Selecting a Restoration Technique to Minimize OCR Error

Los Alamos
National Laboratory

Mike Cannon, Mike Fugate, Don Hush and Clint Scovel
Computer Research and Applications Group, CCS–3
Mail Stop B265

LANL Technical Report: LA–UR–01-6860

Report Date: January 24, 2003

Abstract

This paper introduces a learning problem related to the task of converting printed documents to ASCII text files. The goal of the learning procedure is to produce a function that maps documents to restoration techniques in such a way that on average the restored documents have minimum OCR error. We derive a general form for the optimal function and use it to motivate the development of a nonparametric method based on nearest–neighbors. We also develop a direct method of solution based on empirical error minimization for which we prove a finite sample bound on estimation error that is independent of distribution. We show that this empirical error minimization problem is an extension of the empirical optimization problem for traditional M–class classification with general loss function and prove computational hardness for this problem. We then derive a simple iterative algorithm called `Generalized Multi-Class Ratchet` (`GMR`) and prove that it produces an optimal function asymptotically (with probability 1). To obtain the `GMR` algorithm we introduce a new data map that extends Kesler's construction for the multiclass problem (e.g. see p. 266 in (Duda, Hart, & Stork, 2000)) and then apply an algorithm called `Ratchet` to this mapped data, where `Ratchet` is a modification of the `Pocket` algorithm (Gallant, 1990). Finally we apply these methods to a collection of documents and report on the experimental results.

Acknowledgments

We would like to thank James Theiler and Adam Cannon for their constructive feedback on earlier versions of this document.

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W–7405–ENG–36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.

1 Introduction

We describe a learning problem related to the task of converting printed documents to ASCII text files. Existing optical character reader (OCR) systems can accomplish this with high accuracy when the document is pristine, but tend to perform poorly when the document contains noise or distortion. One solution is to add a stage to the OCR process pipeline that enhances the digitized document before it is converted to ASCII. The operation performed by this stage, which accepts a digitized document and produces an enhanced digitized document, is referred to as "restoration". Restoration techniques work best when they are specialized to the type of noise or distortion present in the document. Indeed, several restoration techniques have been developed with this in mind (Cannon, Hochberg, & Kelly, 1999). With such techniques in hand what remains is to choose the restoration technique most appropriate for a given document. In short, we seek a function that maps digitized documents to restoration techniques in such a way that the restored documents have reduced OCR error rates. It is common to decompose this function into two stages; a feature map which converts digitized documents to feature vectors, and a classifier which maps feature vectors to a choice of restoration technique. The feature map is crafted by the designer with the goal of producing a simplified representation of documents that retains information essential to the task. For example several features that carry information relevant to the selection of a suitable restoration technique are described in (Cannon *et al.*, 1999). Although the map from feature vectors to restoration techniques is a standard multi–class classifier, we will see that the corresponding learning problem is not standard. This learning problem is the main concern of this paper. The document conversion process just described is illustrated in Figure 1.

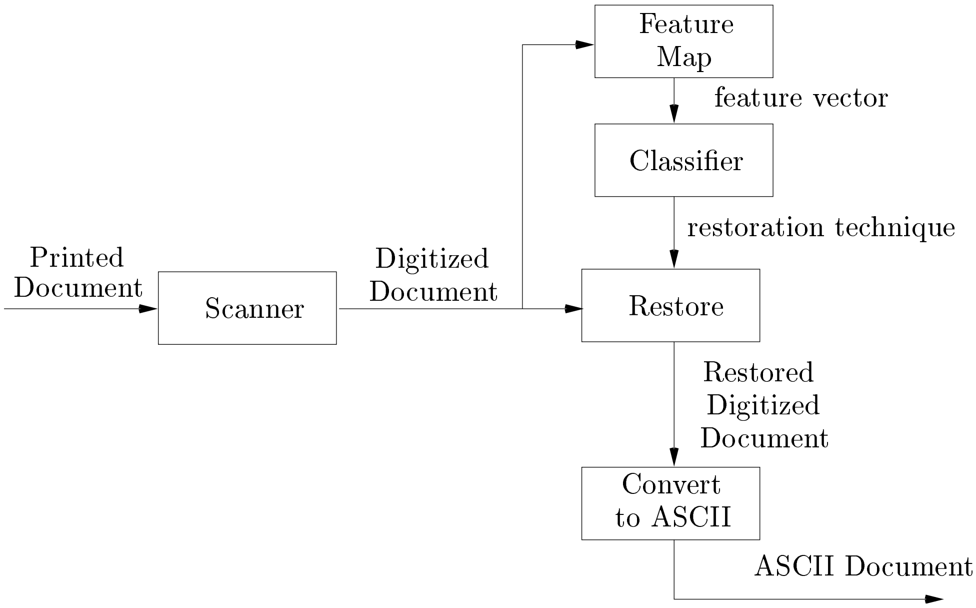

Figure 1: Document Conversion Process

2 Formulation of the Learning Problem

In the process described above printed documents are digitized and converted to feature vectors. Let $\mathcal{X}$ denote the space where the feature vectors live. The map from printed documents to feature vectors is generally many–to–one so that multiple documents may map to the same feature vector. Let M be the number of restoration techniques (including "no restoration") and without loss of generality let $\mathcal{C} = \{1, ..., M\}$ be the space of labels for restoration techniques. Let $\mathcal{Y} = [0, 1]^M$ be the space of OCR error rate vectors where $y \in \mathcal{Y}$ contains one component for each restoration technique. Now consider a real world system that converts printed documents to ASCII files using the process illustrated in Figure 1, and then determines their error rates. We assume that the feature and error rate vectors are *i.i.d.* samples from a random process characterized by a measure P on $\mathcal{X} \times \mathcal{Y}$.

Let $\mathcal{F} : \mathcal{X} \to \mathcal{C}$ denote the class of functions that we wish to consider for our classifier and let $e : \mathcal{F} \to [0, 1]$ be the function that computes average OCR error rate, that is

$$e(f) = E\left[y^{f(x)}\right] \tag{1}$$

where y^i is the i-th component of the vector y. We seek a function $f \in \mathcal{F}$ that minimizes e. However, since P is unknown $e(f)$ is not computable and determination of such a function through direct optimization is not possible. On the other hand it may be possible to use empirical information (e.g. examples of x and y) to produce a near optimal f.

In particular we consider the employment of empirical information obtained as follows. A corpus of n documents is gathered and digitized. Each document is restored using all M techniques, and each restored document is converted to an ASCII file using a conventional OCR system. Then the character error rate for each ASCII file is determined by a human expert. This gives a collection $D_n = ((x_1, y_1), ..., (x_n, y_n))$ of empirical observations where x_i is the feature vector representation of document i and y_i is the M-vector of error rates for document i. Our goal is to determine a learning procedure L that accepts D_n as its input and outputs a function f with $e(f)$ as small as possible.

In Section 3 we present a *nearest neighbor* method. This method is developed in the spirit of the traditional nonparametric approach to statistical inference which determines the general form of the optimal inference rule as a function of the data distribution (or parameters thereof) and then substitutes a nonparametric (e.g. local) estimate of the distribution function (or its parameter values) into this general form. This method is *indirect* in the sense that the learning process is more directly concerned with the estimation of a distribution function (or its parameter values) than with the accuracy of the induced inference rule.

In Section 4 we develop a *direct* method that chooses f to minimize an empirical version of the error function defined in (1). We prove that the excess error due to optimization over a finite sample is bounded and converges to zero as n goes to infinity. Specifically, let e^* be the optimal error

$$e^* = \inf_{f \in \mathcal{F}} e(f)$$

and let f_n be a function determined by minimizing the empirical error. We prove a bound on the *estimation error* $e(f_n) - e^*$ that (with high probability) decreases monotonically to zero with n. We show that the empirical error minimization problem is an extension of the empirical optimization problem for traditional M–class classification with general loss function. We prove computational hardness for this problem, and then develop a simple algorithm called `Generalized Multi-Class Ratchet` (`GMR`) which we prove provides an optimal solution asymptotically (with probability one).

Finally, in Section 6 we present experimental results that compare these methods with previous methods on a real world corpus.

3 A Nearest Neighbor Method

Suppose we wish to determine a function that minimizes the error defined in (1). The following theorem characterizes the form of the optimal solution and motivates the nearest neighbor method described below.

Theorem 1. *Let $\mathcal{X}$ be a set, $\mathcal{Y} = [0,1]^M$ and P be a measure on $\mathcal{X} \times \mathcal{Y}$ with density p. Let*

$$\mu(x) = E\left[y|x\right] \tag{2}$$

be the mean of y given a value x. Define a function $f^* : \mathcal{X} \to \mathcal{C}$ by

$$f^*(x) \in \arg\min_{j \in \mathcal{C}} \mu^j(x)$$

where ties in the $\arg\min_j$ *function are resolved by some fixed rule. Then the function f^* provides an optimal solution to the problem*

$$\min_f \; e(f)$$

where $e(f) = E\left[y^{f(x)}\right]$.

Proof. The criterion e takes the form

$$\begin{aligned} E\left[y^{f(x)}\right] &= \int y^{f(x)} p(x,y) dx dy \\ &= \int \left[\int y^{f(x)} p(y|x) dy\right] p(x) dx \\ &= \int \left[\int y^{f(x)} p(y^{f(x)}|x) dy^{f(x)}\right] p(x) dx \\ &= \int \mu(x)^{f(x)} p(x) dx \end{aligned}$$

where $\mu(x)^{f(x)}$ is the $f(x)$–th component of $\mu(x)$. By definition $\mu(x)^{f^*(x)} \leq \mu(x)^{f(x)}$ giving

$$e(f^*) = \int \mu(x)^{f^*(x)} p(x) dx \leq \int \mu(x)^{f(x)} p(x) dx = e(f)$$

for all f. ♦

Theorem 1 motivates an approach of the form

$$\hat{f}(x) \in \arg\min_{c \in \mathcal{C}} \hat{\mu}^c(x) \tag{3}$$

where $\hat{\mu}(x)$ is an estimate of the mean error vector at x. If $\mathcal{X}$ is a metric space then a nonparametric estimate $\hat{\mu}(x)$ can be formed by applying a nearest neighbor method to the empirical observations in D_n. For example a simple (but crude) estimate of $\hat{\mu}(x)$ takes the form

$$\hat{\mu}(x) = \frac{1}{k} \sum_{i \in \mathcal{K}(x)} y_i \tag{4}$$

where $\mathcal{K}(x)$ is the index set of the k samples from $(x_1, ..., x_n)$ that are closest to x under the metric. We refer to the classifier $\hat{f}$ in (3) with $\hat{\mu}$ given by (4) as the k–nearest neighbor method. We provide no formal analysis of this method, but its simplicity, along with the historic success of this general approach to statistical inference, make it an attractive candidate for our problem and so we include it in our empirical comparisons in Section 6. In the next two sections we develop an alternative method based on empirical error minimization which is the main contribution of this paper.

4 Empirical Error Minimization

Finite sample bounds on estimation error that are independent of the distribution can often be derived for learning procedures that minimize a particular form of empirical error. For our problem we define the empirical error as follows

$$e_n(f) = \frac{1}{n} \sum_{i=1}^{n} y_i^{f(x_i)}. \tag{5}$$

This is simply a Monte Carlo estimate of $e(f)$ when the samples in D_n are *i.i.d.* Our learning procedure then chooses a member of $\mathcal{F}$ that minimizes e_n, i.e.

$$f_n \in \arg\min_{f \in \mathcal{F}} e_n(f). \tag{6}$$

In section 4.1 we give support for this procedure by establishing a bound on estimation error as a function of a shatter coefficient of $\mathcal{F}$. Section 4.2 describes a particular class of functions $\mathcal{F}$ called *linear machines* and derives a shatter coefficient bound for this class. Then, in Section 5 we discuss computational issues related to empirical error minimization and develop a simple algorithm that is later used in our experiments.

4.1 Performance Bounds for Empirical Error Minimization

The theorem below gives a bound on estimation error as a function of the following n-shatter coefficient of $\mathcal{F}$.

Definition 1. Let $\mathcal{X}$ be a set, $\mathcal{C} = \{1, 2, ..., M\}$, and $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathcal{C}$. Let $X_n = (x_1, x_2, ..., x_n) \in \mathcal{X}^n$. We define the restriction f_{X_n} of a function $f \in \mathcal{F}$ to the data sample X_n in the natural way and denote the resulting class of functions

$$\mathcal{F}_{X_n} = \{f_{X_n} : X_n \to \mathcal{C} | f \in \mathcal{F}\}.$$

The n-shatter coefficient for $\mathcal{F}$ is defined

$$S(\mathcal{F}, n) = \sup_{X_n} |\mathcal{F}_{X_n}|$$

The following theorem holds.

Theorem 2. *Let $\mathcal{X}$ be a set, $\mathcal{C} = \{1, 2, ..., M\}$, $\mathcal{Y} = [0,1]^M$ and P be a measure on $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathcal{C}$. Let $D_n = ((x_1, y_1), (x_2, y_2), ..., (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ be n i.i.d. random samples and let P_n denote the corresponding n–fold product measure. Let e be the error function $e(f) = E\left[y^{f(x)}\right]$, with $e^* = \inf_{f \in \mathcal{F}} e(f)$, and empirical error $e_n(f) = \frac{1}{n}\sum_{i=1}^n y_i^{f(x_i)}$. Let f_n be chosen to satisfy*

$$f_n \in \arg\min_{f \in \mathcal{F}} e_n(f). \tag{7}$$

Then for every $\epsilon > 8/n$,

$$P_n\left(e(f_n) - e^* > \epsilon\right) \leq 2S(\mathcal{F}, 2n) e^{-n\epsilon^2/16(e^*+\epsilon)} + e^{-n(\epsilon - 8/n)^2/4(8e^*+\epsilon)}$$

where $S(\mathcal{F}, 2n)$ is the $2n$-shatter coefficient of $\mathcal{F}$.

Proof. Our proof uses a theorem by Bartlett and Lugosi (stated below) which gives an estimation error bound in terms of the following covering numbers. Let $\mathcal{Z}$ be a set and consider a class $\mathcal{G}$ of functions $g : \mathcal{Z} \to [0,1]$. Let $Z_n = (z_1, ..., z_n) \in \mathcal{Z}^n$ and define the distance between two functions in $\mathcal{G}$ to be

$$d_\infty(g_1, g_2, Z_n) = \max_{z_i \in Z_n} |g_1(z_i) - g_2(z_i)|.$$

The covering number $N_\infty(\mathcal{G}, Z_n, \epsilon)$ of $\mathcal{G}$ at scale ϵ is the smallest N for which a set $\mathcal{G}_\epsilon = \{g_1, ..., g_N\}$ exists such that for every $g \in \mathcal{G}$ there exists a $g_i \in \mathcal{G}_\epsilon$ such that $d_\infty(g, g_i, Z_n) < \epsilon$.

The following theorem is proved in (Bartlett & Lugosi, 1999).

Theorem 3 (Bartlett and Lugosi). *Let $D_n = ((x_1, y_1), (x_2, y_2), ..., (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ be n i.i.d. random samples. Let $l(\cdot, \cdot)$ be a function that takes values in $[0,1]$ and let $\mathcal{L}$ be a class of loss functions $\mathcal{L} = \{l(f(\cdot), \cdot) : f \in \mathcal{F}\}$ where $\mathcal{F}$ is a class of functions on $\mathcal{X}$. Let e be the expected loss $e(f) = E[l(f(x), y)]$, with $e^* = \inf_{f \in \mathcal{F}} e(f)$, and empirical loss $e_n(f) = \frac{1}{n}\sum_{i=1}^n l(f(x_i), y_i)$. Let $f_n \in \mathcal{F}$ be chosen to satisfy*

$$e_n(f_n) \leq \inf_{f \in \mathcal{F}} e_n(f) + 1/n \tag{8}$$

Then for every $\varepsilon > 4/n$,

$$P_n(e(f_n) - e^* > 2\varepsilon) \leq 2E_{2n}\left[N_\infty\left(\mathcal{L}, D_{2n}, \frac{\varepsilon}{8}\right)\right] e^{-n\varepsilon^2/(4e^*+8\varepsilon)} + e^{-n(\varepsilon - 4/n)^2/(8e^*+2\varepsilon)}.$$

With the appropriate choice of loss function this theorem applies directly to our problem. In fact this theorem can probably be proved directly in terms of $|\mathcal{F}_{X_{2n}}|$, but we see no virtue in doing so here. To apply this theorem we let $\mathcal{F}$ be the function class defined in Theorem 2 and define $l(u, y) = y^u$ so that the class of loss functions in (x, y) is given by

$$\mathcal{L} = \{(x, y) \mapsto y^{f(x)} : f \in \mathcal{F}\}$$

With this class $\inf_{f\in\mathcal{F}} e_n(f) = \min_{f\in\mathcal{F}} e_n(f)$ and so the function f_n chosen in (7) satisfies condition (8) in Theorem 3. Thus, to prove theorem 2 we need only replace ε with $\epsilon/2$ and show that

$$E_{2n}\left[N_\infty\left(\mathcal{L}, D_{2n}, \frac{\epsilon}{16}\right)\right] \leq S(\mathcal{F}, 2n)$$

Let $\mathcal{F}_{X_{2n}}$ and $f_{X_{2n}}$ be as defined in Definition 1 and $\mathcal{L}_{X_{2n}}$ be the restriction of $\mathcal{L}$ to $\mathcal{F}_{X_{2n}}$. Since $d_\infty(f_1, f_2, X_{2n}) = \max_{x_i\in X_{2n}} |f_1(x_i) - f_2(x_i)|$, then $\mathcal{F}_{X_{2n}}$ is an ϵ–cover of $\mathcal{F}$ with respect to d_∞ for any $\epsilon \geq 0$ and since $|l(f(x_i), y_i) - l(f_{X_{2n}}(x_i), y_i)| = |y_i^{f(x_i)} - y_i^{f_{X_{2n}}(x_i)}| = 0$ this implies $\mathcal{L}_{X_{2n}}$ is an ϵ–cover of $\mathcal{L}$ for any $\epsilon \geq 0$. Thus $N_\infty\left(\mathcal{L}, D_{2n}, \frac{\epsilon}{16}\right) \leq |\mathcal{L}_{X_{2n}}| \leq |\mathcal{F}_{X_{2n}}|$ and

$$E_{2n}\left[N_\infty\left(\mathcal{L}, D_{2n}, \frac{\epsilon}{16}\right)\right] \leq \sup_{X_{2n}} |\mathcal{F}_{X_{2n}}| = S(\mathcal{F}, 2n).$$

♦

The fact that we have obtained the bound in Theorem 2 as a special case of a more general theorem suggests that it may be possible to improve the bound. In addition, to make the final result easier to interpret we have bounded the expected value of N_∞ by its supremum which may contribute to the looseness of the bound. Nevertheless Theorem 2 establishes the first finite sample bound on estimation error for the learning problem described in Section 2 and as a consequence it not only enables a consistency result for the empirical error minimization learning strategy but gives a bound on the rate at which the estimation error goes to zero. Indeed, we now state a simple corollary of this theorem which demonstrates a bound on this rate. With probability $1 - \delta$,

$$e(f_n) \leq e^* + \frac{32}{n}\log\left(\frac{2S(\mathcal{F}, 2n) + e^4}{\delta}\right) + \sqrt{e^*}\sqrt{\frac{32}{n}\log\left(\frac{2S(\mathcal{F}, 2n) + e^4}{\delta}\right)}. \tag{9}$$

To see this let

$$\delta = 2S(\mathcal{F}, 2n)e^{-n\epsilon^2/16(e^*+\epsilon)} + e^{-n(\epsilon-8/n)^2/4(8e^*+\epsilon)} \tag{10}$$

and note that the probability statement in the theorem implies

$$P_n\left(e(f_n) - e^* \leq \epsilon\right) > 1 - \delta. \tag{11}$$

To obtain (9) we use (10) to derive an upper bound on ϵ as a function of δ, which in turn provides a probabilistic bound on $e(f_n) - e^*$ according to (11). In this derivation the restriction $\epsilon > 8/n$ maps to a restriction on δ. For $\epsilon > 8/n$ the right hand side of (10) is monotonically decreasing in ϵ which implies that $\delta < 1 + 2S(\mathcal{F}, 2n)e^{-4/(ne^*+8)}$, which in light of (11) is a trivial restriction.

The second term in (10) is less than

$$e^{\frac{16\epsilon}{4(8e^*+\epsilon)}} e^{-\frac{n\epsilon^2}{4(8e^*+\epsilon)}}$$

and since $\frac{\epsilon}{8e^*+\epsilon} \leq 1$ the second term is less than

$$e^4 e^{-\frac{n\epsilon^2}{4(8e^*+\epsilon)}}$$

and

$$\delta \leq 2S(\mathcal{F}, 2n)e^{-\frac{n\epsilon^2}{16(e^*+\epsilon)}} + e^4 e^{-\frac{n\epsilon^2}{4(8e^*+\epsilon)}}$$

$$\leq (2S(\mathcal{F}, 2n) + e^4)e^{-\frac{n\epsilon^2}{32(e^*+\epsilon)}}$$

so we can say that

$$\delta \leq \acute{\delta} = (2S(\mathcal{F}, 2n) + e^4)e^{-\frac{n\epsilon^2}{32(e^*+\epsilon)}}$$

If we further denote $b = \frac{32}{n} \log \frac{2S(\mathcal{F},2n)+e^4}{\acute{\delta}}$ then

$$\frac{\epsilon^2}{e^* + \epsilon} = b$$

or

$$\epsilon^2 - b\epsilon - be^* = 0$$

and we use the solution

$$\epsilon = \frac{b + \sqrt{b^2 + 4e^*b}}{2}.$$

Applying the inequality

$$\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}, \quad \forall a, b \in \Re^+$$

we obtain

$$\epsilon \leq \frac{b + b + 2\sqrt{e^*}\sqrt{b}}{2} = b + \sqrt{e^*}\sqrt{b}$$

or written out

$$\epsilon \leq \frac{32}{n} \log\left(\frac{2S(\mathcal{F}, 2n) + e^4}{\acute{\delta}}\right) + \sqrt{e^*}\sqrt{\frac{32}{n} \log\left(\frac{2S(\mathcal{F}, 2n) + e^4}{\acute{\delta}}\right)}.$$

Since $e(f_n) - e^* \leq \epsilon$ and $\delta \leq \acute{\delta}$ then with probability $1 - \delta$ we obtain the result in (9). This result guarantees the following. The employment of empirical risk minimization with function classes $\mathcal{F}$ whose shatter coefficient grows subexponentially with n produces a function f_n whose generalization error exceeds the optimal generalization error by an amount that is bounded by an expression that decreases asymptotically to zero with n. In the next section we analyze a specific function class called *linear machines* and show their shatter coefficients to be polynomial in n. This translates into a bound $O(\sqrt{\ln n/n})$ on the rate at which the estimation error goes to zero.

4.2 Linear Machines

We now restrict our attention to classes $\mathcal{F}$ called *linear machines*. Let $\mathcal{X}$ be a set. We construct the space of functions $f : \mathcal{X} \to \mathcal{C}$ in the following way. Let $\mathcal{W}$ denote a real vector space of linear functions $w : \mathcal{X} \to \Re$ of dimension d. Let $\hat{\mathcal{F}}$ denote the space of maps $\hat{f} : \mathcal{X} \to \Re^M$ consisting of M choices $\hat{f}^i \in \mathcal{W}, i = 1, .., M$. The general form of a linear machine is the composition of an element of $\hat{\mathcal{F}}$ and the *winner–take–all* function, i.e.

$$f(x) \in \arg\max_{k \in \mathcal{C}} \hat{f}^k(x) \tag{12}$$

where a rule is provided for breaking ties. For simplicity we employ the tie breaking rule that chooses the largest index involved in the tie. That is we define the linear machine determined by $\hat{f}$ as

$$f_{\hat{f}}(x) = \max_{i \in \mathcal{I}_{\hat{f}}(x)} i. \tag{13}$$

where $\mathcal{I}_{\hat{f}}(x)$ is the set

$$\mathcal{I}_{\hat{f}}(x) = \arg\max_{k \in \mathcal{C}} \hat{f}^k(x).$$

We denote the class of linear machines that results when $\hat{f}$ varies over all of $\hat{\mathcal{F}}$ by $\mathcal{F}$.

We now determine a bound on the shatter coefficient for $\mathcal{F}$. We denote the set of points in $\mathcal{X}$ determined by X_n with the same name $X_n \subset \mathcal{X}$. We define f_{X_n} and $\mathcal{F}_{X_n}$ as in Definition 1. Our goal here is to bound $|\mathcal{F}_{X_n}|$ uniformly in X_n. To this end we utilize an analogue of VC dimension and Sauer Lemma for such classes of sets. We follow Natarajan (Natarajan, 1991).

A subset S is said to be shattered by $\mathcal{F}$ if there exist two functions $f, g \in \mathcal{F}$ such that

1. For all $x \in S, f(x) \neq g(x)$.
2. For all $S_1 \subset S$, there exists an $h \in \mathcal{F}$ such that

$$h(x) = f(x), \quad x \in S_1$$

$$h(x) = g(x), \quad x \in S - S_1.$$

The (now called) Natarajan dimension $\mathcal{N}(\mathcal{F}, X_n)$ of $\mathcal{F}$ with respect to X_n is defined as the size of the largest subset of X_n which is shattered by $\mathcal{F}$. We now bound this dimension in terms of $\mathcal{W}$ and $\mathcal{C}$.

Lemma 1. *Let $\mathcal{W}$, $\mathcal{C}$, and X_n be as above. Then*

$$\mathcal{N}(\mathcal{F}, X_n) \leq M(M-1)d.$$

Proof. Let $S \subset X_n \subset \mathcal{X}$ be a set which is shattered by $\mathcal{F}$. As a consequence of this definition of shattering, there exist two functions f and g with the properties mentioned above. Consider two values $i_1, i_2 \in \mathcal{C}$ where $i_1 < i_2$ and define

$$Q(i_1, i_2) = \{x \in X_n : f(x) = i_1, g(x) = i_2\}$$

and

$$S_{i_1 i_2} = S \cap Q(i_1, i_2).$$

For any subset $S_1 \subset S_{i_1 i_2}$ there exists an $h \in \mathcal{F}$ such that $h(x) = f(x) = i_1$ when $x \in S_1$ and $h(x) = g(x) = i_2$ when $x \in S_{i_1 i_2} - S_1$. Let $\hat{h}$ be a member of $\hat{\mathcal{F}}$ that determines h and let

$$\delta h = \hat{h}^{i_1} - \hat{h}^{i_2}.$$

Then from the definition (13) of the linear machine h

$$S_1 = \{x \in S_{i_1 i_2} : \delta h(x) > 0\}.$$

Consequently under the standard definition of shattering for binary function classes (e.g. see p. 196 in (Devroye, Györfi, & Lugosi, 1996)) $S_{i_1 i_2}$ is shattered by $\mathcal{W}$ and by the theorem of Steele and Dudley (e.g. see p. 221 in (Devroye *et al.*, 1996)) on the VC dimension of classifiers determined from vector spaces of linear functions

$$|S_{i_1 i_2}| \leq d.$$

The same inequality is true when $i_2 < i_1$. Finally since $X_n = \cup_{i_1 \neq i_2} Q(i_1, i_2)$ is a disjoint union

$$|S| = \sum_{i_1 \neq i_2} |S_{i_1 i_2}| = \sum_{i_1 \neq i_2} |S \cap Q(i_1, i_2)| \leq M(M-1)d$$

and the proof is finished. ♦

We are now in a position to bound $|\mathcal{F}_{X_n}|$.

Theorem 4. *With the same assumptions as Lemma* 1,

$$|\mathcal{F}_{X_n}| \leq M^{2M^2 d} n^{M^2 d}$$

Proof. The analogue of the Sauer lemma we use is Lemma 5.1 on page 104 in (Natarajan, 1991);

$$|\mathcal{F}_{X_n}| \leq M^{2\mathcal{N}(\mathcal{F}, X_n)} n^{\mathcal{N}(\mathcal{F}, X_n)}.$$

If we then apply the simplified bound $\mathcal{N}(\mathcal{F}, X_n) \leq M(M-1)d \leq M^2 d$ from Lemma 1 the proof is finished. ♦

5 Algorithms for Empirical Error Minimization

Section 4.1 established a generalization error bound for learning procedures that solve the empirical error minimization problem in (6). This problem contains empirical optimization for the traditional M–class classification problem with general loss function as a special case. To see this consider the traditional M–class classification problem with training data $((x_1, c_1), ..., (x_n, c_n))$ (where $c_i \in \mathcal{C}$). Let L be the $M \times M$ loss matrix where $L_{c,\hat{c}}$ is the loss incurred when a pattern from class c is assigned to class $\hat{c}$. If we set $y_i = (L_{c_i,1}, L_{c_i,2}, ..., L_{c_i,M})$ then $e_n(f)$ is precisely the empirical loss function for the M-class classifier f. The optimization problem in (6) is more general. Since the loss vectors depend on x they may be different for each of the n data samples and therefore may not be the row vectors of any fixed loss matrix. For this reason we call the optimization problem in (6) the *M-Class Classification with Generalized Loss* (MCGL) problem.

We now restrict our study of the MCGL problem to the case where $\mathcal{F}$ is the class of linear machines. We use the acronym MCGL_{LM} for this problem. For computation we must be more specific about $\mathcal{X}$ and $\mathcal{W}$ than we were in Section 4.2. To this end we assume our feature measurements x live in $\mathcal{X} \subseteq \Re^d$ and we choose $\mathcal{W} = \Re^{d+1}$ to be the class of affine functions on $\mathcal{X}$ defined by $w \cdot (1, x)$, where $\cdot$ is the usual inner product. Thus $\mathcal{W}$ is the class of linear functions restricted to the domain $\mathcal{X}_1 = 1 \times \mathcal{X} \subset \Re^{d+1}$. This gives $\hat{\mathcal{F}} = \Re^{M(d+1)}$ and a shatter coefficient for $\mathcal{F}$ that satisfies Theorem 4 with d replaced by $d+1$. To maintain a clear distinction between feature measurements $x \in \mathcal{X}$ and members of the domain $(1, x) \in \mathcal{X}_1$ we use ξ as the domain variable for $\mathcal{X}_1$. In addition we adopt the notation $\omega = (w_1, w_2, ..., w_M) \in \Re^{M(d+1)}$ for $\hat{f}$ so that (13) becomes

$$f_\omega(\xi) = \max_{i \in \mathcal{I}_\omega(\xi)} i \tag{14}$$

where

$$\mathcal{I}_\omega(\xi) = \arg\max_{k \in \mathcal{C}} \omega_k \cdot \xi. \tag{15}$$

In section 5.1 we show that the decision version of MCGL_{LM} is computationally intractable in the sense that there are instances that cannot be solved in polynomial time. Among the paths that might be pursued in this situation we choose the development of an algorithm for which we can prove asymptotic optimality. To this end we develop an adaptation of the `Pocket-with-Ratchet` algorithm in (Gallant, 1990) that we call `Ratchet`. We prove that `Ratchet` is asymptotically optimal for criteria that satisfy a property called PLD that we define in Section 5.2. Then in Section 5.3 we prove that the MCGL_{LM} criterion satisfies this PLD property and use this result to determine a Generalized Multi-Class Ratchet (`GMR`) algorithm for this problem. Although there is no guarantee that `GMR` will produce an optimal solution in practice where we have finite computational resources, it is very simple and tends to perform well in our experiments.

5.1 A Computational Hardness Result for MCGL_{LM}

In this section we show that the following decision version of MCGL_{LM} is NP–Hard.

Definition 2 (DECISION–MCGL_{LM} (DMCGL_{LM})). Given a positive real number ε, positive integers d and M, a finite data sample $\Xi = ((\xi_1, y_1), ..., (\xi_n, y_n))$ where $\xi_i \in 1 \times \Re^d$ and $y_i \in [0,1]^M$, and a class of functions $\mathcal{F} : 1 \times \Re^d \to \{1, 2, ..., M\}$ defined in (14), does there exist an $\omega \in \Re^{M(d+1)}$ such that

$$\sum_{i=1}^{n} y_i^{f_\omega(\xi_i)} \le \varepsilon \ ?$$

Lemma 2. *$DMCGL_{LM}$ is NP-Hard.*

Proof. The proof uses a reduction from the APPROX-HALFSPACES problem which is shown to be NP–Complete in (Höffgen & Simon, 1992).

Definition 3 (APPROX-HALFSPACES). (Höffgen and Simon) Given a data sample $A = ((u_1, s_1), ..., (u_n, s_n))$ where $u_i \in \{0,1\}^d$ and $s_i \in \{-1, 1\}$, and an integer $K \geq 1$. Define $e_{b,\psi}$ to be the number or errors for $(b, \psi) \in \Re^{d+1}$,

$$e_{b,\psi} = |\{i : (\psi \cdot u_i + b \leq 0, s_i = 1) \text{ or } (\psi \cdot u_i + b > 0, s_i = -1)\}|$$

Do there exist parameters b, ψ such that $e_{b,\psi} \leq K$?

Since the APPROX-HALFSPACES problem can easily be formulated as a restriction of the DMCGL$_{LM}$ problem the reduction is straightforward and so we do not present it here. ♦

We note that since DMCGL$_{LM}$ is NP–hard it is possible to create an optimization version of this problem that is very similar to MCGL$_{LM}$ that is also NP–Hard.

5.2 The Ratchet Algorithm

In this section we take a brief departure from our study of the MCGL$_{LM}$ problem to develop an algorithm called `Ratchet`. `Ratchet` is designed to (asymptotically) optimize a more general class of problems whose criteria satisfy a property we call *positive–linear–dependent* (PLD). This section defines the PLD property, develops the `Ratchet` algorithm and proves asymptotic convergence for this algorithm. Section 5.2.1 then provides sufficient conditions for a function to be PLD and Section 5.2.2 works out an important example. To realize `Ratchet` for a particular criterion we must construct a map ϕ (described below) that witnesses the PLD property for this criterion. Section 5.2.2 illustrates this by constructing such a map and proving the PLD property for a two–class weighted error criterion. Our treatment of this criterion allows us to establish a context for `Ratchet` by showing how, on this particular problem, `Ratchet` can be derived as a modification of the Pocket algorithm (Gallant, 1990).

We begin with some definitions. Let $z \in \Re^m$ and $\omega \in \Re^m$. We say that z is ω*–positive* if $\omega \cdot z > 0$. Let $\mathcal{I}$ be a countably infinite set and consider a set $Z = \{(z_1, i_1), ..., (z_n, i_n)\} \subseteq \Re^m \times \mathcal{I}$ where $z_j \in \Re^m$ and $\{i_1, ..., i_n\} \subset \mathcal{I}$ (this definition allows the set Z to have repeated values of z distinguished by their index value i). We use the abbreviated notation $Z = \{z_{i_1}, ..., z_{i_n}\}$ for this set and we call this type of set a *multisample*. In addition we call $\{i_1, ..., i_n\}$ the index set for Z. Similarly we denote the subset $\{(z_j, i_j), ..., (z_k, i_k)\} \subseteq Z$ by $\{z_{i_j}, ..., z_{i_k}\}$ and refer to it as a *subsample* of Z with index set $\{i_j, ..., i_k\} \subseteq \{i_1, ..., i_n\}$. We define $Z^+ \subseteq Z$ to be a *positive linear* (PL) subsample of Z if there exists an $\omega \in \Re^m$ such that all members of Z^+ are ω–positive, and define

$$\Omega^+ = \{\omega : \omega \cdot z_i > 0, \forall z_i \in Z^+\}$$

to be the witness set for Z^+. For technical reasons we define the empty set to be a PL multisample with the whole space as its witness set.

We consider minimization problems with criteria R that satisfy the following definition.

Definition 4. Let $\mathcal{A}$ be a set and let R be a function from $\mathcal{A} \times \Re^m$ to $\Re$. Suppose that for every $A \in \mathcal{A}$, $R_A = R(A, \cdot)$ achieves its infimum on a nontrivial set $\Omega^*(A) \subseteq \Re^m$. Then R is a *positive–linear–dependent* (PLD) function if there exists a map to multisamples $\phi : \mathcal{A} \to\to \Re^m \times \mathcal{I}$, such that for every $A \in \mathcal{A}$ there exists a PL subset of the multisample $\phi(A) = \{z_{i_1}, z_{i_2} ...\}, z_{i_j} \in \Re^m, \{i_1, i_2, ...\} \subset \mathcal{I}$ whose witness set Ω^+ satisfies $\Omega^+ \subseteq \Omega^*(A)$.

In our application of this definition to learning problems $\mathcal{A}$ is the set of all training sets, $\Re^m$ is the classifier parameter space, and R_A is an empirical error function that we wish to minimize with our choice of parameter $\omega \in \Re^m$. We consider PLD criteria because they appear in some important learning problems (e.g. the MCGL$_{LM}$ problem) and can be optimized by a very simple algorithm when a map ϕ is known. Indeed, consider the Randomized Perceptron (`RP`) algorithm acting on a multisample Z as illustrated in Algorithm 1. In the proof of Theorem 5

Algorithm 1 `Randomized Perceptron`

`INPUT`: A multisample $Z = \{z_{i_1}, z_{i_2}, ..., z_{i_n}\}$.

$k \leftarrow 0$
$\omega(0) \leftarrow 0$
loop
  $i \leftarrow$ random sample index drawn uniformly from $\{i_1, i_2, ..., i_n\}$
  if $(\omega(k) \cdot z_i \leq 0)$ **then**
    $\omega(k+1) \leftarrow \omega(k) + z_i$
  else
    $\omega(k+1) \leftarrow \omega(k)$
  end if
  $k \leftarrow k + 1$
end loop

in Appendix Appendix A: we show that with probability 1 the ω visited by `RP` witness every PL subset of Z. Thus, a simple algorithm for optimizing a PLD criterion when ϕ is known is to run `RP` on the multisample $Z = \phi(A)$, compute the criterion value R_A each time ω changes value and save the one with the smallest criterion value. We call this the `Ratchet` algorithm and it is illustrated in Algorithm 2. The following theorem establishes the optimality of this algorithm.

Theorem 5. *Let R be a PLD criterion witnessed by a map ϕ. For every $A \in \mathcal{A}$ consider the sequence $\omega(k), k = 0, 1, ...$ produced by the* `Ratchet` *algorithm with inputs A, R, ϕ. Let $\omega^*(k), k = 0, 1, ...$ be a sequence that satisfies $\omega^*(k) \in \arg\min_{\omega(i):i=0,1,...,k} R_A(\omega(i))$. Then*

$$R_A(\omega^*(k)) \overset{wp1}{\rightarrow} \min_{\omega} R_A(\omega)$$

where wp1 denotes "with probability 1".

Proof. See Appendix Appendix A:. ♦

Algorithm 2 `Ratchet`: ω^* is the ratchet parameter and ω is the parameter for the randomized perceptron algorithm.

`INPUTS`: An element $A \in \mathcal{A}$, a criterion function R, and a map ϕ

{Compute the multisample Z}
$Z = \{z_{i_1}, ..., z_{i_n}\} \leftarrow \phi(A)$

{Initialize parameters.}
Set $\omega(0)$ and ω^* to zero and set $R^* \leftarrow R_A(\omega^*)$.

{Perform the randomized perceptron algorithm and track the best solution.}
$k \leftarrow 0$
loop
  $i \leftarrow$ random sample index drawn uniformly from $\{i_1, i_2, ..., i_n\}$
  if $(\omega(k) \cdot z_i \leq 0)$ **then**
    $\omega(k+1) \leftarrow \omega(k) + z_i$
    if $(R_A(\omega(k+1)) < R^*)$ **then**
      $R^* \leftarrow R_A(\omega(k+1))$
      $\omega^* \leftarrow \omega(k+1)$
    end if
  else
    $\omega(k+1) \leftarrow \omega(k)$
  end if
  $k \leftarrow k+1$
end loop

5.2.1 Sufficient Conditions for PLD

To realize `Ratchet` for a particular criterion we must first determine that the criterion is PLD witnessed by a known map ϕ. The following lemma is often useful in establishing the PLD property once a map ϕ has been proposed.

Lemma 3. *Let $\mathcal{A}$ be a set and let R be a function from $\mathcal{A} \times \Re^m$ to $\Re$. Suppose that for every $A \in \mathcal{A}$, $R_A = R(A, \cdot)$ achieves its infimum on a nontrivial set $\Omega^*(A) \subseteq \Re^m$. Let $\phi : \mathcal{A} \twoheadrightarrow \Re^m \times \mathcal{I}$ be a map to multisamples. For $A \in \mathcal{A}$ let $Z = \phi(A) = \{z_{i_1}, ..., z_{i_n}\}, z_{i_j} \in \Re^m, \{i_1, ..., i_n\} \subset \mathcal{I}$ and let $J^+(\omega) = \{i_j : \omega \cdot z_{i_j} > 0\}$ denote the index set of ω–positive samples from Z. If for every $A \in \mathcal{A}$ and every $\omega \in \Re^m$ there exists an $\acute{\omega} \in \Re^m$ such that*

3.1. $J^+(\acute{\omega}) \supseteq J^+(\omega)$
3.2. $R_A(\acute{\omega}) = R_A(\omega)$
3.3. $\big(\omega_0, \omega_1 \in \Re^m$ *and* $J^+(\acute{\omega}_0) \supseteq J^+(\acute{\omega}_1)\big) \;\Rightarrow\; \big(R_A(\omega_0) \leq R_A(\omega_1)\big)$.

then R is PLD witnessed by ϕ.

Proof. Let $A \in \mathcal{A}$, $\omega^* \in \Omega^*(A)$, and let Ω^+ be the witness set for the samples indexed by $J^+(\acute{\omega}^*)$. For any $\omega_0 \in \Omega^+$ the relation $J^+(\omega_0) \supseteq J^+(\acute{\omega}^*)$ holds and therefore $J^+(\acute{\omega}_0) \supseteq J^+(\acute{\omega}^*)$

holds by condition 3.1. Consequently conditions 3.2 and 3.3 give

$$R_A(\omega_0) = R_A(\acute{\omega}_0) \leq R_A(\acute{\omega}^*) = R_A(\omega^*).$$

Consequently ω_0 is optimal and so $\Omega^+ \subseteq \Omega^*(A)$. ♦

5.2.2 Ratchet from Pocket

In this section we establish the PLD property for a two–class weighted error criterion and show how the `Ratchet` algorithm for this criterion can be derived as a modification of the `Pocket` algorithm (Gallant, 1990).

Gallant introduced `Pocket` for the problem of minimizing the empirical error of a linear classifier. This problem is the optimization version of the APPROX-HALFSPACES problem in Definition 3 with u extended to $\Re^d$. If we let $\xi_i = (1, u_i)$, $\omega = (b, \psi)$ and $A = ((\xi_1, s_1), ..., (\xi_n, s_n)) \in (\mathcal{X}_1 \times \{-1, 1\})^n$ then the criterion to be minimized is defined by

$$R_A(\omega) = \sum_{i=1}^{n} I(\omega \cdot \xi_i > 0, s_i = -1) + I(\omega \cdot \xi_i \leq 0, s_i = 1)$$

where $I(\cdot)$ is the indicator function that takes a value 1 when its argument is true and 0 otherwise. A nonzero contribution from the i–th component of this sum represents an error for the linear classifier defined by $sign(\omega \cdot \xi)$ (where $sign(0) = -1$) on the sample (ξ_i, s_i). The `Pocket` algorithm operates by running the `RP` algorithm on the multisample $Z = \{z_1, ..., z_n\}, z_i = s_i\xi_i$, computing the *run length* for each ω visited (i.e. the number of consecutive ω–positive samples encountered before ω is modified by the algorithm), and retaining the $\omega(k)$ with the largest run length in the "pocket". Gallant also introduces a variation called `Pocket-with-Ratchet` that places a new value of ω in the pocket only when it has both a larger run length and witnesses a smaller criterion value. These `Pocket` algorithms are attractive because the run length is very simple to compute, but they may not be appropriate for other criteria. For example consider the weighted error criterion $\bar{R}$ defined by

$$\bar{R}_A(\omega) = \sum_{i=1}^{n} c_{-1} I(\omega \cdot \xi_i > 0, s_i = -1) + c_1 I(\omega \cdot \xi_i \leq 0, s_i = 1)$$

where c_{-1}, c_1 are the costs for the two types of classification error, and consider the obvious adaptation of the `Pocket-with-Ratchet` algorithm that operates on the same multisample Z and replaces the value of ω in the pocket when the run length is larger and the criterion value $\bar{R}_A(\omega)$ is smaller. With $c_{-1} = c_1$ the criterion $\bar{R}_A = R_A$ is minimized when the number of positive samples in Z is maximized and so values of ω with larger run lengths are more likely to have smaller criterion values, but this is not necessarily true when $c_{-1} \neq c_1$. In fact it seems unlikely that any statistic computed on ω–positive samples only can be used to order Ω when $c_{-1} \neq c_1$. More generally the determination of a suitable replacement for the run length rule remains an open problem. The `Ratchet` algorithm is obtained by removing the run length rule from `Pocket-with-Ratchet` so that a value of ω with the smallest criterion value is saved in the pocket. This requires that the criterion value be computed each time ω is modified and therefore requires more computation than the `Pocket` algorithms, but it yields a viable algorithm. Indeed, Lemma 4 below verifies that the criterion $\bar{R}$ is PLD witnessed by a map that gives $z_i = s_i\xi_i$, and is therefore optimized asymptotically (wp1) by the realization of `Ratchet` just discussed.

Lemma 4. *Let $\bar{R}$ be a function from $\mathcal{A} \times \Re^{d+1}$ to $\Re$ where $\mathcal{A} = (\mathcal{X}_1 \times \{-1,1\})^n$. For any $A = ((\xi_1, s_1), ..., (\xi_n, s_n)) \in \mathcal{A}$ where $\xi_i \in \mathcal{X}_1$, $s_i \in \{-1,1\}$ let $\bar{R}_A = \bar{R}(A, \cdot)$ be defined by*

$$\bar{R}_A(\omega) = \sum_{i=1}^{n} c_{-1} I(\omega \cdot \xi_i > 0, s_i = -1) + c_1 I(\omega \cdot \xi_i \leq 0, s_i = 1)$$

where $0 < c_{-1}, c_1 < \infty$. Let $\mathcal{N}$ be the set of natural numbers. Then $\bar{R}$ is PLD witnessed by the map $\phi : \mathcal{A} \to\to \big(\{-1,1\} \times \mathcal{X}\big) \times \mathcal{N}$ defined by $\phi(A) = \{z_1, ..., z_n\}, z_i = s_i\xi_i$.

Proof. See Appendix Appendix A:. ♦

5.3 The Generalized M-Class Ratchet Algorithm

In this section we show how the `Ratchet` algorithm can be used to solve the MCGL$_{LM}$ optimization problem. We do this by constructing a map ϕ that witnesses the PLD property of the MCGL$_{LM}$ criterion. The map ϕ is then used to determine a realization of `Ratchet` for the MCGL$_{LM}$ problem.

In the MCGL$_{LM}$ optimization problem we minimize the criterion defined in (5) over the class of linear machines defined at the beginning of Section 5. Recall that for this class of functions the criterion is defined on the data $((\xi_1, y_1), ..., (\xi_n, y_n))$ which is obtained from the original data $((x_1, y_1), ..., (x_n, y_n))$ by applying the map $x \mapsto (1, x)$ to each x_i. The criterion R is defined by

$$R_{(\Xi,Y)}(\omega) = \frac{1}{n}\sum_{i=1}^{n} y_i^{f_\omega(\xi_i)} \tag{16}$$

where $\Xi = (\xi_1, ..., \xi_n) \in \mathcal{X}_1^n$, $Y = (y_1, ..., y_n) \in ([0,1]^M)^n$, $\omega = (w_1, w_2, ..., w_M) \in \Re^{M(d+1)}$ and the functions f_ω are defined by (14).

To prove that R is PLD we construct a map ϕ that witnesses this property. We show that under ϕ the data sample Ξ maps to a multisample $Z = \{z_1, ..., z_{nM(M-1)}\}, z_i \in \mathcal{X}_1^M$ such that for every $\omega \in \Re^{M(d+1)}$ there exists an $\acute{\omega} \in \Re^{M(d+1)}$ such that the value $R_{(\Xi,Y)}(\omega)$ of the criterion $R_{(\Xi,Y)}$ is equal to $R_{(\Xi,Y)}(\acute{\omega})$ and is determined by a subsample of the $\acute{\omega}$–positive samples from Z, i.e. a PL subsample of Z. Thus there exists a collection of PL subsamples of Z that determine all values of $R_{(\Xi,Y)}$ and $R_{(\Xi,Y)}$ is constant on any of their witness sets. The PLD property follows directly.

We begin by defining a map ϕ which is an extension of Kesler's construction for the multiclass problem (see p. 266 in (Duda *et al.*, 2000), pp. 87–93 in (Nilsson, 1990), and (Smith, 1969)).

Definition 5. Let $\rho : \mathcal{X}_1 \to \mathcal{Z}^{M(M-1)}, \mathcal{Z} = \mathcal{X}_1^M$ be the function that maps a point ξ to a point $(..., z_{jk}, ...)$, $1 \leq j \leq M$, $k : 1 \leq k \leq M, k \neq j$ where $z_{jk} \in \mathcal{Z}$ is the vector obtained by concatenating M vectors as follows: ξ is placed in the j–th position, $-\xi$ in the k–th position, and zero vectors are placed in the other $M-2$ positions as illustrated below,

$$z_{jk} = (0...0 \underbrace{\xi}_{j^{th}} 0...0 \underbrace{-\xi}_{k^{th}} 0...0)$$

Now let $\gamma : \mathcal{X}_1^n \to \mathcal{Z}^{nM(M-1)}$ be the map defined by

$$\gamma(\Xi) = \big(\rho(\xi_1), ..., \rho(\xi_n)\big).$$

We adopt the notation z_{ijk} for the jk–th member of $\rho(\xi_i)$ so that

$$\gamma(\Xi) = (..., z_{ijk}, ...).$$

The map $\phi : \mathcal{X}_1^n \to\to \mathcal{Z} \times \mathcal{N}^3$ is then defined as the composition of γ with a map to multisamples so that $\phi(\Xi) = \{..., (z_{ijk}, ijk), ...\}$, or with our abbreviated notation

$$\phi(\Xi) = \{..., z_{ijk}, ...\}.$$

Our next step is to develop an expression for the criterion value $R_{(\Xi,Y)}(\omega)$ in terms of basic operations on the mapped samples z_{ijk}. We accomplish this by developing an expression for the values of $f_\omega|_\Xi$ and substituting into (16). The following properties are useful in this regard and are easily verified for all $1 \le i \le n$, $1 \le j \le M$, $k : 1 \le k \le M, k \ne j$ and any $a = (a_1, a_2, ..., a_M)$ where $a_i \in \Re^{d+1}$.

1. <u>pairwise comparison</u>

$$\begin{aligned}
\Big(a_j \cdot \xi_i > a_k \cdot \xi_i\Big) &\text{ if and only if } \Big(a \cdot z_{ijk} > 0\Big) \\
\Big(a_j \cdot \xi_i < a_k \cdot \xi_i\Big) &\text{ if and only if } \Big(a \cdot z_{ijk} < 0\Big) \\
\Big(a_j \cdot \xi_i = a_k \cdot \xi_i\Big) &\text{ if and only if } \Big(a \cdot z_{ijk} = 0\Big)
\end{aligned} \tag{17}$$

2. <u>winner–take–all with ties</u>
Recall from (15) that $\mathcal{I}_a(\xi) = \arg\max_k a_k \cdot \xi$ and let us define $T_a(i) = \{j : \min_k a \cdot z_{ijk} = 0\}$. Then

$$\Big(|\mathcal{I}_a(\xi_i)| > 1\Big) \text{ if and only if } \Big(\max_j \min_k a \cdot z_{ijk} = 0\Big) \text{ if and only if } \Big(T_a(i) = \mathcal{I}_a(\xi_i)\Big) \tag{18}$$

3. <u>winner–take–all without ties</u>

$$\big(\mathcal{I}_a(\xi_i) = \{j\}\big) \text{ if and only if } \Big(a \cdot z_{ijk} > 0, \forall k \ne j\Big) \text{ if and only if } \Big(T_a(i) = \emptyset\Big) \tag{19}$$

4. <u>symmetry</u>

$$z_{ijk} = -z_{ikj} \tag{20}$$

Using properties (18)-(19) in (14)

$$\begin{aligned} f_\omega(\xi_i) &= \max_{j \in \mathcal{I}_\omega(\xi_i)} j \\ &= \sum_{j=1}^{M} j \left(I\big(\omega \cdot z_{ijk} > 0, \forall k \neq j\big) + I\left(j = \max_{l \in T_\omega(i)} l\right)\right). \end{aligned} \tag{21}$$

The first indicator function treats the case where there are no ties and the second treats ties. The criterion value in (16) can be expressed

$$R_{(\Xi,Y)}(\omega) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{M} y_i^j I(f_\omega(\xi_i) = j)$$

and substituting the expression for f_ω from (21) gives

$$R_{(\Xi,Y)}(\omega) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{M} y_i^j \left(I\big(\omega \cdot z_{ijk} > 0, \forall k \neq j\big) + I\left(j = \max_{l \in T_\omega(i)} l\right)\right). \tag{22}$$

To prove that R is PLD we verify conditions 3.1-3.3 in Lemma 3. We begin with a lemma that confirms that for every $\omega \in \Re^{M(d+1)}$ there exists an $\acute{\omega} \in \Re^{M(d+1)}$ without ties on Ξ that satisfies conditions 3.1-3.2.

Lemma 5. *Let* $(\Xi, Y) \in \left(\mathcal{X}_1 \times [0,1]^M\right)^n$ *and let* $Z = \phi(\Xi)$ *where* ϕ *is given by Definition 5. Let* $R_{(\Xi,Y)}$ *be the criterion function defined by (22). For every* $\omega \in \Re^{M(d+1)}$ *there exists an* $\acute{\omega} \in \Re^{M(d+1)}$ *such that following properties hold;*

5.1. $T_{\acute{\omega}}(i) = \emptyset, \forall i$

5.2. $J^+(\acute{\omega}) \supseteq J^+(\omega)$

5.3. $R_{(\Xi,Y)}(\acute{\omega}) = R_{(\Xi,Y)}(\omega)$.

Proof. Let $\omega = (w_1, w_2, ..., w_M) \in \Re^{M(d+1)}$ and construct $\acute{\omega} = (\acute{w}_1, \acute{w}_2, ..., \acute{w}_M)$ by adding a small positive constant to each of the M offset parameters of ω, i.e.

$$\acute{w}_j = w_j + (\delta_j, 0), \;\; 0 \in \Re^d, \;\;\; 1 \leq j \leq M$$

where δ_j is defined as follows. Let $Z_1 = \{z_{ijk} : z_{ijk} \in Z, |\omega \cdot z_{ijk}| > 0\}$ and let

$$\delta = \begin{cases} \min_{z_{ijk} \in Z_1} |\omega \cdot z_{ijk}| & |Z_1| > 0 \\ 1 & |Z_1| = 0 \;. \end{cases}$$

Now define

$$\delta_j = \left(\frac{j-1}{M}\right)\delta, \;\; 1 \leq j \leq M.$$

This gives

$$\begin{aligned}\acute{\omega} \cdot z_{ijk} &= \acute{w}_j \cdot \xi_i - \acute{w}_k \cdot \xi_i \\ &= w_j \cdot \xi_i - w_k \cdot \xi_i + \delta_j - \delta_k \\ &= \omega \cdot z_{ijk} + \delta_j - \delta_k.\end{aligned}$$

Since $|\delta_j - \delta_k| < \delta$ it follows that $\omega \cdot z_{ijk} > 0 \Rightarrow \acute{\omega} \cdot z_{ijk} > 0$ and $\omega \cdot z_{ijk} < 0 \Rightarrow \acute{\omega} \cdot z_{ijk} < 0$, and since $j > k \Leftrightarrow \delta_j > \delta_k$ it follows that $(\omega \cdot z_{ijk} = 0$ and $j > k) \Rightarrow \acute{\omega} \cdot z_{ijk} > 0$ and $(\omega \cdot z_{ijk} = 0$ and $j < k) \Rightarrow \acute{\omega} \cdot z_{ijk} < 0$. This verifies properties 5.1 and 5.2, and also verifies that $\acute{\omega}$ makes the same contribution as ω to the sum in (22) when there is no tie. Thus, to verify property 5.3 all that remains is to show that $\acute{\omega}$ makes the same contribution as ω to the sum in (22) when there is a tie. For $j \in T_\omega(i)$ property (18) implies $\omega \cdot z_{ijk} > 0, \forall k \notin T_\omega(i)$ and we have just shown that this implies $\acute{\omega} \cdot z_{ijk} > 0, \forall k \notin T_\omega(i)$. Also since $j > k \Leftrightarrow \delta_j > \delta_k$ it follows that for the winner j^* of a tie $(\omega \cdot z_{ij^*k} = 0$ and $j^* = \max_{l \in T_\omega(i)} l) \Rightarrow (\acute{\omega} \cdot z_{ij^*k} > 0, \forall k \in T_\omega(i), k \neq j^*)$ from which we conclude that $\acute{\omega} \cdot z_{ij^*k} > 0, \forall k \neq j^*$ and our proof is complete. ♦

Properties 5.1 and 5.3 allow us to rewrite (22) as

$$R_{(\Xi,Y)}(\omega) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{M} y_i^j I(\acute{\omega} \cdot z_{ijk} > 0, \forall k \neq j) \tag{23}$$

which gives the criterion value in terms of the subsample of Z with index set $J^+(\acute{\omega}) = \{ijk : \acute{\omega} \cdot z_{ijk} > 0\}$. However, to verify condition 3.3 we must look more closely at how the criterion value is affected by the structure of this index set.

Let $\epsilon > 0$ and define

$$\Delta_{ijk} = \left\{ \begin{array}{ll} \epsilon, & y_i^j = y_i^k \\ y_i^k - y_i^j, & \text{otherwise} \end{array} \right., \quad 1 \leq i \leq n, 1 \leq j \leq M, k : 1 \leq k \leq M, k \neq j. \tag{24}$$

Define the *target* index sets

$$J_{ij} = \{ijk : \Delta_{ijk} > 0, 1 \leq k \leq M, k \neq j\}, \quad 1 \leq i \leq n, 1 \leq j \leq M \tag{25}$$

Let the $\acute{\omega}$–positive subsets of these sets be

$$J_{ij}^+(\acute{\omega}) = \{ijk : ijk \in J_{ij} \text{ and } \acute{\omega} \cdot z_{ijk} > 0\}$$

and let

$$J_i^+(\acute{\omega}) = \cup_{j=1}^{M} J_{ij}^+(\acute{\omega}). \tag{26}$$

The following lemma establishes properties of these index sets that are used to verify condition 3.3.

Lemma 6. *Consider the definitions in (24)-(26) with $\acute{\omega}$ satisfying Lemma 5. The following properties hold for all $i = 1, 2, ..., n$.*

6.1. *if* $\big(\Delta_{ijl} > 0$ *and* $\Delta_{ilk} > 0\big)$ *then* $\big(\Delta_{ijk} > 0\big)$

6.2. $\Big(|J_{il}| \geq |J_{im}|\Big)$ *if and only if* $\Big(y_i^l \leq y_i^m\Big)$

6.3. $\Big(j^* = \arg\max_{j : J_{ij} \subseteq J_i^+(\acute{\omega})} |J_{ij}|\Big)$ *if and only if* $\Big(\acute{\omega} \cdot z_{ij^*k} > 0, \forall k \neq j^*\Big)$.

Proof. We start with Property 6.1. If $y_i^j = y_i^k$ then by definition $\Delta_{ijk} = \epsilon > 0$ so assume that $y_i^j \neq y_i^k$. In this case we can write

$$\Delta_{ijk} = y_i^k - y_i^j = (y_i^l - y_i^j) + (y_i^k - y_i^l)$$

and the assumptions $\Delta_{ijl} > 0$ and $\Delta_{ilk} > 0$ imply that both terms on the right are ≥ 0. Also since $y_i^j \neq y_i^k$ at least one of them is strictly positive which completes the proof of Property 6.1.

The definitions in (24) and (25) insure that the number of elements in J_{il} equals the number of components of y_i that are greater than or equal to y_i^l, i.e.

$$|J_{il}| = |\{k : y_i^k \geq y_i^l\}|$$

Property 6.2 follows directly.

Now we prove Property 6.3. The left–hand expression for j^* is legitimate only if $J_i^+(\acute{\omega})$ contains at least one target set and the largest target set it contains is unique. We begin by showing that it contains at least one target set. Let l^* be the distinct index for which $\acute{\omega} \cdot z_{il^*k} > 0$ for all $k \neq l^*$. This implies that $J_{il^*}^+(\acute{\omega}) = J_{il^*}$ and therefore that $J_{il^*} \subseteq J_i^+(\acute{\omega})$. Note that this includes the case where $J_{il^*} = J_i^+(\acute{\omega}) = \emptyset$ so that the largest subset of $J_i^+(\acute{\omega})$ has size 0. The fact that the largest target set contained in $J_i^+(\acute{\omega})$ is unique follows from the uniqueness of l^* and the remainder of our proof.

Let $j^* \in \arg\max_{j:J_{ij} \subseteq J_i^+(\acute{\omega})} |J_{ij}|$ and suppose that $l^* \neq j^*$. Then $\acute{\omega} \cdot z_{il^*j^*} > 0$ and from symmetry (20) it follows that $\acute{\omega} \cdot z_{ij^*l^*} < 0$, and since $J_{ij^*} \subseteq J_i^+(\acute{\omega})$ it follows that $l^* \notin J_{ij^*}$ so that $\Delta_{ij^*l^*} < 0$. Therefore the definition of Δ in (24) implies that $\Delta_{il^*j^*} > 0$. Consider an index $ij^*k \in J_{ij^*}$. Since $\Delta_{ij^*k} > 0$ and $\Delta_{il^*j^*} > 0$ it follows from Property 6.1 that $\Delta_{il^*k} > 0$ and so $il^*k \in J_{il^*}$ so that $|J_{ij^*}| \leq |J_{il^*}|$. However since $\acute{\omega} \cdot z_{il^*j^*} > 0$ and $\Delta_{il^*j^*} > 0$ the set J_{il^*} contains the index il^*j^* which is not in J_{ij^*} so that $|J_{ij^*}| < |J_{il^*}|$. Since $J_{il^*} = J_{il^*}^+(\acute{\omega}) \subseteq J_i^+(\acute{\omega})$ this contradicts the definition of j^*. ♦

We can now state and prove the main theorem.

Theorem 6. *The $MCGL_{LM}$ criterion defined by (16) is PLD witnessed by ϕ in Definition 5.*

Proof. We need only verify the conditions in Lemma 3. For any (Ξ, Y) and any $\omega \in \Re^{M(d+1)}$ the criterion value $R_{(\Xi,Y)}(\omega)$ is a finite sum and therefore $R_{(\Xi,Y)}$ achieves its infimum on a nontrivial subset $\Omega^*((\Xi, Y)) \subseteq \Re^{M(d+1)}$. Lemma 5 verifies conditions 3.1 and 3.2. To finish the proof we must verify condition 3.3, i.e.

$$\big(\omega_0, \omega_1 \in \Re^{M(d+1)} \text{ and } J^+(\acute{\omega}_0) \supseteq J^+(\acute{\omega}_1)\big) \;\Rightarrow\; \big(R_{(\Xi,Y)}(\omega_0) \leq R_{(\Xi,Y)}(\omega_1)\big)$$

Applying Property 6.3 to (23) gives

$$R_{(\Xi,Y)}(\omega) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{M} y_i^j I(j = \arg\max_{l:J_{il} \subseteq J_i^+(\acute{\omega})} |J_{il}|). \tag{27}$$

Let $j_i = \arg\max_{l:J_{il}\subseteq J_i^+(\acute{\omega}_1)} |J_{il}|$. If $J^+(\acute{\omega}_0) \supseteq J^+(\acute{\omega}_1)$ then since $J^+ = \cup_i J_i^+$ is a disjoint union $J_i^+(\acute{\omega}_0) \supseteq J_i^+(\acute{\omega}_1)$ and so by (27) $R_{(\Xi,Y)}(\omega_0)$ can deviate from $R_{(\Xi,Y)}(\omega_1)$ only if $J^+(\acute{\omega}_0)$ contains additional target sets J_{il} where $|J_{il}| > |J_{ij_i}|$ for one or more i. If $J^+(\acute{\omega}_0)$ contains additional target sets that satisfy this condition then by Property 6.2 and (27) we have $R_{(\Xi,Y)}(\omega_0) \leq R_{(\Xi,Y)}(\omega_1)$ and our proof is complete. ♦

As a consequence of Theorem 6 we can apply the `Ratchet` algorithm with arguments (Ξ, Y), R and ϕ to provide a solution to the MCGL$_{LM}$ criterion. The following observations lead to a more efficient practical implementation. First it is possible to prove that the PLD property is also witnessed by a map $\bar{\phi}$ that yields the multisample $\bar{Z} \subseteq Z$ defined by

$$\bar{Z} = \{z_{ijk} : z_{ijk} \in Z, \Delta_{ijk} > 0\}.$$

Although we do not provide the proof, it it easy to see that this will not affect the solution since from (27) it is clear that criterion values can be expressed exclusively in terms of these samples. Second, since multiplying each point in $\bar{Z}$ by a positive scalar has no effect on its PL subsamples it is possible to prove that the PLD property is also witnessed by a map $\bar{\phi}'$ that yields the multisample

$$\bar{Z}' = \{z'_{ijk} : z'_{ijk} = \Delta_{ijk} z_{ijk}, z_{ijk} \in \bar{Z}\}.$$

Applying the `Ratchet` algorithm to this set leads to the *Generalized Multi–class Ratchet* (`GMR`) algorithm in Algorithm 3. We present this algorithm in terms of its operation on the original data $((\xi_1, y_1), ..., (\xi_n, y_n))$, i.e. the map $\bar{\phi}'$ is implemented implicitly.

6 Experimental Results

In this section we present the results of experiments with a collection of documents called the Q7 corpus. The Q7 corpus consists of 1445 printed documents (varying in size from 106 to 6290 characters) that were digitized and restored using 9 different methods. Each original document, along with its 9 restored versions, was converted to an ASCII text file using an existing OCR system. Then the character error rate for each of the 14,450 ASCII text files was determined manually. In addition, each digitized document is represented by a $d = 7$ dimensional feature vector when presented to the classifier. The components of these feature vectors represent 7 different *document image quality measures* designed to quantify notions like “speckle”, “broken characters”, “touching characters”, etc. Each quality measure is a real number in the range $[0, 1]$ and is designed so that smaller values represent better quality. A detailed description of the methodology used to design the 9 restoration methods and the 7 quality measures can be found in (Cannon *et al.*, 1999) (although the specific restoration methods and quality measures designed in (Cannon *et al.*, 1999) are different from those used here). In summary the Q7 corpus is a data sample $D_n = ((x_1, y_1), ..., (x_n, y_n))$ with $n = 1445$ samples, where $x \in [0, 1]^d$ with $d = 7$, and $y \in [0, 1]^M$ with $M = 10$ (9 restorations and 1 original).

Table 1 shows the average OCR error rates (averaged over the corpus) for the unrestored corpus and for the corpus restored using each of the 9 methods. It also shows that average OCR error rates when the corpus is restored using the method that gives the best and worst

Algorithm 3 `GMR`: Generalized M-Class Ratchet Algorithm. ω are the ratchet weights and $(w_1, w_2, ..., w_M)$ are the weights for randomized perceptron algorithm.

`INPUTS:` $MaxIter$, and a data sample $((\xi_1, y_1), ..., (\xi_n, y_n))$
`OUTPUT:` ω

{Initialization: (e_{min} is a lower bound on the error).}
$\omega \leftarrow 0, \quad (w_1, ..., w_M) \leftarrow (0, ..., 0)$
$e_\omega \leftarrow \frac{1}{n} \sum_j y_j^{f_\omega(\xi_j)}$
$\epsilon \leftarrow \min\left(\frac{1}{n},\ \min_{ijk: y_i^j \neq y_i^k} |y_i^k - y_i^j|\right)$
$e_{min} = \frac{1}{n} \sum_{i=1}^{n} \min_c y_i^c$

{Perform the randomized multi-class perceptron algorithm and track the best solution.}
for $iter = 1$ to $MaxIter$ **do**
  $i \leftarrow$ random sample index drawn uniformly from $\{1, 2, ..., n\}$
  $(j, k) \leftarrow$ random index pair drawn uniformly from $\{(l, m) : 1 \leq l, m \leq M, l \neq m\}$
  if $(y_i^k = y_i^j)$ **then**
    $\Delta = \epsilon$
  else
    $\Delta \leftarrow y_i^k - y_i^j$
  end if
  if $(\Delta(w_j - w_k) \cdot \xi_i \leq 0)$ **then**
    $w_j \leftarrow w_j + \Delta\xi_i$
    $w_k \leftarrow w_k - \Delta\xi_i$
    if $\left(\frac{1}{n} \sum_j y_j^{f_{(w_1,...,w_M)}(\xi_j)} < e_\omega\right)$ **then**
      $e_\omega \leftarrow \frac{1}{n} \sum_j y_j^{f_{(w_1,...,w_M)}(\xi_j)}$
      $\omega \leftarrow (w_1, ..., w_M)$
    end if
    if $(e_\omega = e_{min})$ **then**
      return(ω)
    end if
  end if
end for
return(ω)

OCR error rate for each document. The entries in this table represent estimates of error rates for a general document population characterized by a fixed (but unknown) distribution. Each restoration method is designed for a specific type of distortion and when applied to documents without that distortion can actually degrade the document quality resulting in an increased OCR error rate. On this corpus, no one of the 9 restoration methods provides an improvement in error rate over that for the unrestored documents. In fact, in several cases the error rate is significantly worse. On the other hand, each of the 9 methods leads to an improved OCR error rate for a significant fraction of the 1445 documents as shown in the third column of the table. In total, the OCR error rate can be improved for 81% of the documents using at least

Type of Restoration	**Average Error Rate for Corpus**	**% Documents with Improved OCR Error**
No restoration	.1105	-
Method 1	.1127	41
Method 2	.3150	22
Method 3	.4132	19
Method 4	.1206	32
Method 5	.2082	31
Method 6	.1139	45
Method 7	.1180	29
Method 8	.1151	22
Method 9	.1352	30
Best	.0810	81
Worst	.4740	0

Table 1: Summary of Q7 Corpus Statistics

one of the 9 restoration methods (as indicated by the value in the third column for "Best"). The extent to which these improvements can be realized in practice however depends on the accuracy with which we can chose an appropriate restoration technique.

The average OCR error rate for unrestored documents is approximately .1105. If a perfect classifier exists under P, i.e. one that can always choose the best restoration method based on the document representation x, then the result in Table 1 indicates that the average error rate can be reduced to approximately .081. This then serves as an estimate of the lower bound on the achievable error rate. At the other extreme, if a classifier exists that can always choose the worst restoration method then the result in Table 1 indicates that the average error rate would climb to approximately .474. This serves as an estimate of the upper bound on error rate. We consider both these bounds to be loose, since it is unlikely that either a perfect or worst–case classifier exists. Nevertheless, they provide useful insight. For example, the fact that the average OCR error rate for the unrestored corpus is approximately .36 below the upper bound but only about .03 above the lower bound suggests that it may be difficult to find classifiers that give improved OCR error rates for this corpus.

We now describe experimental results for the following classifier design methods.

kNN: The k–nearest neighbor method described in Section 3 with the Euclidean metric.

LM: An M–class linear machine trained with `GMR` to minimize the empirical error e_n defined in (5).

QM: An M–class quadratic machine trained with `GMR` to minimize the empirical error e_n defined in (5). This method is identical to the linear machine above expect that the 7–dimensional feature vectors $x = (x^1, ..., x^7)$ are extended to 35 dimensions by augmenting them with all second order terms formed from the pairwise products of the original 7 features.

LM01: An M–class linear machine trained to minimize "classification error". With this method each sample is considered to be "misclassified" if it is assigned a restoration method other than one that gives a minimum OCR error rate. The goal in the training process is to produce a classifier that chooses a restoration method that gives the minimum individual OCR error rate for as many samples as possible. This corresponds to treating our problem as a traditional M–class classifier design problem with zero–one loss. To use the `GMR` algorithm let

$$c_i \in \arg\min_{c \in \mathcal{C}} y_i^c$$

and then set the c_i–th component of y_i to 0 and the rest to 1, i.e.

$$y_i = (11...1 \underbrace{0}_{c^i} 1...1)$$

This loss function makes no distinction between restoration methods that do not give the minimal OCR error rate for individual samples.

QM01: An M–class quadratic machine trained with `GMR` to minimize the same classification error as LM01 above.

MV01: An M–class classifier formed as a majority vote of $M(M-1)/2$ 2–class classifiers, each trained with the `Pocket` algorithm to minimize the 2–class classification error (zero–one loss). This was the method applied to this problem previously in (Cannon *et al.*, 1999).

In all cases the results reported are an average of ten 10–fold cross validation runs. The number of iterations used in the `GMR` and `Pocket` algorithms is 10,000,000. For the kNN method the value of k was chosen for each cross validation run using a second 10–fold cross validation on the training subsample for each $k \in \{1, 2, ...30\}$ and choosing the value with the smallest error estimate.

Table 2 summarizes the average OCR error estimates for the classifier design methods. Earlier we mentioned that there is little room for improvement with this corpus (less than approximately .03). In addition the error rate estimates obtained with this size of corpus not accurate enough to distinguish between values that differ by less than .03 with high confidence (e.g. see the estimated standard deviation of the OCR error estimates in the third column of the table). Thus our interpretation of the results in Table 2 are made with this caveat. With this in mind we note that all methods provide improved estimates over no restoration. In addition all the methods introduced in this paper give improved estimates over the MV01 approach taken previously. The best error estimate is reported for the Linear Machine (LM), although the kNN and Quadratic Machine (QM) have similar performance. The Quadratic Machine (QM) performance is slightly worse than the Linear Machine (LM) suggesting that it may suffer from overfitting. Methods trained with the zero–one loss (LM01 and QM01) perform consistently worse than their counterparts trained with the generalized loss (LM and QM). As expected however, the average classification errors for LM01 and QM01 are less than for LM and QM respectively. Specifically, the average classification error estimates are LM01=.767, LM=.778. QM01=.756, QM=.777. This demonstrates that an optimal classifier for the document restoration problem is not achieved by optimizing classification error. On the other hand, classifiers designed to optimize classification error (MV01, LM01 and QM01) appear to do surprisingly well on this corpus. Finally we report that the estimate of the fraction of documents whose OCR error rate is improved is 0.49 for all four methods LM, QM, LM01 and QM01. This is .32 below the upper bound of .81 in Table 1.

Method	**Average OCR Error Rate**	**Standard Deviation**
No Restoration	.1105	.014
MV01	.1018	-
kNN	.0983	.012
LM	.0977	.011
QM	.0988	.013
LM01	.1019	.013
QM01	.1002	.013

Table 2: Comparison of methods. The third column gives the standard deviation estimate for the average OCR error rate estimates.

7 Acknowledgements

We would like to thank James Theiler and Adam Cannon for their constructive feedback on earlier versions of this document.

Appendix A: Proofs

Proof of Theorem 5. Our proof is motivated by the analysis of the `Pocket` algorithm in (Muselli, 1997). The foundation for this proof hinges on two well-known results for the randomized perceptron algorithm. First, the values of ω visited by the randomized perceptron algorithm remain bounded, and second if the sequence of samples drawn in the main loop includes a specific type of subsequence exclusively from a PL multisample Z^+ then the algorithm will produce a value of ω from the witness set Ω^+. Our proof proceeds by showing that for any PL multisample Z^+, and in particular for an optimal Z^+, the sequence of samples drawn in the main loop includes such a subsequence wp1.

Let $Z = \phi(A)$ and let D satisfy $|z_j| < D, \forall z_j \in Z$. The perceptron cycling theorem guarantees that $|\omega(k)|$ remains bounded for all iterations (Block & Levin, 1970). In particular this result is true (and nontrivial) in cases where the algorithm iterates indefinitely. Such cases occur when there is no value of ω for which all samples in Z are ω–positive. Let B be this bound when `RP` is initialized with $\omega(0) = 0$.

Consider the collection of PL subsamples $Z_i^+, i = 1, 2, ...$ of Z and let $\Omega_i^+, i = 1, 2, ...$ be their witness sets. Define the *margin* for subsample Z_i^+ to be

$$\rho_i = \max_{\omega \in \Omega_i^+} \min_{z_j \in Z_i^+} \frac{\omega \cdot z_j}{|\omega|}.$$

Let $r_i = |Z_i^+|$ and consider a sequence of samples from Z_i^+ of length r_i in which each point from Z_i^+ appears exactly once. We define a *cyclic* sequence of length κ from Z_i^+ to be the concatenation of $\lceil \kappa / r_i \rceil$ such sequences truncated to length κ. The perceptron convergence theorem guarantees that when the RP algorithm encounters a cyclic sequence of length $\kappa = \lfloor \frac{D^2 + |\omega(k)|^2}{\rho_i^2} \rfloor$ from Z_i^+ beginning at iteration k then $\omega(k + \kappa) \in \Omega_i^+$ (Novikoff, 1962; Vapnik, 1998). Let

$$\kappa_i = \lfloor \frac{D^2 + B^2}{\rho_i^2} \rfloor.$$

Since B bounds the size of any $\omega(k)$, κ_i represents a lower bound on the number of iterations sufficient to guarantee that a member of Ω_i^+ is produced when a cyclic sequence from Z_i^+ is encountered during the algorithm.

Since R is PLD there exists a PL subset of Z whose witness set witnesses the minimal value of R. Let i^* be the index of such an optimal PL subsample and define the event

A**:** a cyclic sequence of length $l \geq \kappa_{i^*}$ from $Z_{i^*}^+$ is presented to `Ratchet` during the first k iterations.

and its complement

$\bar{A}$**:** no cyclic sequence of length $l \geq \kappa_{i^*}$ from $Z_{i^*}^+$ is presented to `Ratchet` during the first k iterations.

Let $R^* = \min_\omega R_A(\omega)$ and $R_k = R_A(\omega^*(k))$. Let $\epsilon > 0$ be so small that

$$\{R_A(\omega) \neq R^*\} = \{R_A(\omega) - R^* > \epsilon\}$$

Then

$$\begin{aligned} P(R_k - R^* > \epsilon) \quad &= P(R_k \neq R^*) \\ &= 1 - P(R_k = R^*) \\ &\leq 1 - P(A) \\ &= P(\bar{A}) \end{aligned} \tag{28}$$

Define the event

B**:** an *i.i.d.* random sequence of length κ_{i^*} from Z is a cyclic sequence from $Z_{i^*}^+$

Since the total number of sequences of length κ_{i^*} is finite the probability of this event is strictly greater than zero. Let $q_{i^*} > 0$ denote this probability.

Now consider a sequence of length $k \geq \kappa_{i^*}$ and partition it into adjacent (nonoverlapping) subsequences of length κ_{i^*}. Let σ be the set of $\lfloor k/\kappa_{i^*} \rfloor$ subsequences of length κ_{i^*} formed in this way. If there is no cyclic sequence of length $l \geq \kappa_{i^*}$ from $Z_{i^*}^+$ in the first k iterations then there will be no cyclic sequence of length κ_{i^*} from $Z_{i^*}^+$ in σ, and since the sequences in σ are independent

$$P(\bar{A}) \leq (1 - q_{i^*})^{\lfloor k/\kappa_{i^*} \rfloor}.$$

Substituting into (28) gives

$$P(R_k - R^* > \epsilon) \leq (1 - q_{i^*})^{\lfloor k/\kappa_{i^*} \rfloor}.$$

Since κ_{i^*} is finite and q_{i^*} is strictly greater than zero

$$\sum_{k=1}^{\infty} P(R_k - R^* > \epsilon) \leq \sum_{k=1}^{\infty} (1 - q_{i^*})^{\lfloor k/\kappa_{i^*} \rfloor} < \infty.$$

This implies that (e.g. see (Serfling, 1980), p.10)

$$P(\lim_{k \to \infty} R_k = R^*) = 1$$

and the proof is finished. ♦

Proof of Lemma 4. For any $\omega \in \Re^{d+1}$ the criterion value is a finite sum and therefore the criterion achieves its infimum on a nontrivial set $\Omega^*(A) \subseteq \Re^{d+1}$. If we let

$$c_i = \begin{cases} c_1, & s_i = 1 \\ c_{-1}, & s_i = -1 \end{cases}.$$

and define $C = \sum_i c_i$ then $\bar{R}_A(\omega)$ in terms of correctly classified samples is

$$\begin{aligned} \bar{R}_A(\omega) &= C - \sum_{i=1}^{n} \Big(c_{-1} I(\omega \cdot \xi_i \leq 0, s_i = -1) \ + \ c_1 I(\omega \cdot \xi_i > 0, s_i = 1) \Big) \\ &= C - \sum_{i=1}^{n} \Big(c_i I(\omega \cdot (s_i \xi_i) > 0) \ + \ c_{-1} I(\omega \cdot (s_i \xi_i) = 0, s_i = -1) \Big) \end{aligned}$$

and from the definition of ϕ

$$\bar{R}_A(\omega) = C - \sum_{i=1}^{n} \Big(c_i I(\omega \cdot z_i > 0) + c_{-1} I(\omega \cdot z_i = 0, s_i = -1) \Big). \tag{29}$$

To complete the proof we verify conditions 3.1-3.3 in Lemma 3. For any $\omega \in \Re^{d+1}$ let

$$\delta = \begin{cases} 1, & \omega \cdot z_i = 0 \text{ for all } z_i \in Z \\ \min_{z_i \in Z, \omega \cdot z_i \neq 0} |\omega \cdot z_i|, & \text{otherwise} \end{cases}$$

and let

$$\acute{\omega} = \omega - (\delta/2, 0)\ , \quad 0 \in \Re^d.$$

This gives

$$\begin{array}{ll} \acute{\omega} \cdot z_i \geq \delta/2 > 0, & \text{when } (\omega \cdot z_i > 0) \text{ or } (\omega \cdot z_i = 0, s_i = -1) \\ \acute{\omega} \cdot z_i \leq -\delta/2 < 0, & \text{when } (\omega \cdot z_i < 0) \text{ or } (\omega \cdot z_i = 0, s_i = 1) \end{array}$$

and therefore condition 3.1 holds and (29) can be written

$$\bar{R}_A(\omega) = \bar{R}_A(\acute{\omega}) = C - \sum_{i=1}^{n} c_i I(\acute{\omega} \cdot z_i > 0) = C - \sum_{i \in J^+(\acute{\omega})} c_i.$$

which verifies condition 3.2. The right hand side of this expression also establishes a monotonic relation between nested sets J^+ and the values of $\bar{R}$ which verifies condition 3.3 and completes our proof. ♦

References

Bartlett, P., & Lugosi, G. (1999). An inequality for uniform deviations of sample averages from their means. *Statistics and Probability Letters*, *44*, 55–62.

Block, H., & Levin, S. (1970). On the boundedness of an iterative procedure for solving a system of linear inequalities. *Proceedings of the American Mathematical Society*, *26:2*, 229–235.

Cannon, M., Hochberg, J., & Kelly, P. (1999). Quality Assessment and Restoration of Typewritten Document Images. *International Journal on Document Analysis and Recognition*, *2*(2/3), 80–89.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, New York, NY.

Duda, R., Hart, P., & Stork, D. (2000). *Pattern Classification.* Wiley, New York, NY.

Gallant, S. (1990). Perceptron–based learning algorithms. *IEEE Transactions on Neural Networks*, *1*(2), 179–191.

Höffgen, K.-U., & Simon, H.-U. (1992). Robust trainability of single neurons. In *Proceedings of the Computational Learning Theory (COLT) Conference*, pp. 428–438.

Muselli, M. (1997). On convergence properties of pocket algorithm. *IEEE Transactions on Neural Networks*, *8*(3), 623–629.

Natarajan, B. (1991). *Machine Learning: A theoretical Approach.* Morgan Kaufmann, San Mateo, CA.

Nilsson, N. (1990). *The Mathematical Foundation of Learning Machines.* Morgan–Kaufmann, San Mateo, CA.

Novikoff, A. (1962). On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, Vol. XII, pp. 615–622 Polytechnic Institute of Brooklyn.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley and Sons, Inc., New York.

Smith, F. (1969). Design of multicategory pattern classifiers with two–category classifier design procedures. *IEEE Transactions on Computers*, *C–18*, 548–551.

Vapnik, V. (1998). *Statistical Learning Theory.* John Wiley and Sons, Inc., New York, NY.