal{D}_i(\mathbf{x}_p)\ \log\{\mathcal{O}_i(\mathbf{x}_p,\ \boldsymbol{\theta})\} + \left(1\ -\ \mathcal{D}_i(\mathbf{x}_p)\right)\ \log\{1\ -\ \mathcal{O}_i(\mathbf{x}_p,\ \boldsymbol{\theta})\}\right] + n_{p\bar{i}}\ \cdot\ \left[\overline{\mathcal{D}_i}(\mathbf{x}_p)\ \log\{\mathcal{O}_i(\mathbf{x}_p,\ \boldsymbol{\theta})\} + \left(1\ -\ \overline{\mathcal{D}_i}(\mathbf{x}_p)\right)\ \log\{1\ -\ \mathcal{O}_i(\mathbf{x}_p,\ \boldsymbol{\theta})\}\right]\Big\} \tag{36}$$

Following the litany of section 3.1.2, one can express the asymptotic cross entropy as

$$\lim_{n_t \to \infty} CE = -\sum_{p=1}^{P} \mathrm{P}(\mathbf{x}_p) \sum_{i=1}^{N} \Big\{ \mathrm{P}(\omega_i \,|\, \mathbf{x}_p) \left[\mathcal{D}_i(\mathbf{x}_p) \log\{\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})\} + \left(1 - \mathcal{D}_i(\mathbf{x}_p)\right) \log\{1 - \mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})\} \right] + \mathrm{P}(\overline{\omega_i} \,|\, \mathbf{x}_p) \cdot \left[\overline{\mathcal{D}_i}(\mathbf{x}_p) \log\{\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})\} + \left(1 - \overline{\mathcal{D}_i}(\mathbf{x}_p)\right) \log\{1 - \mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})\} \right] \Big\} \quad (37)$$

From this asymptotic form, one can show that the necessary condition for minimum Cross Entropy is

$$\lim_{n_t \to \infty} \mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}) = \mathcal{D}_i(\mathbf{x}_p) \cdot \mathrm{P}(\omega_i \,|\, \mathbf{x}_p) + \overline{\mathcal{D}_i}(\mathbf{x}_p) \cdot \mathrm{P}(\overline{\omega_i} \,|\, \mathbf{x}_p) \;\; \forall\, \mathbf{x}_p \quad (38)$$

— precisely the same condition required for minimizing the MSE objective function. For this reason, the comments following (29) and Figure 3 accurately describe the dependence of information theoretic classifier outputs on target values: binary targets yield classifier outputs that asymptotically equate to the *a posterioris* $\mathrm{P}(\omega_i \,|\, \mathbf{x}_p)$.

Returning to (37) one can derive the asymptotic Cross Entropy (binary targets: $\mathcal{D}_i = 1$, $\overline{\mathcal{D}_i} = 0$) in terms of the aggregate approximation error to the Bayesian discriminant function (expressed in (34) and (35)). Using derivational procedures analogous to those of equations (18) – (21), one finds

$$\begin{aligned} \lim_{n_t \to \infty} CE &= -\sum_{p=1}^{P} \mathrm{P}(\mathbf{x}_p) \sum_{i=1}^{N} \Big\{ \mathrm{P}(\omega_i \,|\, \mathbf{x}_p) \log\{\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})\} + \mathrm{P}(\overline{\omega_i} \,|\, \mathbf{x}_p) \log\{1 - \mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})\} \Big\} \quad (39) \\ &\asymp -\sum_{i=1}^{N} \Big\{ \mathrm{P}(\omega_i)\, \mathrm{E}\left[\log\{\mathcal{O}_i(\mathbf{x}, \boldsymbol{\theta})\} \,|\, \omega_i\right] + \mathrm{P}(\overline{\omega_i})\, \mathrm{E}\left[\log\{1 - \mathcal{O}_i(\mathbf{x}, \boldsymbol{\theta})\} \,|\, \overline{\omega_i}\right] \Big\} \\ &= -\sum_{i=1}^{N} \Big\{ \int_{\mathbf{X}} \log\{\mathcal{O}_i(\mathbf{x}, \boldsymbol{\theta})\}\, \rho(\mathbf{x}, \omega_i)\, d\mathbf{x} + \int_{\mathbf{X}} \log\{1 - \mathcal{O}_i(\mathbf{x}, \boldsymbol{\theta})\}\, \rho(\mathbf{x}, \omega_i)\, d\mathbf{x} \Big\} \end{aligned}$$

$$\begin{aligned} \lim_{n_t \to \infty} CE &= \sum_{i=1}^{N} \Big\{ \int_{\mathbf{X}} \log\{\mathcal{O}_i(\mathbf{x}, \boldsymbol{\theta})\} \underbrace{P(\omega_i \,|\, \mathbf{x})}_{g_i(\mathbf{X})} \rho(\mathbf{x})\, d\mathbf{x} + \int_{\mathbf{X}} \log\{1 - \mathcal{O}_i(\mathbf{x}, \boldsymbol{\theta})\} \underbrace{P(\overline{\omega_i} \,|\, \mathbf{x})}_{1 - g_i(\mathbf{X})} \rho(\mathbf{x})\, d\mathbf{x} \Big\} \quad (40) \\ &= \epsilon \end{aligned}$$

A comparison of (40) and (34) confirms that each of the N terms in (40) is equivalent to the cross entropy approximation error term of (34). Thus, minimizing the cross entropy objective function of (36) (binary targets) also minimizes the cross entropy approximation errors of (34) and (35). As with the MSE objective function, it is necessary that the MLP's functional capacity exceed the functional complexity of all the class-conditional densities $\rho(\mathbf{x} \,|\, \omega_i)$ in order for ϵ in (35) to be zero (see section 3.1.2).

Note that CE, much like its MSE counterpart, is a weighted integral sum of the cross entropy between between the MLP outputs $\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})$ and their corresponding discriminant functions. As with the MSE objective function, the form of ϵ_i for the CE objective function in (40) indicates that the approximation error minimization process focuses on the mode(s) of $\mathbf{x}$, where $\rho(\mathbf{x})$ is large (see section 5).

4 CLASSIFICATION FIGURES OF MERIT: LIMITED BAYESIAN PERFORMANCE WITHOUT EXPLICIT ESTIMATION OF *A POSTERIORI* PROBABILITIES

The N-monotonic CFM objective function [8] is given by

$$CFM_{mono} \triangleq \frac{1}{n_t} \sum_{p=1}^{P} \sum_{i=1}^{N} \left\{ n_{pi} \cdot \sigma\left[\Delta_i(\mathbf{x}_p, \boldsymbol{\theta})\right] \right\} \quad (41)$$

where

$$\Delta_i(\mathbf{x}_p, \boldsymbol{\theta}) \triangleq \mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta}) - max\, \mathcal{O}_{j \neq i}(\mathbf{x}_p) \quad (42)$$

Thus, for n_{pi} cases of the prototype $\mathbf{x}_p$, output $\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})$ represents the correct class ω_i, while output $\mathcal{O}_{j \neq i}(\mathbf{x}_p)$ is the most active output representing an incorrect class $\omega_{j \neq i}$. The function $\sigma\left[\Delta_i(\mathbf{x}_p, \boldsymbol{\theta})\right]$ is typically a strictly increasing continuously differentiable function of $\Delta_i(\mathbf{x}_p, \boldsymbol{\theta})$. The asymptotic form for CFM_{mono} is

$$\lim_{n_t \to \infty} \mathrm{CFM}_{mono} =$$

Table 3: A ranking of the *N a posterioris* (and their corresponding CFM terms) of $\mathrm{CFM}_{mono}(\mathbf{x}_p)$ in (43).

$$\left.\begin{array}{lcl} \mathrm{P}(\omega_{r1}\,|\,\mathbf{x}_p) & : & \sigma\big[\Delta_{r1}(\mathbf{x}_p,\,\boldsymbol{\theta})\big] \\ \mathrm{P}(\omega_{r2}\,|\,\mathbf{x}_p) & : & \sigma\big[\Delta_{r2}(\mathbf{x}_p,\,\boldsymbol{\theta})\big] \\ \mathrm{P}(\omega_{r3}\,|\,\mathbf{x}_p) & : & \sigma\big[\Delta_{r3}(\mathbf{x}_p,\,\boldsymbol{\theta})\big] \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \mathrm{P}(\omega_{rN}\,|\,\mathbf{x}_p) & : & \sigma\big[\Delta_{rN}(\mathbf{x}_p,\,\boldsymbol{\theta})\big] \end{array}\right\} = \left\{\begin{array}{l} \sigma(\underline{\mathcal{O}_{r1}} - \mathcal{O}_{r2}) \\ \sigma(\underline{\mathcal{O}_{r2}} - \mathcal{O}_{r1}) \\ \sigma(\underline{\mathcal{O}_{r3}} - \mathcal{O}_{r1}) \\ \cdot \\ \cdot \\ \cdot \\ \sigma(\underline{\mathcal{O}_{rN}} - \mathcal{O}_{r1}) \end{array}\right.$$

$$\sum_{p=1}^{P} \mathrm{P}(\mathbf{x}_p) \underbrace{\sum_{i=1}^{N} \mathrm{P}(\omega_i\,|\,\mathbf{x}_p) \cdot \sigma\big[\Delta_i(\mathbf{x}_p,\,\boldsymbol{\theta})\big]}_{\mathrm{CFM}_{mono}(\mathbf{x}_p)} \quad (43)$$

$$\asymp \sum_{i=1}^{N} \mathrm{E}\left[\mathrm{P}(\omega_i\,|\,\mathbf{x}) \cdot \sigma\big[\Delta_i(\mathbf{x},\,\boldsymbol{\theta})\big]\right] \quad (44)$$

$$= \sum_{i=1}^{N} \int_{\mathbf{X}} \underbrace{\mathrm{P}(\omega_i\,|\,\mathbf{x})}_{g_i(\mathbf{X})} \cdot \sigma\big[\Delta_i(\mathbf{x},\,\boldsymbol{\theta})\big]\ \rho(\mathbf{x})\,d\mathbf{x} \quad (45)$$

Because $\sigma[\Delta]$ is a strictly increasing function of Δ, $\sigma'[\Delta] > 0$ and it is impossible to find the maximum of (43) by solving for the zero of its gradient with respect to the outputs $\{\mathcal{O}\}$. Furthermore, the identity of $\mathcal{O}_{j\neq i}(\mathbf{x}_p)$ in (42) is stochastic, so (43) is not a continuously differentiable function of the classifier outputs. As a result one cannot analytically determine the maximum CFM_{mono} values of the classifier outputs $\mathcal{O}_i(\mathbf{x},\,\boldsymbol{\theta})$. Nevertheless, it is useful to consider how $\mathrm{CFM}_{mono}(\mathbf{x}_p)$ — the CFM for a single prototype $\mathbf{x}_p$ — in (43) is maximized. Table 3 depicts the *a posteriori* and N $\mathrm{CFM}_{mono}(\mathbf{x}_p)$ terms from (43) associated with the prototype $\mathbf{x}_p$. The *a posterioris* are ranked in decreasing order; their associated CFM_{mono} terms are ranked in the same order. Thus $\mathrm{P}(\omega_{r1}\,|\,\mathbf{x}_p)$ is the largest *a posteriori* for $\mathbf{x}_p$ while $\mathrm{P}(\omega_{rN}\,|\,\mathbf{x}_p)$ is the smallest, and $\sigma\big[\Delta_{r1}(\mathbf{x}_p,\,\boldsymbol{\theta})\big]$ is the term involving output $\mathcal{O}_{r1}$ and its largest competitor $\mathcal{O}_{r2}$.

We wish to show that if the *a posterioris* are ranked as shown in table 3, then the classifier output $\mathcal{O}_{r1}$ (underlined in the last column in table 3) corresponding to class ω_{r1} will be the most active of all outputs when $\mathrm{CFM}_{mono}(\mathbf{x}_p)$ is maximized.

4.1 ASYMPTOTIC PERFORMANCE OF CFM_{mono} FOR $\sigma[\Delta] = \mathbf{u}[\Delta]$

Figure 4 illustrates three different functions (normalized so that $-1 \leq \sigma[\Delta] \leq 0$) one might use to implement CFM_{mono}. The first of these functions is the Heaviside step function (denoted by $\mathrm{u}[\Delta]$). Clearly this function is an exception to the general rule stating that CFM_{mono} is a strictly increasing continuously differentiable function of Δ. This functional form is of interest for two reasons. First, it is the MLP analog of the original perceptron learning criterion (e.g., [6], pg. 141); second, it leads to a very simple determination of the maximum $\mathrm{CFM}_{mono}(\mathbf{x}_p)$ values for the classifier outputs. When this objective function is used to implement CFM_{mono} learning, one can see readily from table 3 that CFM_{mono} will be maximized if output $\mathcal{O}_{r1}$ is marginally bigger than any of its competitors. Because $\mathrm{u}'[\Delta] = 0\ \forall\ \Delta \neq 0$, there is no numerical incentive for $\mathcal{O}_{r1}$ to be made any more than marginally larger than its competitors. The relative activation of outputs $\mathcal{O}_{r2} \rightarrow \mathcal{O}_{rN}$ is irrelevant beyond their being less that $\mathcal{O}_{r1}$. Thus, the Heaviside step functional form of CFM_{mono} implements the Bayesian discriminant function — albeit marginally – for asymptotically large training sets.

4.2 ASYMPTOTIC PERFORMANCE OF CFM_{mono} FOR STRICTLY INCREASING DIFFERENTIABLE FUNCTIONS OF Δ

In practice, learning with a discontinuous $\sigma[\Delta]$ like the Heaviside step is unstable. One can achieve stable learning using strictly increasing continuously differentiable functions of Δ [8]. One can analyze the asymptotic behavior of these functions by considering their effect on a set of classifier outputs in the initial equilibrium state for which all outputs are equal. If we define σ_e as the value of the CFM_{mono} function $\sigma[\Delta]$ when its argument $\Delta = 0$, and σ'_e as the derivative of the CFM_{mono} function at this same point (see inset in Figure 4), we can observe how the outputs will be perturbed from the equilibrium point as CFM_{mono} is maximized. A differential positive change in the value of $\mathcal{O}_{r1}$ results in a change to the over-all CFM_{mono} of

$$\frac{d\mathrm{CFM}_{mono}(\overline{\Delta} = 0)}{d\mathcal{O}_{r1}\uparrow} = \sigma'_e \cdot \mathrm{P}(\omega_{r1}\,|\,\mathbf{x}_p) - \sigma'_e \cdot \mathrm{P}(\overline{\omega_{r1}}\,|\,\mathbf{x}_p) \quad (46)$$

$$> 0 \ \textit{iff}\ \mathrm{P}(\omega_{r1}\,|\,\mathbf{x}_p) > \mathrm{P}(\overline{\omega_{r1}}\,|\,\mathbf{x}_p) \quad \text{or}\ \ \mathrm{P}(\omega_{r1}\,|\,\mathbf{x}_p) > 0.5$$

A differential negative change in the value of $\mathcal{O}_{r1}$ results in a re-ordering of the output rankings; $\mathcal{O}_{r1}$ becomes the *least* active output (all the other outputs remain unchanged), so all the terms in the right-most column of table 3 are

altered to reflect this change in the identity of the most active competitor (refer back to (42)), and the net change in CFM_{mono} is given by

$$\frac{d\text{CFM}_{mono}\,(\overline{\Delta} = 0)}{d\mathcal{O}_{r1}\downarrow} \;=\; -\sigma_e' \cdot \text{P}(\omega_{r1} \mid \mathbf{x}_p) \;<\; 0 \quad (47)$$

One can show that altering any of the outputs $\mathcal{O}_{r2}, \mathcal{O}_{r3}, \ldots \mathcal{O}_{rN}$ independent of any alteration to $\mathcal{O}_{r1}$ from the equilibrium point always results in a net reduction in CFM_{mono}

$$\frac{d\text{CFM}_{mono}\,(\overline{\Delta} = 0)}{d\mathcal{O}_{rj \neq r1}\uparrow} = -\sigma_e' \cdot \text{P}(\omega_{rj} \mid \mathbf{x}_p) + \sigma_e' \cdot \text{P}(\overline{\omega_{rj}} \mid \mathbf{x}_p) \;<\; 0 \quad (48)$$

$$\frac{d\text{CFM}_{mono}\,(\overline{\Delta} = 0)}{d\mathcal{O}_{rj \neq r1}\downarrow} \;=\; -\sigma_e' \cdot \text{P}(\omega_{rj} \mid \mathbf{x}_p) \;<\; 0 \quad (49)$$

Equations (46) – (49) do not indicate what the optimum values for $\mathcal{O}_{r1} \rightarrow \mathcal{O}_{rN}$ are when $\text{P}(\omega_{r1} \mid \mathbf{x}_p) > 0.5$. These optimal values depend on the specific functional form of $\sigma[\Delta]$. When $\sigma[\Delta]$ is a linear function of Δ, the optimal values of the outputs are $\mathcal{O}_{r1} = 1$ and $\mathcal{O}_{rj \neq r1} = 0$. Non-linear functional forms (such as the "maximally flat" one shown in Figure 4) tend to produce non-binary outputs $\mathcal{O}_{r1} < 1$ and $\mathcal{O}_{rj \neq r1} > 0$.

Equations (46) – (49) show that the equilibrium point yields sub-optimal CFM_{mono} only if the largest *a posteriori* is greater than 0.5. Since the class boundaries for an N-class problem are defined as the connected set of points on $\mathbf{x}$ at which all non-zero *a posterioris* are equal, continuously differentiable CFM_{mono} functions will (in theory) fail to form decision boundaries in regions on $\mathbf{x}$ where there are more than two non-zero *a posterioris* for asymptotically large training sets. Moreover, these equations indicate that continuously differentiable CFM_{mono} functions will set all classifier outputs equal in all regions on $\mathbf{x}$ where no *a posteriori* exceeds 0.5 — conceivably a large fraction of the domain of $\mathbf{x}$ when the number of classes N is large. While these asymptotic limitations would seem to render the CFM_{mono} class of objective functions useless for classification, in practice the functions compare quite favorably with reasonable error measures. Equation (44) may provide some insight into this apparent inconsistency.

Using the definition of correlation (denoted by ϕ), we can define the correlation between the ith CFM_{mono} term and its corresponding *a posteriori* by

$$\phi_i(\boldsymbol{\theta}) = \Big(\text{E}\left[\text{P}(\omega_i \mid \mathbf{x}) \cdot \sigma\left[\Delta_i(\mathbf{x}, \boldsymbol{\theta})\right]\right] - \text{E}\left[\text{P}(\omega_i \mid \mathbf{x})\right]\ \text{E}\left[\sigma\left[\Delta_i(\mathbf{x}, \boldsymbol{\theta})\right]\right]\Big) \cdot \Big(\text{Var}\left[\text{P}(\omega_i \mid \mathbf{x})\right] \cdot \text{Var}\left[\sigma\left[\Delta_i(\mathbf{x}, \boldsymbol{\theta})\right]\right]\Big)^{-1/2} \quad (50)$$

As a result, it is possible to express (44) as

$$\text{CFM}_{mono} \asymp \sum_{i=1}^{N} \left\{ \sqrt{\text{Var}\left[\text{P}(\omega_i \mid \mathbf{x})\right]} \cdot \phi_i(\boldsymbol{\theta}) \cdot \sqrt{\text{Var}\left[\sigma\left[\Delta_i(\mathbf{x}, \boldsymbol{\theta})\right]\right]} + \text{P}(\omega_i) \cdot \text{E}\left[\sigma\left[\Delta_i(\mathbf{x}, \boldsymbol{\theta})\right]\right] \right\} \quad (51)$$

Since the terms $\sqrt{\text{Var}[\text{P}(\omega_i \mid \mathbf{x})]}$ and $\text{P}(\omega_i)$ are not functions of the classifier's parameters $\boldsymbol{\theta}$, they are constants vis-a-vis optimizing CFM_{mono} in (51). Thus, maximizing CFM_{mono} tends to maximize the correlation between $\text{P}(\omega_i \mid \mathbf{x})$ and $\sigma\left[\Delta_i(\mathbf{x}, \boldsymbol{\theta})\right]$, along with the expectation and variance of $\sigma\left[\Delta_i(\mathbf{x}, \boldsymbol{\theta})\right]$ for each class ω_i over the entire domain of $\mathbf{x}$. In fact empirical studies bear this out. Classifiers trained with CFM_{mono} objective functions and relatively small sets of statistically independent samples tend to yield outputs with higher variance than their reasonable error measure counterparts; CFM_{mono}-trained classifiers also tend to yield multiple outputs with high activations when uncertain as to the class of a test sample. There is strong correlation between $\text{P}(\omega_i \mid \mathbf{x})$ and $\sigma\left[\Delta_i(\mathbf{x}, \boldsymbol{\theta})\right]$ as indicated by the CFM_{mono} classifiers' median error rate of 1.5% on a speaker-dependent /b,d,g/ phoneme recognition task [8].

5 COMMENTS ON THE APPLICABILITY OF THESE PROOFS TO THE STUDY OF GENERALIZATION IN MLP CLASSIFIERS

When one sets out to classify an RV $\mathbf{x}$, the total number of statistically independent training samples n_t and the functional capacity (denoted by $\boldsymbol{\theta}$; see section 3.1.2) of one's classifier are two factors that will determine in large part the classifier's ultimate performance. We reproduce expressions for the MSE and information theoretic reasonable error measures and the analogous expression for the CFM_{mono} objective functions in order to illustrate the importance of these factors. Probabilities that rely on the asymptotic statistics of the training data are now presented as *estimates* (denoted by brackets "$\langle\rangle$") of the true underlying probabilities:

$$MSE \asymp$$

$$\sum_{i=1}^{N} \left\{ \int_{\mathbf{X}} \left[\mathcal{O}_i(\mathbf{x},\, \boldsymbol{\theta}) \;-\; \underbrace{\langle \mathrm{P}(\omega_i \,|\, \mathbf{x}) \rangle}_{g_i(\mathbf{X})} \right]^2 \langle \rho(\mathbf{x}) \rangle \, d\mathbf{x} \right.$$
$$- \int_{\mathbf{X}} \langle \underbrace{(\mathrm{P}(\omega_i \,|\, \mathbf{x}))^2}_{g_i(\mathbf{X})} \rangle \, \langle \rho(\mathbf{x}) \rangle \, d\mathbf{x}$$
$$\left. + \; \langle \mathrm{P}(\omega_i) \rangle \right\} \tag{52}$$

$$CE \;\asymp$$
$$-\sum_{i=1}^{N} \int_{\mathbf{X}} \left[\underbrace{\langle \mathrm{P}(\omega_i \,|\, \mathbf{x}) \rangle}_{g_i(\mathbf{X})} \, \log \{ \mathcal{O}_i(\mathbf{x},\, \boldsymbol{\theta}) \} \right.$$
$$\left. + \left(1 \;-\; \underbrace{\langle \mathrm{P}(\omega_i \,|\, \mathbf{x}) \rangle}_{g_i(\mathbf{X})} \right) \log \{ 1 \;-\; \mathcal{O}_i(\mathbf{x},\, \boldsymbol{\theta}) \} \right]$$
$$\cdot \, \langle \rho(\mathbf{x}) \rangle \, d\mathbf{x} \tag{53}$$

$$\mathrm{CFM}_{mono} \;\asymp$$
$$\sum_{i=1}^{N} \int_{\mathbf{X}} \underbrace{\langle \mathrm{P}(\omega_i \,|\, \mathbf{x}) \rangle}_{g_i(\mathbf{X})} \cdot \, \sigma \left[\Delta_i(\mathbf{x},\, \boldsymbol{\theta}) \right] \, \langle \rho(\mathbf{x}) \rangle \, d\mathbf{x} \tag{54}$$

Equations (52) – (54) indicate that the optimization of all of these objective functions depends on accurate estimates of the *a posterioris* and PDF of $\mathbf{x}$ — functions of the training set itself. Optimization also depends on sufficient functional capacity in $\boldsymbol{\theta}$. Finally, optimization — and thereby approximation of Bayesian discriminant performance — relies on the behavior of each objective function given some fixed number of training samples n_t and some fixed parameterization of the classifier $\boldsymbol{\theta}$.

In simplistic terms, there are four possible circumstances one will encounter:

- $n_t \rightarrow \infty$; $\boldsymbol{\theta}$ sufficient: For the case in which one has plenty of independent training samples, one's training data will yield accurate estimates of the *a posterioris* and PDF of $\mathbf{x}$ over its entire domain. Furthermore, one's classifier will have sufficient functional capacity to model these estimates, and one will achieve Bayesian discriminant performance.
- $n_t \rightarrow \infty$; $\boldsymbol{\theta}$ insufficient: For this case, the estimates of the *a posterioris* and PDF of $\mathbf{x}$ will be accurate, but the classifier will have insufficient functional capacity to model them. As a result, the classifier will not achieve Bayesian performance.
- $n_t \ll \infty$; $\boldsymbol{\theta}$ sufficient: In practice one rarely has access to a sufficient amount of training data. In such cases the data will constitute poor estimates of the *a posterioris* and PDF of $\mathbf{x}$. If the classifier has sufficient parameterization it will learn these inaccurate probabilistic estimates and will generalize poorly on disjoint test data ([6], section 3.8).
- $n_t \ll \infty$; $\boldsymbol{\theta}$ insufficient: For the case in which one has insufficient data, it is often advantageous to train a classifier with reduced parameterization. In such cases the data will constitute poor estimates of the *a posterioris* and PDF of $\mathbf{x}$, but the classifier will not have sufficient functional capacity to learn the less representative features of these inaccurate probabilistic estimates. Instead it will have only enough capacity to learn the gross features of these poor estimates, and generalization to disjoint test data will be as good as warranted by the training data.

  This case has been studied in great detail from a number of different perspectives. PAC[4] analysis of learning using VC dimension applies a worst case analysis to the problem of learning from examples by deriving bounds on the number of exemplars needed to attain (with a desired probability) a desired accuracy. The VC dimension is defined only for concept classes that are discrete, or in connectionist parlance, for binary outputs (but potentially continuous inputs). Discrete concept classes require that the class-conditional densities of the input RV $\rho(\mathbf{x} \,|\, \omega_i)$ be non-overlapping. For this reason the PAC formalism does not apply to our situation, in which estimates of *a posteriori* probabilities, or of the "best guess classification" (where the best guess might not be very accurate) are required.

  VC PAC analysis has been applied to feedforward networks of binary threshold elements, to which we direct interested readers' attention [2].

The study of classifier generalization is typically viewed as a problem of determining the optimal parameterization for a classifier, given some fixed number of training samples. Interest in the functional form of the objective function used to train the classifier has rarely gone beyond establishing its validity as a learning metric on some information theoretic basis. But (52) – (54) clearly illustrate that different objective functions approximate the Bayesian discriminant function in markedly different ways. In this sense, given fixed n_t and $\boldsymbol{\theta}$, each objective function represents a different estimator of the Bayesian discriminant function. In detection and estimation theory, estimators are evaluated in terms of their bias and variance for finite sample sizes. Good estimators are those that are unbiased with minimal variance (as determined by the Cramér-Rao bound [5]); such estimators are characterized as "efficient". We feel that the study of generalization in MLP classifiers can be

[4] Probably Approximately Correct.

advanced by placing more emphasis on theoretical comparisons of objective functions as *estimators of the Bayesian discriminant function*. The derivations of this paper serve as a point of departure for such comparisons.

6 SUMMARY

Multi-Layer Perceptrons can be trained with two broad classes of objective functions to yield Bayesian discriminant performance. Reasonable error measures yield MLP outputs that (under conditions summarized below) asymptotically converge to the *a posteriori* probabilities associated with the input RV. Mean-squared error and information theoretic objective functions prove to be reasonable error measures. Classification Figures of Merit (CFM_{mono}) yield MLP outputs that generally reflect the identity of the maximum *a posteriori* associated with any sample of the input RV for asymptotically large training sets.

The conditions necessary for MLP Bayesian discriminant performance (given that the classifier is trained with a reasonable error measure or a CFM_{mono} objective function) are

- The class-conditional densities of the input RV must be "well-behaved" to the extent that their complexity must be bounded (section 3.1.1).
- The functional capacity of the MLP classifier, expressed in its parameterization θ, must be sufficient to model the class-conditional densities of the input RV (section 3.1.2).
- The MLP training set must contain an asymptotically large number of statistically independent training samples.

Given these results, we are inclined to view architecturally identical MLPs trained with different Bayesian objective functions as alternative *estimators* of the Bayesian discriminant function. We offer these results as a basis for evaluating MLP classifier generalization in the context of traditional detection and estimation theory.

Acknowledgments

This research was funded by grants from Bell Communications Research, ATR Interpreting Telephony Research Laboratories, and the National Science Foundation (NSF grant number EET-8716324). We wish to thank Professor B. V. K. Vijaya Kumar[5] and Dr. Dave Touretzky[6] for their critical review of this paper.

References

[1] A. Barron, "Statistical Learning Networks: A Unifying View" presented at 1988 Symposium on the Interface: Statistics and Computer Science.

[2] E. B.Baum and D. Haussler, "What Size Net Gives Valid Generalization?" *Neural Computation*, vol. 1, pp. 151-160, spring, 1989.

[3] H. Bourlard and C. Wellekens, "Links Between Markov Models and Multilayer Perceptrons," in *Advances in Neural Information Processing Systems, vol. 1*, Dave Touretzky, ed., San Diego: Morgan-Kaufmann, pp. 502-510, 1988.

[4] G. Cybenko "Approximation by Superpositions of a Sigmoidal Function," *Mathematics of Control, Signals and Systems*, vol 2, pp. 303-314, 1989.

[5] M. H. DeGroot, *Probability and Statistics, 2nd ed.*, Reading: Addison-Wesley, pp. 424-429, 1986.

[6] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, New York: John Wiley & Sons, 1973.

[7] H. Gish, "A Probabilistic Approach to the Understanding and Training of Neural Network Classifiers," in *Proceedings of the 1990 IEEE International Conference on Acoustics, Speech, and Signal Processing*, April, 1990, vol. 3, pp. 1361-1364.

[8] J. B. Hampshire II and A. H. Waibel, "A Novel Objective Function for Improved Phoneme Recognition Using Time-Delay Neural Networks", *IEEE Trans. Neural Networks*, vol. 1, pp. 216-228, June, 1990.

[9] S. J. Hanson and D. J. Burr, "Minkowski-*r* Back-Propagation: Learning in Connectionist Model with Non-Euclidean Error Signals", *Proceedings of the 1987 Neural Information Processing Conference, Am. Institute of Physics, D.*, D. Anderson, ed., pp. 348-357, 1988.

[10] G. E. Hinton, "Connectionist Learning Procedures", *Carnegie Mellon University Technical Report CMU-CS-87-115 (version 2)*, Dec. 1987, p. 14.

[11] R. Lippmann, "Pattern Classification Using Neural Networks" in *IEEE Communications Magazine*, vol. 27, No. 11, 1989.

[5] Dept. of Electrical and Computer Engineering, Carnegie Mellon University.

[6] School of Computer Science, Carnegie Mellon University.

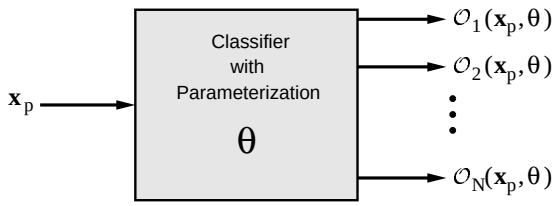

Figure 1: The general *N*-class classification problem.

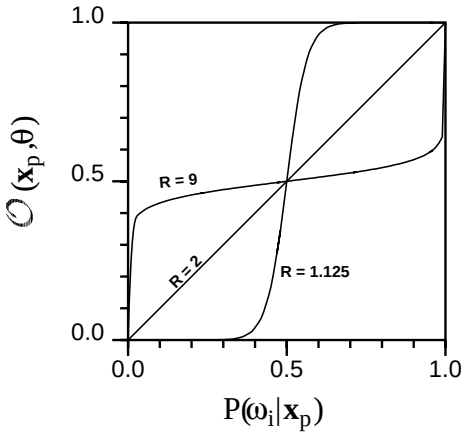

Figure 2: The L_R minimum error value of $\mathcal{O}_i(\mathbf{x}_p, \boldsymbol{\theta})$ is plotted as a function of $P(\omega_i \mid \mathbf{x}_p)$ for $R = 1.125, 2.0, 9.0$.

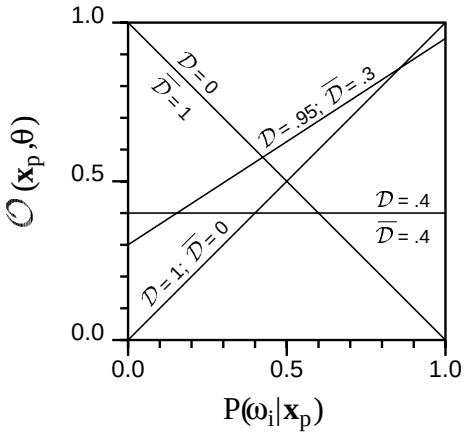

Figure 3: The minimum MSE value of $\mathcal{O}_i(\mathbf{x}_p,\ \theta)$ is plotted as a function of $P(\omega_i \mid \mathbf{x}_p)$ for various training target values.

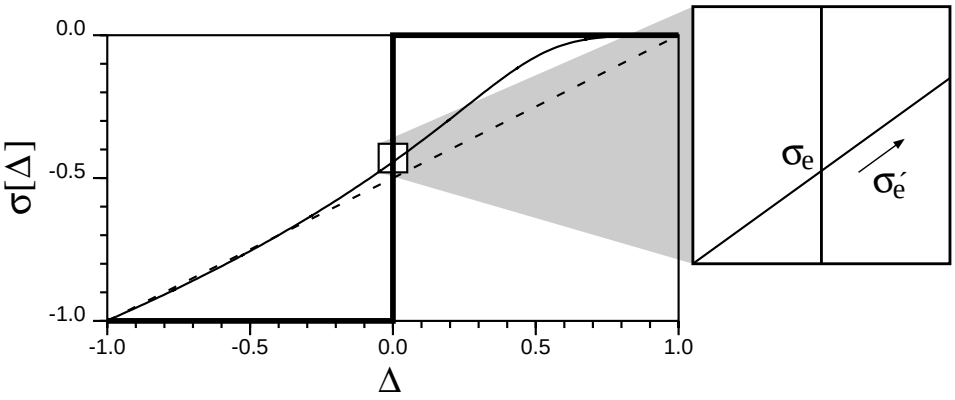

Figure 4: Three different functional implementations of the CFM$_{mono}$ learning rule: Heaviside step, linear, and "maximally flat". Inset: a CFM$_{mono}$ function in the vicinity of $\Delta = 0$.