THE ERROR-REJECT TRADEOFF

Lars Kai Hansen
CONNECT, Electronics Institute B349
The Technical University of Denmark
DK-2800 Lyngby, Denmark,
lkhansen@ei.dtu.dk

Christian Liisberg
Applied Bio Cybernetics
Overdrevsvej
DK-3390 Hundested, Denmark
abcyabc@pop.denet.dk

Peter Salamon
Dept. of Mathematical Sciences
San Diego State University
San Diego CA 92182 USA,
salamon@math.sdsu.edu

Abstract

We investigate the error versus reject tradeoff for classifiers. Our analysis is motivated by the remarkable similarity in error-reject tradeoff curves for widely differing algorithms classifying handwritten characters. We present the data in a new scaled version that makes this universal character particularly evident. Based on Chow's theory of the error-reject tradeoff and its underlying Bayesian analysis we argue that such universality is in fact to be expected for general classification problems. Furthermore, we extend Chow's theory to classifiers working from finite samples on a broad, albeit limited, class of problems. The problems we consider are *effectively binary*, i.e., classification problems for which almost all inputs involve a choice between the right classification and at most one predominant alternative. We show that for such problems at most half of the initially rejected inputs would have been erroneously classified. We show further that such problems arise naturally as small perturbations of the PAC model for large training sets. The perturbed model leads us to conclude that the dominant source of error comes from pairwise overlapping categories. For infinite training sets, the overlap is due to noise and/or poor preprocessing. For finite training sets there is an additional contribution from the inevitable displacement of the decision boundaries due to finiteness of the sample. In either case, a rejection mechanism which rejects inputs in a shell surrounding the decision boundaries leads to a universal form for the error-reject tradeoff. Finally we analyze a specific reject mechanism based on the extent of consensus among an ensemble of classifiers. For the ensemble reject mechanism we find an analytic expression for the error-reject tradeoff based on a maximum entropy estimate of the problem difficulty distribution.

Keywords: error-reject tradeoff, handwritten digits, neural networks, ensembles.

1 Introduction

Characterization of the error-reject tradeoff for neural classifiers is a problem of significant practical importance (see eg. [11]). Nevertheless, remarkably little attention has been devoted to this problem in the neural net literature. In a large scale evaluation of devices for recognition of handwritten characters, particular attention was focused on the relation between the reduction in the number of generalization errors and the number of rejected inputs [21]. The evaluation took place as part of a conference held by the U. S. National Institute of Standards and Technology (NIST) wherein training sets and test sets were provided in a competition designed to assess the performance of different classifier systems on the benchmark problem of character recognition. The systems participating in the competition employed widely different classifier schemes insofar as these schemes were revealed. Nevertheless, striking uniformity was observed among the various error-reject tradeoff graphs, the reporting of which was an integral part of the competition. In figure 1, we show examples adapted from the proceedings and from other independent experiments on digit recognition. Geist and Wilkinson [10] noted the similarity of the trade-off graphs and obtained a good fit with a three parameter phenomenological model. Our goal in the present paper is to explain these tradeoff curves, and their universality from a more theoretical perspective.

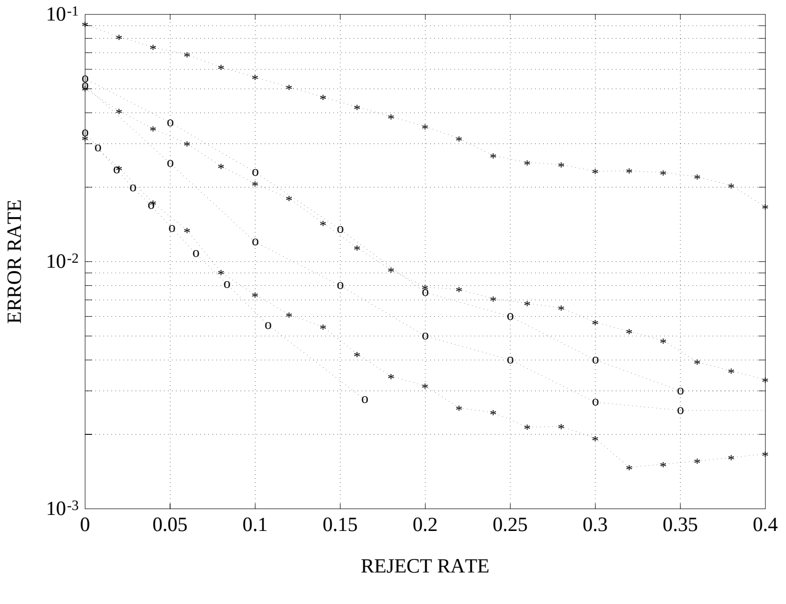

Figure 1: Error versus reject rates for different classifiers. The three $*$-marked sets of rates are three systems presented at the NIST Consensus conference.The three sets of open circles are derived from the experiments of Lee.

Intuitively, a rejection rule is based on the "degree of certainty" that the operator feels concerning a classification. Most classifier implementations come naturally equipped with a scale to estimate at least an ordinal degree of certainty. In general, however, the decision to reject a pattern can be based on a completely separate algorithm from the one which classifies, i.e. the extent of certainty represents an independent degree of freedom. Some of the implementations in the NIST competition in fact trained separate neural networks just to predict the certainty of a classification [5].

The criterion of *rationality* specifies "the right" choice of rejection rule as the one which minimizes the expected number of generalization errors [7, 6]. It follows that the rational measure of the degree of certainty is the misclassification probability, m. The rejection rule based on m in the presence of perfect information is traditionally known as the *Bayes optimal reject rule*. Provided that the natural estimate of the degree of certainty provided with a classifier is monotonic in m, such a classifier implements the Bayes optimal reject rule. In this paper we argue that well trained classifier systems operate close to the Bayes optimal limit and hence classify approximately according to the correct class distributions and reject approximately according to the misclassification probability m to the extent allowed by the finiteness of the sample. We extend the classical theory of the error-reject tradeoff due to C.K. Chow [6] to *near optimal* classifiers whose decisions are based on large, albeit finite, datasets. The extension relies on a model scenario for almost perfectly learnable problems.

The motivation for our model scenario comes from the remarkably simple phenomenology of the error-reject tradeoff in the context of handwritten digit recognition. In particular, we argue several ways that this problem is effectively a binary decision problem: a generic system either gets the digit right or chooses one predominant alternative. One of the results from this effectively binary character is that for small reject rates one generically expects a tradeoff of the form:

$$\delta E(R) \equiv E(R) - E(0) = -\frac{1}{2}R \tag{1}$$

where $E(R)$ is the error rate at a reject rate of R. The expression (1) suggests that $E(R)/E(0)$ should be plotted against $R/E(0)$ providing a "universal" linear aproximation to $E(R)$ for various classifiers. The success of this suggestion is illustrated in figure 2. It is seen that the naive scaling shows a universal error-reject tradeoff for several independent experiments and this scaling turns out to be a key ingredient for the universal error-reject curves described in section 6. The coefficient $(\frac{1}{2})$ of the reject rate on the right hand side of (1) is the fraction of the rejected patterns which would have been incorrectly classified and will be refered to as the error-reject ratio and its marginal counterpart, dE/dR as the marginal error-reject ratio. Note that this marginal error-reject ratio is the same in the scaled coordinates, i.e.

$$\frac{d(E/E_0)}{d(R/E_0)} = \frac{dE}{dR}. \tag{2}$$

In the NIST proceedings, Geist and Wilkinson [10] called for a "perfect" reject mechanism with an error-reject ratio of one, i.e.,

$$\delta E(R) \equiv -R. \tag{3}$$

Note that such a perfect mechanism rejects *only* inputs which would have been misclassified. Such efficiency is, however, not compatible on the average with the implicit assumption that the classifier is well optimized. The ideal mechanism (3) implies that we can identify a set of inputs where all decisions are wrong. In that case a better classifier could be obtained simply by letting the classification for all these inputs be random. Since the probability of the correct class coming up for any given input is the reciprocal of the number of classes n, the average error rate would be $1 - 1/n$. Hence, this modified rule would be better (at zero reject rate) than the "optimized" classifier. While this can occur for small training sets, it is an effect which must disappear as the size of the training set gets large.

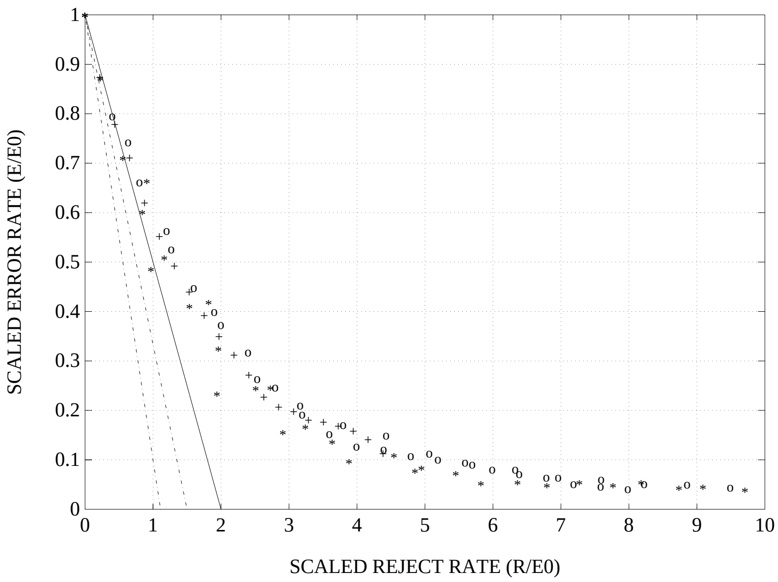

Figure 2: Scaled error-reject rates for the classifiers in figure 1. The solid line is the relation in equation (1) valid for small reject rates, the two dashed lines correspond to degrees of confusion of $n_{\text{eff}} = 3$ and $n_{\text{eff}} = 10$.

This kind of reject rule, and the bounding average error rate $1 - 1/n$ are familiar from the usual reject decision facing a student taking any one of the standardized multiple choice examinations whose scoring includes a penalty to cancel the effect of random guessing. The error-reject tradeoff in equation (1) can be interpreted in this metaphor as saying that the classifiers were able to narrow the field of possibilities to two before resorting to guessing. This illustrates a general relationship between the error-reject ratio and the *effective degree of confusion*, a measure based on consensus performance [14]. This relationship is discussed further in section 6.

A theory of the reject mechanism for the one-layer perceptron with a noiseless perceptron teacher has been developed by Parrondo and Van den Broeck [13]. It too complies with the generic first order approximation in equation (1).

A general discussion based on optimal rejection for Bayesian classifiers, including multi-class problems, was published in 1970 by Chow [6]. In the next section we review Chows results. In section four, we go on to discuss the implications of these results for classifiers trained on finite training sets for a class of problems dubbed the simplest scenario model described in section three. We then consider ensembles of classifiers in section six followed by a treatment of the ensemble reject mechanism in section seven. This leads us to a universal one parameter family of error-reject curves for effectively binary problems wherein the parameter is specified by E_0, the error rate at zero reject. For highly accurate classifiers the universal error-reject curves become independent of the proficiency. The universal character of these curves derives from two facts: errors occur primarily in areas of binary overlap between class probability distributions and rejection occurs by eliminating patterns in the vicinity of decision boundaries.

2 Chow's theory of the error-reject tradeoff

Chow's error-reject analysis is based on Bayesian decision theory [7] and consequently operates from the *ideal* class probability distributions. In this sense it may be considered as the *teacher* for the classification problem. We begin with a review of this ideal case in the present section, leaving the analysis of empirical classifiers operating from *approximate* class probability distributions estimated from a finite sample $D = \{(\mathbf{x}_k, i_k)\}$ for a later section.

Consider a classification problem with n classes $i = 1, ..., n$ which form the possible categories for input vectors $\mathbf{x}$. Complete information about the problem consists of the joint distribution $P(\mathbf{x}, i)$ which determines the conditional distribution $P(i|\mathbf{x})$, the marginal distribution $P(\mathbf{x})$ and the class distributions $P(\mathbf{x}|i)$. The ideal Bayes classifier chooses the classification

$$i(\mathbf{x}) = \arg\max_{\mathrm{i=1,..,n}} (P(i|\mathbf{x})) \tag{4}$$

for a given $\mathbf{x}$. Note that this choice corresponds to rational decisions in the sense that it minimizes the expected number of misclassifications. (See figure 3.)

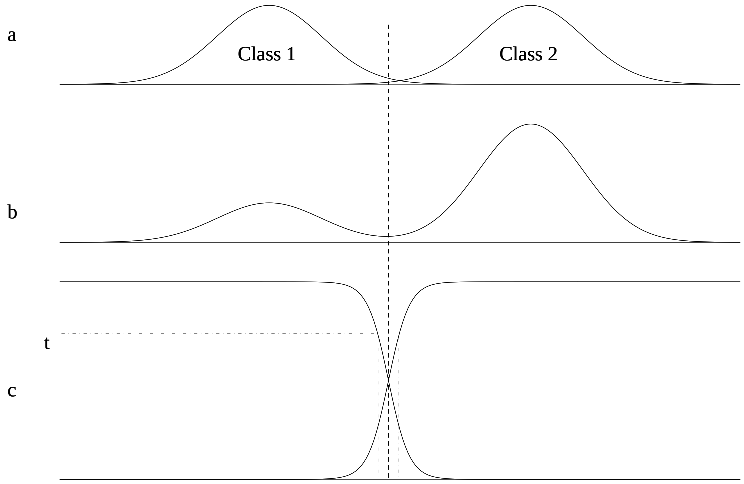

Figure 3: The figures illustrate the various distributions involved for a simple one dimensional example with two categories. (a) shows the two class distributions $P(\mathbf{x}|i)$, (b) shows the input distribution $P(\mathbf{x})$ assuming that class 2 is three times as frequent as class 1, and (c) shows the conditional distributions $P(i|\mathbf{x})$. All three curves show the region of rejected inputs for the threshold t=0.6.

The function $i(\mathbf{x})$ corresponds to the usual notion of teacher which is typically discussed in the case when the $P(i|\mathbf{x})$ take on only the values 0 and 1 almost everywhere relative to the measure defined by $P(\mathbf{x})$. Our formulation of classification is constructed with an eye towards overlapping categories, i.e. problems that are not perfectly learnable in the sense that the error rate need not vanish even with complete information about the problem. In terms of the formalism, this corresponds to the existence of regions in $\mathbf{x}$ where more than one $P(i|\mathbf{x})$ has appreciable mass. One common source of this feature in real examples comes from the presence of noise in almost any form. The fact that humans classifying the data in the NIST competition had an error rate of 2.5% implies that our approach is appropriate; perfect teachers able to achieve error free classification do not exist for handwritten digit recognition. "Noise" is present e.g. due to sampling and human rendering.

Denoting the proficiency of the Bayes classifier by

$$r(\mathbf{x}) = P(i(\mathbf{x})|\mathbf{x}) = \max_{\mathrm{i}=1,..,\mathrm{n}} (P(i|\mathbf{x})) \tag{5}$$

and the misclassification rate by

$$m(\mathbf{x}) = 1 - r(\mathbf{x}) \tag{6}$$

the zero reject error rate becomes

$$E = \int P(\mathbf{x}) m(\mathbf{x}) d\mathbf{x} \tag{7}$$

Introducing the threshold t on the misclassification rate m, we write the reject mechanism as

$$\begin{aligned} m(\mathbf{x}) &\leq t \Rightarrow \text{accept}, & (8) \\ m(\mathbf{x}) &> t \Rightarrow \text{reject}. & (9) \end{aligned}$$

With this criterion we can write the reject-rate and the error-rate in terms of the parameter t as

$$\begin{aligned} R(t) &= \int P(\mathbf{x}) H\left(m(\mathbf{x}) - t\right) d\mathbf{x} & (10) \\ E(t) &= \int P(\mathbf{x}) H\left(t - m(\mathbf{x})\right) m(\mathbf{x}) d\mathbf{x} & (11) \end{aligned}$$

where the two Heaviside functions $H(\cdot)$ are non-zero for rejected and accepted inputs respectively[1]. Based on simple properties of probability distributions, Chow proved a number of relations between these two functions. For completeness, we sketch the proofs of the following three relations in an appendix.

- $E(t)$, $R(t)$ are monotonic in t
- For differentiable rates: $dE/dR = -t$
- E is a convex function of R.

An important corollary can be derived from the second relation. The corollary bounds the fraction dE/dR of rejected inputs which would be incorrectly classified by the ideal Bayes classifier. We begin by noting that as a consequence of $\sum_i P(i|\mathbf{x}) = 1$, and $P(i|\mathbf{x}) \geq 0$, it follows that $r(\mathbf{x}) = \max_{\mathrm{i=1,..,n}}(P(i|\mathbf{x}))$ must be at least 1/n, i.e., $m(\mathbf{x}) \leq 1 - 1/n$. Decreasing t from $t = 1$, nothing is rejected until $t = 1 - 1/n$. Thus in the regime where $dR \neq 0$, $t \leq 1 - 1/n$, giving the bound

[1] $E(t)$ is here a parametric form of the function introduced ealier: E(R)

$$dE/dR = -t \quad \geq \quad -(1 - 1/n) \tag{12}$$

or, alternatively,

$$|dE/dR| \quad \leq \quad 1 - 1/n \tag{13}$$

As t deceases further, more is rejected but the additional amounts rejected contain progressively smaller fractions of patterns which would have been erroneously classified. This law of diminishing returns is expressed in Chow's third relation.

The value $t = 1/2$ seems special for several reasons even for problems with $n \neq 2$. In fact, the NIST competition data followed the relation (1), which is the $n = 2$ version of equation (13), even though the actual number of categories was much larger than 2. We believe that this is the case for most problems which can be learned to a very high accuracy. Motivated by this, we introduce a scenario which leads to such behavior.

3 A Model Scenario

As stated above, our interest is in problems with some overlap among input categories. On the other hand, such overlap should be fairly small. We now describe the following "model scenario" which we believe to be typical of problems which are learnable to a very low error rate. We construct this scenario as a perturbation of the simpler problem of a Boolean function on the space of $\mathbf{x}$'s. The asymptotic theory of Boolean functions under the rubric of the PAC model has attracted much attention [3]. Appropriate for our analysis is the more general framework described by Haussler [4], which requires only that one be able to train the problem to within a tolerance of the ideal Bayes proficiency. Since we expect that such proficiency is close to one, we consider only a slightly perturbed version of the PAC model. For the PAC problem, the teacher (in our sense) has the property that $P(i(\mathbf{x})|\mathbf{x}) = 1$ for almost all $\mathbf{x}$ relative to the measure defined by the input distribution $P(\mathbf{x})$. We envision that the space of $\mathbf{x}$'s is tiled by simple regions each of which is characterized by its value for the correct classification $i(\mathbf{x}) = i_0$. Note that this forces the regions to be disjoint. Our perturbation of this problem allows each region to be surrounded by a boundary layer of thickness δ in which $P(i_0|\mathbf{x})$ drops from unity to 0. In this scenario, $m(\mathbf{x})$ rises to about 0.5 near the decision boundary and rises to higher values only in the neighborhood of the (generically lower dimensional) intersections between different decision boundaries. To lowest order in δ, the volume of the regions without overlaps is constant, the region where two regions overlap is linear in δ and, in general, the region where k regions overlap goes as δ^{k-1}. One example of how such a scenario can arise is an error free problem in which the inputs are subjected to additive noise.

EXAMPLE 1: In a study designed to test the usefulness of consensus decisions by ensembles of neural networks [14], we introduced the following toy problem of classifying a number of regions in the 20-dimensional hypercube. The regions are defined in terms of 10 randomly chosen corners of the hypercube which are designated as "pure" patterns. The i-th pure pattern represents class i, $i = 1, ..., 10$, and samples are generated for classification by perturbing one of the pure patterns by bit inversion with a specified probability p.

Letting $\mathbf{x}_i$ represent the i-th pure pattern, the class probabilities are binomial

$$P(\mathbf{x}|i) = \frac{20!}{\Delta!(20-\Delta)!} p^{\Delta}(1-p)^{20-\Delta} \tag{14}$$

where $\Delta = \Delta(\mathbf{x}, \mathbf{x}_i)$ is the Hamming distance between $\mathbf{x}$ and $\mathbf{x}_i$.

For sizeable p, the volumes of the pure regions are too small and the thicknesses of the boundary layers are too large to match our scenario. This is further revealed by the effective degree of confusion (discussed in section 5) which was experimentally determined to be 4, 7, and 9 for p values of 0.05, 0.10, and 0.15 respectively. For p sufficiently small the effective degree of confusion becomes 2 and fits the scenario above.

The example illustrates a Bayesian view of the problem in which the input stream consists of a sequence $i_1, ..., i_t, ...$. As each i_t is received for classification, it is interpreted as a feature vector $\mathbf{x}$ according to the class distribution $P(\mathbf{x}|i_t)$. Note that the class distributions must have small overlaps with each other if the problem is to be solvable to high accuracy. In terms of handwritten digits, the tilings represent pure regions corresponding to perfect printer fonts while the noise process corresponds to the human rendering and subsequent digitization of the character.

It is convenient to divide the input space into the following two regions:

$$\text{Majority Region} = \{\mathbf{x}| \;\; r(\mathbf{x}) \geq 0.5\} \tag{15}$$

$$\text{Plurality Region} = \{\mathbf{x}| \;\; r(\mathbf{x}) < 0.5\} \tag{16}$$

where the names have been chosen by analogy to "votes" by the evidence. Note that the ideal Bayes classifier rejects all plurality regions before rejecting any majority regions. One possible reason for seeing (1) empirically is that $P(\mathbf{x})$ assigns very low probability to the plurality region[2]. Since in the model scenario described above, plurality regions are restricted to a neighborhood of a lower dimensional domain (intersections between decision boundaries), any smooth input distribution $P(\mathbf{x})$

[2]If the problem is solvable with a low error rate, such probability cannot be too large.

will assign arbitrarily small volume to plurality regions for sufficiently small values of the thickness of the boundary layers, δ. (See figure 4.)

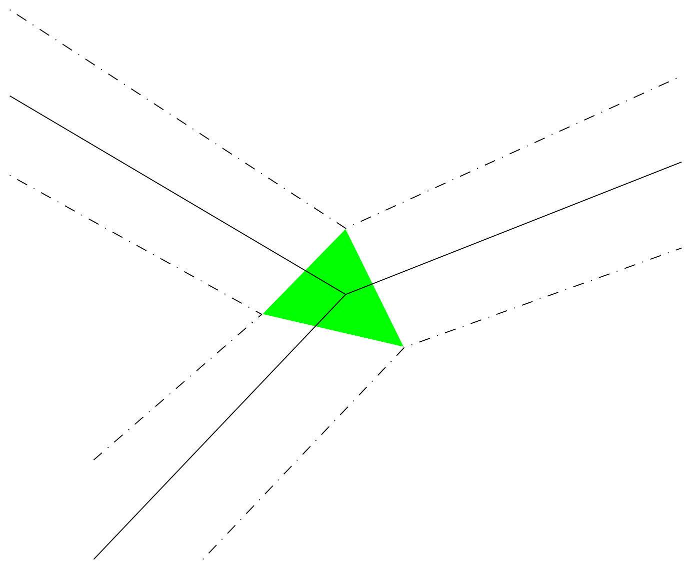

Figure 4: The figure shows a schematic two dimensional input space with the decision boundaries (solid lines) fattened to thickness δ (the region between the dash-dotted lines is rejected). The region of intersection between the fattened boundaries is shaded and shrinks to zero as δ^2.

In fact, requiring the model scenario to hold over all of the input space is overly restrictive. Certainly, for the handwritten digit recognition problem, there exist immense regions of $\mathbf{x}$ which make no sense as digits and have $P(\mathbf{x}) = 0$. It is in fact sufficient to require the simplest scenario only in the majority region along with a requirement that the probability of the plurality region be less than some tolerance ϵ.

$$\int_{\text{Plurality Region}} P(\mathbf{x}) d\mathbf{x} < \epsilon \tag{17}$$

We assume that ϵ is small enough to be negligible. This model scenario also has the property that (after the ϵ rejected for the plurality region) *the first rejected patterns sit exactly in a shell around the decision boundaries.* Since all but two of the $P(i|\mathbf{x})$ vanish for most $\mathbf{x}$ along such boundaries, $m(\mathbf{x}) = 0.5$. Thus the ideal Bayes classifier rejects initially with an error-reject tradeoff of 0.5. We have accounted in part for the NIST reject performance data. Note however that the ideal m drops to zero rapidly if the boundary layers are thin, hinting that there is more to the story. The remaining explanation must however be sought in the behavior of classifiers working from finite training sets rather than from ideal distributions.

4 Finite Training Sets

We now assume that the best information available to the classifier is the *a posteriori* distribution

$$P_{\mathrm{D}}(i|\mathbf{x}) = \mathrm{Prob}(i|\mathbf{x}, D, P_0) \tag{18}$$

where P_{D} is the estimated *a posteriori* probability distribution based on the learning set $D = \{(\mathbf{x}_k, i_k); k = 1, ...K\}$ and the prior distribution $P_0(\mathbf{x}, i)$ [2]. The optimal Bayes classifier chooses

$$i_{\mathrm{D}}(\mathbf{x}) = \arg\max_{\mathrm{i=1,..,n}} \left(P_{\mathrm{D}}(i|\mathbf{x})\right). \tag{19}$$

The corresponding misclassification rate is

$$m_{\mathrm{D}}(\mathbf{x}) = 1 - P(i_{\mathrm{D}}(\mathbf{x})|\mathbf{x}), \tag{20}$$

while the *estimated* misclassification rate is

$$\widehat{m}(\mathbf{x}) = 1 - P_{\mathrm{D}}(i_{\mathrm{D}}(\mathbf{x})|\mathbf{x}). \tag{21}$$

Thresholding on the value of $\widehat{m}$ yields the rejection and error rates

$$R(t) = \int P(\mathbf{x}) H\left(\widehat{m}(\mathbf{x}) - t\right) d\mathbf{x} \tag{22}$$

$$E(t) = \int P(\mathbf{x}) H\left(t - \widehat{m}(\mathbf{x})\right) m_{\mathrm{D}}(\mathbf{x}) d\mathbf{x}. \tag{23}$$

We now examine what happens to Chow's relations in this context. The first relation holds without any change since the more stringent one makes the threshold t, the more patterns are rejected and rejecting any patterns can only decrease the number of erroneously classified patterns. The second relation is however altered to

$$dE/dR = -\langle m_{\mathrm{D}}\rangle_{\widehat{m}=t} \tag{24}$$

where

$$< m_{\mathrm{D}} >_{\widehat{m}=t} = \frac{\int_{\{\mathbf{x}|\widehat{m}(\mathbf{x})=t\}} m_{\mathrm{D}}(\mathbf{x}) P(\mathbf{x}) d\mathbf{x}}{\int_{\{\mathbf{x}|\widehat{m}(\mathbf{x})=t\}} P(\mathbf{x}) d\mathbf{x}}. \tag{25}$$

the average value of m_{D} on the hypersurface in input space where the estimated misclassification rate equals t. (See also the discussion in the appendix.) Thus, in the modified relation, the fraction of additional rejected inputs which would have been erroneously classified as t decreases to $t - dt$ is given by the mean value of the posterior misclassification rate m_{D} over these inputs. Since the mean value in (25) need no longer be monotonic in t, some deviation from the convexity of E in R is possible although, on the average, one still expects (and sees) steadily diminishing fractions of erroneously classified patterns among the rejected inputs.

Most classifiers do not estimate $m(\mathbf{x})$ directly. In fact, such an estimate is not required since thresholding on any monotonic function of m is equivalent. As mentioned above, most implementations naturally provide some parameters whose expected values are monotonic in m. Haussler [4] shows the uniform convergence of many loss estimates to the ideal case as the size K of the dataset approaches infinity. We discuss some examples of such measures in the following sections.

We now assume that the problem has the structure of the simplest scenario described in the previous section. We further assume that the classifiers working from the finite sample D also end up with a classification scheme which follows this scenario, albeit with less narrow and less accurately placed boundary regions. In this context we note that a finite sample will not locate such boundaries precisely even with perfectly crisp categories, i.e. even for PAC problems. We therefore expect that such boundaries are displaced slightly relative to the tiling defined by the ideal Bayes classifier.

The effect of such misplaced decision boundaries on the error-reject tradeoff provides the explanation for the NIST results. Assuming that the reject rule behaves like the ideal and once again ignoring the plurality region, the first patterns rejected will lie in the immediate neighborhood of the decision boundaries (see figure 5a). Fattening boundaries which are very well located gives an error-reject tradeoff of 0.5 by the argument in the previous section. Less well located boundaries have patterns on one side which are correctly classified and patterns on the other side which are incorrectly classified. Fattening these boundaries again gives an initial error-reject ratio of 0.5 which persists until the fattened imperfectly placed boundary becomes fat enough to reach the real boundary at which point it becomes even smaller (see figure 5b). We believe that this mechanism is the dominant one responsible for the observed tradeoff in the NIST competition. Further corroboration of this "effectively" binary character of the problem is discussed in the section on ensemble performance.

We remark that rejecting patterns from any region $\mathbf{x}$ where the problem is effectively binary[3] by fattening a decision boundary always leads to an error-reject tradeoff of 0.5 to first order (figure 5c). Letting r and $1 - r$ be the probabilities of the two classes, we see that fattening a decision boundary correctly rejects r and incorrectly rejects $1 - r$ on one side, while correctly rejecting $1 - r$ and incorrectly rejecting r

[3]Formally, effectively binary refers to regions where at most two classes have appreciable probabilities.

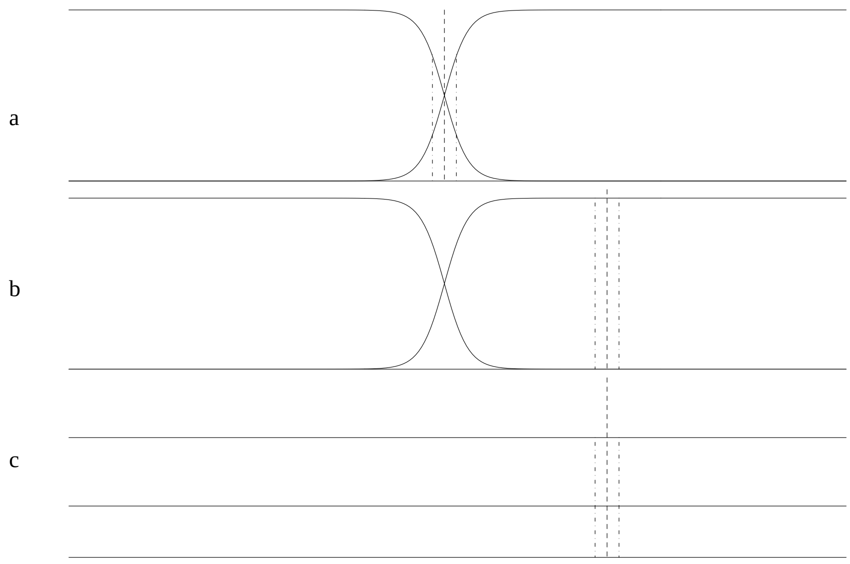

Figure 5: The figures illustrate the initial error-reject tradeoff for various positions of the decision boundary. (a) shows the ideal Bayes position and (b) shows a decision boundary misplaced slightly relative to the Bayes decision. (c) shows that for constant class probabilities the error-reject tradeoff is always 1/2 for binary decisions.

on the other side.

We conclude this section with an example which sheds light on the performance and consequent error-reject tradeoff on problems where the simplest scenario does not apply. This can arise for example from poor preprocessing which gives rise to nontrivial overlap between the class distributions. The following example deals with such a region abstracted to a single point $\mathbf{x}_0$.

<u>EXAMPLE 2:</u> Consider the toy problem with the input space a single point $\{\mathbf{x}_0\}$. The teacher distribution $P(\mathbf{x}_0, i)$ is multinomial and the best guess $i_{\mathrm{D}}(\mathbf{x}_0)$ is just the most frequent category observed. Thus for $K = |D|$ observations, the frequencies $k_1, ...k_n$ occur in D with probability

$$\frac{K!}{\prod_{i=1}^{n} k_i!} \prod_{i=1}^{n} P(\mathbf{x}_0, i)^{k_i} \tag{26}$$

and so the probability α that the right classification is chosen is

$$\alpha = Prob(i_{\mathrm{D}}(\mathbf{x}_0) = i(\mathbf{x}_0)) = \sum_{k_{i(\mathbf{x}_0)} > k_j,\ j \neq k_{i(\mathbf{x}_0)}} \frac{K!}{\prod_{i=1}^{n} k_i!} \prod_{i=1}^{n} P(\mathbf{x}_0, i)^{k_i} \tag{27}$$

To simplify the illustration further, let us consider a binary classification with $i(\mathbf{x}_0) = 1$ and $r = P(\mathbf{x}_0, 1)$. The probability α based on a sample of size K is just the probability that k_1 is greater than $K/2$.

$$\alpha = Prob(i_{\mathrm{D}}(\mathbf{x}_0) = 1) = \sum_{k_1 = \lceil K/2 \rceil}^{K} \frac{K!}{k_1!(K-k_1)!} r^{k_1} (1-r)^{K-k_1} \tag{28}$$

This probability is 0.5 if $K = 0$, r if $K = 1$ and, provided r is appreciably different from 0.5, it converges rapidly to one as K becomes large.

The ideal misclassification probability $m(\mathbf{x}_0) = 1 - r$. The misclassification probability based on the sample, $m_{\mathrm{D}}(\mathbf{x}_0)$, is $1 - r$ if $i_{\mathrm{D}}(\mathbf{x}_0) = 1$, i.e. with the probability α or r with probability $1 - \alpha$. Thus

$$< m_{\mathrm{D}}(\mathbf{x}_0) >= \alpha(1-r) + (1-\alpha)r \tag{29}$$

Note that a suboptimal choice of the classification, i.e. $i_{\mathrm{D}}(\mathbf{x}_0) \neq 1$, is more likely for a fixed K the closer r is to 0.5. In this case however, r must be close to $1 - r$ and so either value of m_{D} is close to m [7].

The estimated misclassification probability $\widehat{m}(\mathbf{x}_0)$ takes on possible values $j/K, j = 0, 1, ...K$ with probabilities

$$Prob(\widehat{m}(\mathbf{x}_0) = j/K) = \frac{K!}{j!(K-j)!} r^{K-j}(1-r)^j \tag{30}$$

Thus $\widehat{m}(\mathbf{x}_0)$ is binomially distributed with mean $1 - r = m$. Similar conclusions can be made concerning the multinomial version of the example albeit with much more technical effort. The formalism is very similar to the formalism for prediction of ensemble performance for which both the binomial and the multinomial case are discussed in reference [14]. Our discussion of this example continues as example 4 below.

Returning to the case of a general input space, we would generically expect to have few if any samples at any specific $\mathbf{x}_0$ and would expect the classifier to choose $i_{\mathrm{D}}(\mathbf{x}_0)$ by generalizing from other $\mathbf{x}$ values. Insofar as this approximates sampling the distribution $P(i|\mathbf{x})$, the reasoning in the example applies. This corresponds to considering our information from D and P_0 regarding $i(\mathbf{x}_0)$ as equivalent to information obtained from a direct sample at $\mathbf{x}_0$ of a certain size K. Such "voting" for the correct classification at $\mathbf{x}_0$ by the data can be made precise for PAC problems in the sense of Denker et al. [1] wherein each new data point eliminates classifiers whose outputs are inconsistent with the point. For some classifiers, e.g.the k-nearest neighbors algorithm [7], only the k data points nearest to $\mathbf{x}_0$ get a vote [4]. For others, such as feedforward neural networks, some datapoints have more "votes" in deciding $i_{\mathrm{D}}(\mathbf{x}_0)$ although simple nearness in input space is not the criterion. For these the relevant "nearness" is measured in the hidden representations. It is not our goal in the present paper however to probe the exact mechanism whereby the evidence in a dataset is translated into a classification by different algorithms. We merely present Example 2 as an instructive caricature of such a mechanism.

Example 3 below illustrates how the natural reject rules associated with feedforward neural networks implement a finite sample version of Bayes classification insofar as it functions by fattening boundaries.

<u>EXAMPLE 3:</u> Neural classifiers

This example interprets the results of this section in the context of feedforward neural networks. For an artificial neural network trained by the usual least squares procedure, e.g., standard backprop, it is known that the network outputs asymptotically (large networks and large training sets) approximate the ideal teacher probabilities [12]

$$y_{\mathrm{i}}(\mathbf{x}) \sim P(i|\mathbf{x}) \tag{31}$$

[4] One of the contestants in the NIST competition (ATT1) in fact employed a version of the k-nearest neighbors algorithm. This algorithm can be shown to converge to the ideal Bayes classifier as the number of neighbors used approaches infinity [7].

where y_i are the output units coding for classes $i = 1, .., n$ respectively. To enforce a decision from the network, these output units are compared and the largest value is used in a "winner takes all" decision. Hence if the network is trained optimally, it implements a Bayesian classifier. In real world applications with finite training sets, the output units are only able to implement the posterior probabilities approximately and the results in the previous section apply.

Le Cun et al. [11] used two reject mechanisms in their seminal work on neural handwritten digit recognition. The first is a mechanism corresponding to the one discussed here,

$$\begin{aligned} \mu(\mathbf{x}) &\geq 1 - t \Rightarrow \text{accept} & (32) \\ \mu(\mathbf{x}) &< 1 - t \Rightarrow \text{reject}, & (33) \end{aligned}$$

where $\mu(\mathbf{x})$ is the output of the maximally activated output unit. Insofar as the expected value of μ tends asymptotically to $r(\mathbf{x})$ by equation (31), thresholding on μ is equivalent to thresholding on $\widehat{m}$. The second mechanism used by Le Cun et al. is a threshold on the difference between $\mu(\mathbf{x})$ and the output of the "runner up" output unit. While for binary classification with perfect data the two thresholds are redundant, they give independent measures of "degree of confidence" for finite data sets even in the binary case and even for perfect data in the m-ary case. To interpret this second threshold, we note that for perfect information it begins rejecting around the region

$$P(i(\mathbf{x})|\mathbf{x}) = P(i_2(\mathbf{x})|\mathbf{x}) \tag{34}$$

where

$$i_2(\mathbf{x}) = \arg \max_{\mathrm{i=1,..,n,\ i \neq i(\mathbf{x})}} (P(i|\mathbf{x})) \tag{35}$$

Note further that this is exactly a decision boundary (albeit possibly a degenerate case of one). Furthermore, given equation (34) we are either in a plurality region or a binary region. Insofar as it works by fattening boundaries, this rejection mechanism also fits our discussion above.

5 Ensembles of Classifiers

We next examine the performance of an ensemble of classifiers employing a consensus scheme. Liisberg [9], used such a voting ensemble of lookup table networks in the NIST competition. We will see that there is a striking similarity between counting votes in favor of a certain classification whether such votes be cast by the evidence embodied in a training set or by the members of an ensemble of classifiers. While our present interest is the ensemble reject mechanism, we begin with a review of basic results concerning ensemble performance.

Collective decisions arrived at by voting can be traced back to antiquity [8]. Ensembles of classifiers were introduced for neural networks [14] as a way to eliminate some of the generalization errors based on finite training sets. The efficacy of the method is explained by the following argument. Most classifier systems share the feature that the solution space is highly degenerate. The *post training* distribution of classifiers trained on different training sets chosen according to $P(\mathbf{x})$ will be spread out over a multitude of nearly equivalent solutions. The ensemble is a particular sample from the set of these solutions. The basic idea of the ensemble approach is to eliminate some of the generalization errors using the differentiation within the realized solutions to the learning problem. The variability of the errors made by the members of the ensemble has shown that the *consensus* improves significantly on the performance of the best individual in the ensemble[5].

In [17], we used the digit recognition problem to illustrate how the consensus of an ensemble of lookup networks may outperform individual networks. We found that the ensemble consensus outperformed the best individual of the ensemble by $20-25\%$. However, due to correlation among errors made by the participating networks, the marginal benefit obtained by increasing the ensemble size was low once this size reached about 15 networks.

In [14], a device was introduced to model the dominant cause of correlation. The model is built on the assumption that correlation of erroneous classification on an input $\mathbf{x}$ is caused by the difficulty of $\mathbf{x}$; most classifiers will get the right answer on "easy" inputs while many classifiers will make mistakes on "difficult" inputs. Within the model, the difficulty of an input $\mathbf{x}$ is defined as $\theta(\mathbf{x}, K)$, the fraction of classifiers that erroneously classify $\mathbf{x}$. θ is computed with an ensemble of networks in the limit that the size of the ensemble tends to infinity. Furthermore, the members of the ensemble are each trained on independently chosen training sets of K samples selected according to $P(\mathbf{x})$. Finally, note that the difficulty is defined on inputs and so the fraction must be averaged over different instances of the input $\mathbf{x}$, i.e.

[5]While this is certainly true for the situation described here wherein the members of the ensemble see different training patterns, the analysis in [14] shows that it can be an effective technique even when all the classifiers were trained using the same training set. In the latter case, the stochastic ingredient in the algorithm typically comes from the choice of the initial values of the classifier parameters.

with the distribution $P(i|\mathbf{x})$. For $K = \infty$, all the classifiers will vote for the Bayes classification $i(\mathbf{x})$ and the error rate for a given $\mathbf{x}$ is just the fraction of the time that the input $\mathbf{x}$ corresponds to a classification other than $i(\mathbf{x})$. This shows that

$$\theta(\mathbf{x}, \infty) = m(\mathbf{x}). \tag{36}$$

For finite K, the relation between m and θ is more complicated albeit still monotonic. This is illustrated in the following continuation of Example 2.

<u>EXAMPLE 4</u>

Consider once again our toy example in which the input space consists of the single point $\{\mathbf{x}_0\}$. We once again restrict ourselves to binary classification with category 1 as the ideal Bayes response which is correct a fraction $r = 1 - m > 0.5$ of the time that input $\mathbf{x}_0$ is seen.

Now consider an ensemble of classifiers each of which sees K samples and decides that $i_{\mathrm{D}}(\mathbf{x}_0) = 1$ with probability α given by

$$\alpha(K, m) = Prob(i_{\mathrm{D}}(\mathbf{x}_0) = 1) = \sum_{k_1 = \lceil K/2 \rceil}^{K} \frac{K!}{k_1!(K - k_1)!}(1 - m)^{k_1} m^{K - k_1} \tag{37}$$

(c.f. equation (28)). The fraction of erroneous classifications on many trials of $\mathbf{x}_0$ gives

$$\theta(\mathbf{x}_0, K) = \alpha(K, m)m + (1 - \alpha(K, m))(1 - m) = \langle m_{\mathrm{D}}(\mathbf{x}_0) \rangle. \tag{38}$$

The consensus among N classifiers makes the ideal Bayes choice with probability

$$\beta(N, K, m) = \sum_{n_1 = \lceil N/2 \rceil}^{N} \frac{N!}{n_1!(N - n_1)!} \alpha(K, m)^{n_1} (1 - \alpha(K, m))^{N - n_1} \tag{39}$$

and thus will have an error rate of

$$E = \beta m + (1 - \beta)(1 - m). \tag{40}$$

We conclude the example with several observations. First, note the striking similarity between the expressions for α and β in equations (37) and (39). Second, note that as $K \to \infty$, $\alpha \to 1$ and thus $\theta(\mathbf{x}_0, K) \to m(\mathbf{x}_0)$ as argued generally above. Finally, we remark that sharing KN samples at $\mathbf{x}_0$ among N networks actually hurts performance slightly in this trivial example.

Returning to our general discussion of ensemble decisions, we next introduce the distribution of problem difficulty

$$\mu(\theta) = \int \delta(\theta(\mathbf{x}) - \theta) P(\mathbf{x}) d\mathbf{x}. \tag{41}$$

Using $\mu(\cdot)$ and the approximation that the networks perform independently on a problem with difficulty θ enables us to predict the error rate of a consensus decision. For example, an ensemble of three classifiers will have the error rate

$$E = \int_0^1 (\theta^3 + 3\theta^2(1-\theta))\mu(\theta) d\theta. \tag{42}$$

The prediction of ensemble performance agrees well with experiment [14]. To predict the improvement for more than three networks on n-fold classification problems, we need more information concerning the tendency of networks to pick the same wrong answer. To avoid considering the (many parameter) details of response probabilities over the n classes, we follow reference [14] in introducing the effective degree of confusion, n_{eff}. n_{eff} is estimated using a model in which classifiers which have the wrong classification choose with equal probabilities from among $n_{\text{eff}} - 1$ equally likely alternatives. Note that modeling the n-fold classification as an n_{eff}-fold classification modifies Chow's result (12) to

$$dE/dR \geq \; -(1 - 1/n_{\text{eff}}) \tag{43}$$

Using the effective degree of confusion it is possible to predict the error versus ensemble size[6]. How consensus performance improves with the size N of the ensemble can reveal a great deal. By comparing the predicted and the observed error rates as a function of the ensemble size, we are able to estimate the effective degree of confusion n_{eff}. The fact that this number turned out to be two for the ensembles of lookup networks on the NIST data is independent corroboration for the effectively binary character of the digit recognition problem. We confirmed this by explicit examination of the performance on the NIST data: for misclassified digits, there is on the average only one dominant alternative considered[17]. Using $n_{\text{eff}} = 2$ in equation (43) above completes this line of argument leading to equation (1).

The above predictors of ensemble performance require knowledge of the problem difficulty distribution, $\mu(\theta)$. For a finite ensemble of size N, the experimental difficulty, $\hat{\theta}$, takes discrete values: $\hat{\theta}(\mathbf{x}) = 0, 1/N, 2/N, ..., (N-1)/N, 1$. The empirical difficulty distribution, $\hat{\mu}(\hat{\theta})$, is then concentrated at these $N+1$ values.

An often useful estimate of $\hat{\mu}(\cdot)$ can be obtained in a robust fashion by choosing the *maximum entropy* $\hat{\mu}(\cdot)$ consistent with a given mean performance p of each

[6] See reference [14], equations (8) and (11).

classifier[14, 15]. Following the standard procedure [19] it is found that the distribution for an ensemble of N devices is a simple discrete exponential:

$$\mu_{\lambda,N}\left(\frac{j}{N}\right) = A_{\lambda,N}\exp\left(-\lambda\frac{j}{N}\right) \qquad j = 0, ..., N \tag{44}$$

with the normalization given by

$$A^{-1}_{\lambda,N} = \sum_{j=0}^{N}\exp\left(-\lambda\frac{j}{N}\right) = \frac{1-e^{-(N+1)\lambda/N}}{1-e^{-\lambda/N}} \tag{45}$$

and with the proviso that λ has to be adjusted so that the distribution corresponds to the correct mean individual performance as in reference [14] or ensemble performance as illustrated in the next section.

In the infinite ensemble limit, i.e. as $N \to \infty$, μ becomes

$$\mu_{\lambda}(\theta) = \left(\frac{\lambda}{1-e^{-\lambda}}\right)\exp(-\lambda\theta) \tag{46}$$

In [14], good correspondence was found between actual measured data and the proposed simple model. We use this maximum entropy estimate of μ in the following section to give a universal error-reject curve for low error rate classifiers on effectively binary problems.

6 The Ensemble Reject Mechanism

For a system using consensus decisions the natural reject mechanism is based on the extent of consensus. This is exactly what defines decision boundaries and thus thresholding on the extent of consensus fattens such boundaries. Given N classifiers indexed by $j = 1, ...N$, denote the classification of the j-th classifier by $i^{(j)}(\mathbf{x})$. Letting $v(i|\mathbf{x})$ be the number of votes for category i given $\mathbf{x}$, i.e.,

$$v(i|\mathbf{x}) = |\{j : i^{(j)}(\mathbf{x}) = i\}|, \tag{47}$$

the consensus decision chooses

$$i_{\mathrm{D}}(\mathbf{x}) = \arg\max_{\mathrm{i=1,..,n}} v(i|\mathbf{x}). \tag{48}$$

The extent of consensus on an input $\mathbf{x}$ is then

$$\gamma(\mathbf{x}) = v(i_{\mathrm{D}}(\mathbf{x})|\mathbf{x})/N. \tag{49}$$

We now argue that thresholding on γ is a practical, finite K approximation to the ideal Bayes rejection rule. Recall that this rule calls for a threshold on the error probability $m(\mathbf{x})$. As argued above, this is equivalent to thresholding on any monotonic function of m and θ is such a function. θ itself is not directly observable and we have to content ourselves with the empirical estimate $\widehat{\theta}$. Even $\widehat{\theta}$ however is only observable on labeled inputs. For unlabeled inputs, we have only the extent of consensus γ. While γ does not equal $\widehat{\theta}$ in general, for effectively binary problems the expected value of θ is monotonic in γ. In light of the arguments above and for the sake of convenience, we restrict ourselves to binary classification in which case $\widehat{\theta} = \gamma$ or $\widehat{\theta} = 1-\gamma$. Note that $\gamma > 0.5$ while $m \leq 0.5$ for effectively binary problems. To see that $\langle\theta\rangle$ is monotonic in γ, let $Prob(\theta = 1-\gamma) = p_1$ be the probability that the consensus chooses the correct answer. Then $Prob(\theta = \gamma) = 1 - p_1$ and $\langle\theta\rangle = (1-\gamma)p_1 + \gamma(1-p_1) = 1 + \gamma(1-2p_1)$. Provided that the consensus choosing the right answer is more likely than vice versa, $p_1 > 1/2$, thus $1 - 2p_1 > 0$. Note that the value of p_1 can be written in terms of the difficulty distribution μ as,

$$p_1 = \frac{\mu(1-\gamma)}{\mu(\gamma) + \mu(1-\gamma)} \tag{50}$$

and that a maximum entropy estimate of $\mu(\cdot)$ implies $p > 1/2$ for $\lambda > 0$. While for finite datasets there exist patterns with $\theta > 1/2$, there is no practical way to identify such patterns and we must content ourselves with the largest available $\langle\theta\rangle$. Rejecting patterns which have $t \geq 1-\gamma$ rejects the patterns with $\widehat{\theta}$ in the interval $[t, 1-t]$. Decreasing t from 1 as before, no patterns are rejected until t reaches $1/2$. For calculational convenience, we assume that the ensemble is large. For $t = 1/2$, we have two counts of votes: the errors (E) and the correct decisions ($C = 1 - E$). In terms of the difficulty distribution the corresponding rates are given by:

$$E_0 = E(1/2) = \int_{1/2}^{1} \mu(\theta)\mathrm{d}\theta \tag{51}$$

$$C(1/2) = \int_{0}^{1/2} \mu(\theta)\mathrm{d}\theta \tag{52}$$

For $t < 1/2$, the rates of accepted errors and correct decisions are given by:

$$E(t) = \int_{1-t}^{1} \mu(\theta)\mathrm{d}\theta \tag{53}$$

$$C(\tau) = \int_{0}^{t} \mu(\theta)\mathrm{d}\theta \tag{54}$$

while the number of rejected inputs is given by

$$R(t) = 1 - (E(t) + C(t)) = \int_t^{1-t} \mu(\theta) d\theta. \tag{55}$$

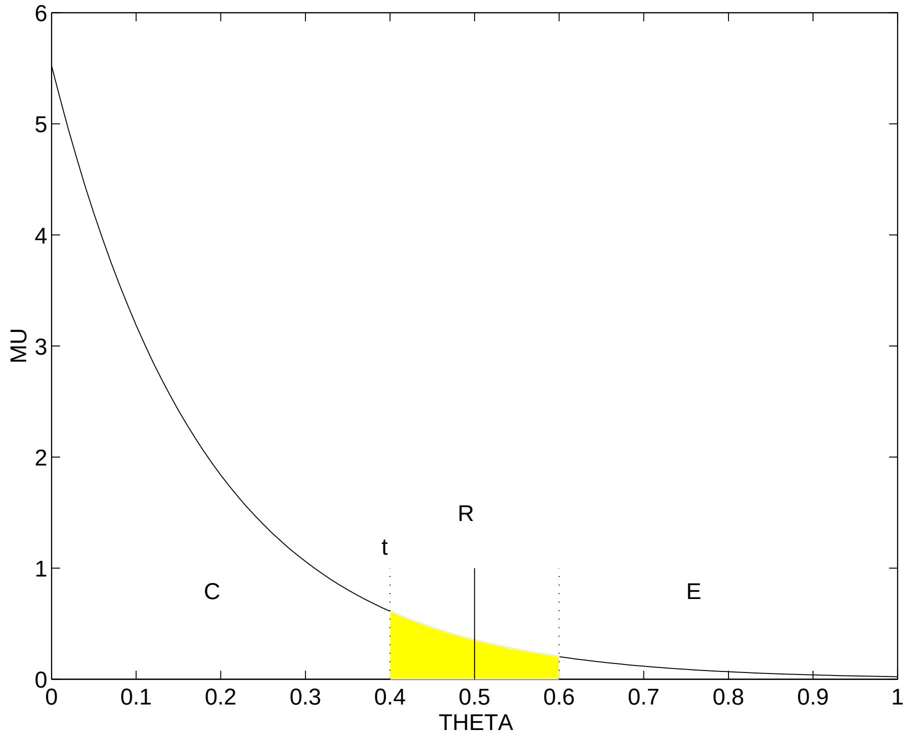

Figure 6: The figure shows the maximum entropy difficulty distribution for $\lambda = 5.5$ illustrating the region rejected using a threshold t.

This is illustrated in figure 6. Using the maximum entropy approximation (46) to the difficulty distribution $\mu(\theta)$, we find

$$E(t) = \frac{e^{-\lambda(1-t)} - e^{-\lambda}}{1 - e^{-\lambda}} \tag{56}$$

and

$$R(t) = \frac{e^{-\lambda t} - e^{-\lambda(1-t)}}{1 - e^{-\lambda}} \tag{57}$$

The appropriate value of λ in this infinite ensemble limit is most easily obtained using the implied value of

$$E_0 = \frac{e^{\lambda/2} - 1}{e^{\lambda} - 1} \tag{58}$$

giving

$$\lambda = 2\ln\left(\frac{1}{E_0} - 1\right). \tag{59}$$

The above equations (56) and (57) can be solved to give an analytic albeit cumbersome expression for E as a function of R; the parametric form presented here is more convenient for most purposes. The family of error-reject curves for various λ values are shown in figure 7 along with the data from several experiments. Three of these experiments are adapted from the NIST report [21], while two are from the benchmark test of Lee [20], and a final one is derived from the NIST data base using a small part of the training set for testing purposes [9].

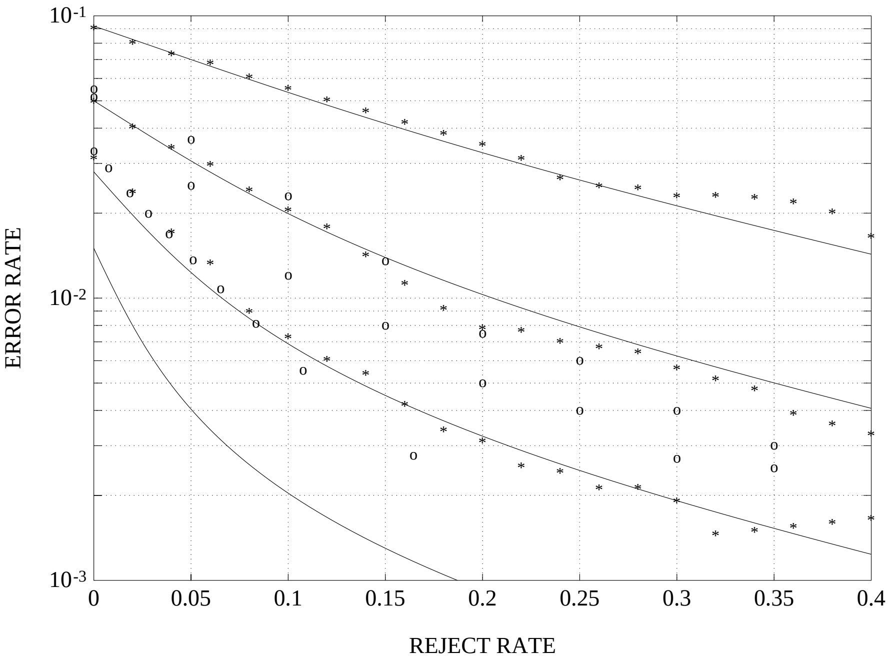

Figure 7: Error versus reject rates for different classifiers and the family of ensemble theory curves. The latter is parameterized by the zero reject error rate (full lines). The three sets of star-marked rates are three systems presented at the NIST Consensus conference. The three sets of open circles are derived from the experiments of Lee.

More interesting however is the scaled plot of these relations showing E/E_0 versus R/E_0. Figure 8 shows the same data scaled in this fashion along with the theoretical curve for $\lambda = 5.5$ which corresponds to $E_0 = 0.06$. It is interesting to note that in the large λ limit, i.e. for $\exp(\lambda/2) >> 1$,

$$E/E_0 = e^{\lambda(\frac{1}{2} - t)} \tag{60}$$

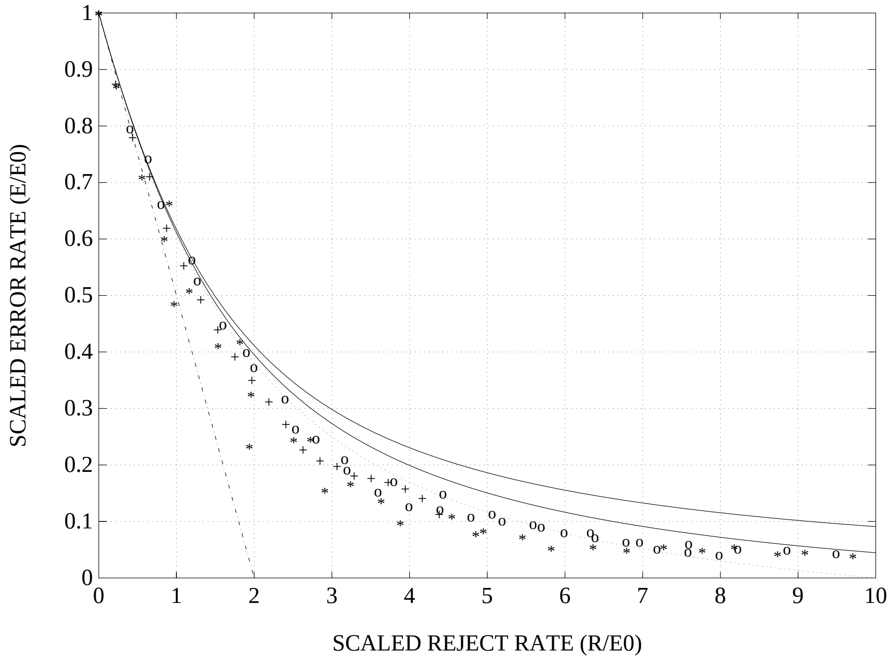

Figure 8: Experimental data plotted using the naive scaling relation overlaid with the ensemble error-reject tradeoff prediction for $E_0 = 0.01$ (dotted line), $E_0 = 0.06$ (solid line), and $E_0 = 0.10$ (dashed line). The dash-dotted line is the tradeoff prediction in relation (1).

while

$$R/E_0 = e^{\lambda(t-\frac{1}{2})} - e^{\lambda(\frac{1}{2}-t)} \tag{61}$$

and thus

$$E/E_0 = f(R/E_0) \tag{62}$$

with f given by

$$f(x) = \sqrt{1 + x^2/4} - x/2. \tag{63}$$

7 Conclusion

In this paper we analyzed the error-reject tradeoff for handwritten character recognition. By means of a simple scaling relationship suggested by Chow's theory of the error-reject tradeoff, we showed that the error-reject data from widely differing classifier algorithms show a universal structure and that this universality is explained by postulating that the problem has an effectively binary character, i.e. that when a classifier misclassifies a pattern, only one predominant alternative is considered. This postulate was confirmed several ways for digit recognition. Furthermore, we argue that most classification problems which can be learned to a high degree of proficiency will also exhibit such effectively binary character. We introduced a model scenario model which leads to such effectively binary character as a perturbation of the PAC model.

Our picture explains the universality of the scaled error-reject structure in effectively binary problems with finite datasets. The ambiguous inputs occur near the decision boundaries in input space. Reasonable reject rules reject patterns in the vicinity of such decision boundaries. Insofar as these boundaries are (on the average) placed similarly by the different classifiers, fattening them results in similar error-reject tradeoff curves.

Since we expect a universal shape for the error-reject curves, we can calculate these curves using any reasonable error-reject mechanism. We carry this out for the reject mechanism based on consensus among an ensemble of classifiers by using a maximum entropy estimate of the problem difficulty distribution. Analytic forms of the error-reject curve are derived and provide an excellent fit to the data for digit recognition using only a single parameter: the mean error rate at zero rejection. In the limit of very well trained networks, the scaled error-reject relationship assumes the very simple form

$$E/E_0 = \sqrt{1 + \left(\frac{R}{2E_0}\right)^2} - \frac{R}{2E_0}. \tag{64}$$

8 Acknowledgments

We wish to acknowledge the inspiring discussions of the 1992 and 1993 workshops on neural networks and complexity at the Telluride Summer Research Center. LKH thanks C. Van den Broeck for most enjoyable email conversations on the subject matter. PS would like to thank B. Andresen and M. Huleihil for helpful discussions related to example 2. This work is supported by the Danish Natural Science and Technical Research Councils through the Computational Neural Network Center (CONNECT). LKH acknowledge a generous donation from the Danish "Radio-Parts Fonden".

A Proofs of Chow's results

A.1 Monotonicity of E(t),R(t)

The two function are defined as integrals of positive integrands:

$$
\begin{aligned}
E(t) &= \int d\mathbf{x} P(\mathbf{x}) \theta\left(t - m(\mathbf{x})\right)\left(m(\mathbf{x})\right) = \int_{\{\mathbf{x}; m(\mathbf{x}) \leq t\}} m(\mathbf{x}) P(\mathbf{x}) d\mathbf{x} && (65)\\
R(t) &= \int d\mathbf{x} P(\mathbf{x}) \theta\left(m(\mathbf{x}) - t\right) = \int_{\{\mathbf{x}; m(\mathbf{x}) \geq t\}} P(\mathbf{x}) d\mathbf{x} && (66)
\end{aligned}
$$

The monotonicity follows by noting that the only t dependence is in the region of integration and these shrink and grow monotonically in t.

A.2 For differentiable rates: $dE/dR = -t$

On decreasing the threshold from t to $t + \Delta t$, the corresponding changes in E and R are

$$\Delta E = -\int_{\{\mathbf{x}; t+\Delta t \leq m(\mathbf{x}) \leq t\}} m(\mathbf{x}) P(\mathbf{x}) d\mathbf{x} \tag{67}$$

$$\Delta R = \int_{\{\mathbf{x}; t+\Delta t \leq m(\mathbf{x}) \leq t\}} P(\mathbf{x}) d\mathbf{x} \tag{68}$$

Since $P(\mathbf{x}) \geq 0$, we see from these that

$$-(t + \Delta t)\Delta R \leq \Delta E \leq -t \Delta R. \tag{69}$$

Provided that $\Delta R \neq 0$ for Δt sufficiently small, i.e. provided there exist $\mathbf{x}$ with $P(\mathbf{x}) \neq 0$ and $m(\mathbf{x})$ in the interval $[t + \Delta t, t]$, then dE/dR is defined and equals $-t$. If in the other hand $\Delta R = 0$ for a range of t values, then ΔE must also vanish for this same range. Thinking of the error-reject curve parametrized by the threshold t, the point on the curve sits still for such ranges of t values although the tangent turns leaving us with a continuous curve with possible corners.

A.3 E is a convex function of R

This follows immediately from the argument in the previous paragraph since the slope is a strictly increasing function along the curve. Note that we have one sided differentiability everywhere.

A.4 Working from $\widehat{m}(\mathbf{x})$

The argument for property A.1 remains unchanged since the region $\{\mathbf{x}; \widehat{m}(\mathbf{x}) \le t\}$ grows or shrinks with t exactly as the analogous region defined by $m(\mathbf{x})$ did. Note that while the definition of the regions of integration switches to $\widehat{m}$, the integrand for E is still m. In this case, the ratio $\Delta E/\Delta R$ becomes the average value of m in the region $\{\mathbf{x}; t + \Delta t \le \widehat{m}(\mathbf{x}) \le t\}$. In the limit $\Delta t \to 0$ this becomes the average of m over the hypersurface $\{\mathbf{x}; \widehat{m}(\mathbf{x}) = t\}$ as in equation (25).

References

[1] J. Denker, D. Schwartz, B. Wittner, S. Solla, R. Howard, and L. Jackel: *Large Automatic Learning, Rule Extraction, and Generalization* Complex Systems , (1987)

[2] D.J.C MacKay: *Bayesian Interpolation*, Neural Computation **4**, 415-447, (1992).

[3] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth: *Learnability and the Vapnik-Chervonenkis Dimension*. Journ. of the ACM **36**, 929 (1989).

[4] D. Haussler: *Decition Theoretic Generalization of the PAC Model for Neural Net and Other Learning Applications*. Preprint, Baskin Center for Computer Engineering and Information Science, University of California Santa Cruz.

[5] F.J. Śmieja, *Multiple network systems (MINOS) modules: Task division and module discrimination*, Proceedings of the 8th AISB Conference on Artificial Intelligence, Leeds, 13-25, (1991).

[6] C.K. Chow: *On Optimum Recognition Error and Reject Tradeoff* IEEE Transactions on Information Theory **IT-16**, 41 (1970).

[7] R.O. Duda and P.E. Hart: *Pattern Classification and Scene Analysis* Wiley-Interscience, New York (1973).

[8] R.T. Clemen *Combining forecast: A review and annotated bibliography.* Journal of Forecasting **5**, 559 (1989).

[9] Chr. Liisberg: *SYSTEM: RISO.* In In Eds. R.A. Wilkinson et al. "The First Census Optical Character Reconition System Conference". US. Dept. of Commerce: NISTIR 4912 (1992).

[10] J. Geist and R.A. Wilkinson: *System Error Rates Versus Rejection Rates.* In Eds. R.A. Wilkinson et al. "The First Census Optical Character Reconition System Conference". US. Dept. of Commerce: NISTIR 4912 (1992).

[11] Y. Le Cun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jakel: *Handwritten Digit Recognition with a Back-Propagation Network*, In Advances in Neural Information Processing Systems II (Denver 1989) ed. D.S.Touretzsky, 396-404. San Mateo: Morgan Kaufman. (1990)

[12] D.W. Ruck, S.K. Rogers, M. Kabrisky, M. Oxley, and B. Suter: *The Multilayer Perceptron as an Approximation to a Bayes Optimal Discriminant Function.* IEEE Transactions on Neural Networks **1**, 296-298, (1990).

[13] J.M.R. Parrondo and C. Van den Broeck: *Error Versus Rejection Curve for the Perceptron*, preprint (1992)

[14] L.K Hansen and P. Salamon: *Neural Network Ensembles*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **12**, 993-1001 (1990).

[15] P. Salamon, L. K. Hansen, B. E. Felts III., and C. Svarer: *The Ensemble Oracle*, AMSE Conference on Neural Networks. San Diego 1991.

[16] L. K. Hansen and P. Salamon: *Self-Repair in Neural Network Ensembles*, AMSE Conference on Neural Networks. San Diego 1991.

[17] L.K. Hansen, C. Liisberg, and Peter Salamon: *Ensemble Methods for Recognition of Handwritten Digits.* In 'Neural Networks For Signal Processing'; Proceedings of the 1992 IEEE-SP Workshop, (Eds. S.Y. Kung, F. Fallside, J. Aa. Sørensen, and C.A. Kamm), IEEE Service Center, Piscataway NJ, 540-549, (1992).

[18] V.K. Govindan and A.P. Shivaprasad: *Character recognition - a review*, Pattern Recognition **23** (1990).

[19] Y. Tikochinsky, N.Z. Tishby, and R.D. Levine: *Alternative approach to maximum-entropy inference.* Phys.Rev. **A-30**, 2638 (1984).

[20] Y. Lee: *Handwritten Digit Recognition Using K Nearest-Neighbor, Radial-Basis Function,and Backpropagation Neural Networks* Neural Computation **3**, 440-449, (1991)

[21] National Institute of Standards and Technology: *NIST Special Data Base 3, Handwritten Segmented Characters of Binary Images*, HWSC Rel. **4-1.1** (1992).

[22] D.B. Schwartz, V.K. Salalam, S.A. Solla, and J.S. Denker: *Exhaustive Learning*, Neural Computation **2**, 371-382 (1990).