The Fundamental Theory of Optimal "Anti-Bayesian" *Parametric* Pattern Classification Using Order Statistics Criteria*

A. Thomas† and B. John Oommen‡

Abstract

The gold standard for a classifier is the condition of optimality attained by the Bayesian classifier. Within a Bayesian paradigm, if we are allowed to compare the testing sample with only *a single* point in the feature space from each class, the *optimal* Bayesian strategy would be to achieve this based on the (Mahalanobis) distance from the corresponding means. The reader should observe that, in this context, the mean, in one sense, is the most *central* point in the respective distribution. In this paper, we shall show that we can obtain optimal results by operating in a diametrically opposite way, i.e., a so-called "anti-Bayesian" manner. Indeed, we assert a completely counter-intuitive result that by working with a *very few* points *distant* from the mean, one can obtain remarkable classification accuracies. The number of points can sometimes be as small as *two*. Further, if these points are determined by the *Order Statistics* of the distributions, the accuracy of our method, referred to as Classification by Moments of Order Statistics (CMOS), attains the optimal Bayes' bound. This claim, which is totally counter-intuitive, has been proven for many uni-dimensional, and some multi-dimensional distributions within the exponential family, and the theoretical results have been verified by rigorous experimental testing. Apart from the fact that these results are quite fascinating and pioneering in their own right, they also give a theoretical foundation for the families of Border Identification (BI) algorithms reported in the literature.

Keywords : *Pattern Classification, Order Statistics, Reduction of training patterns, Prototype Reduction Schemes, Classification by Moments of Order Statistics*

*The authors are grateful for the partial support provided by NSERC, the Natural Sciences and Engineering Research Council of Canada. We are also very grateful for the comments made by the Associate Editor and the anonymous Referees. Their input helped in improving the quality of the final version of this paper. A preliminary version of this paper will be presented as a Keynote/Plenary talk at CIARP'12, the 2012 Iberoamerican Congress on Pattern Recognition, Buenos Aires, Argentina, in September 2012.

†This author can be contacted at: School of Computer Science, Carleton University, Ottawa, Canada. E-mail address: athomas1@scs.carleton.ca.

‡*Chancellor's Professor*; *Fellow: IEEE* and *Fellow: IAPR*. This author can be contacted at: School of Computer Science, Carleton University, Ottawa, Canada : K1S 5B6. This author is also an *Adjunct Professor* with the University of Agder in Grimstad, Norway. E-mail address: oommen@scs.carleton.ca.

1 Introduction

Pattern classification is the process by which unknown feature vectors are categorized into groups or classes based on their features. The age-old strategy for doing this is based on a Bayesian principle which aims to maximize the *a posteriori* probability. It is well known that when expressions for the latter are simplified, the classification criterion which attains the Bayesian optimal lower bound often reduces to testing the sample point using the corresponding distances/norms to the *means* or the "central points" of the distributions.

In this paper, we shall demonstrate that we can obtain optimal results by operating in a diametrically opposite way, i.e., a so-called "anti-Bayesian" manner. Indeed, we shall show the completely counter-intuitive result that by working with a *few* points *distant* from the mean, one can obtain remarkable classification accuracies. The number of points referred to can be as small as *two* in the uni-dimensional case. Further, if these points are determined by the *Order Statistics* of the distributions, the accuracy attains the optimal Bayes' bound. Thus, put in a nut-shell, we introduce here the theory of optimal pattern classification using order statistics of the features rather than the distributions of the features themselves. Thus, we propose a novel methodology referred to as Classification by Moments of Order Statistics (CMOS). It turns out, though, that this process is computationally not any more complex than working with the latter distributions.

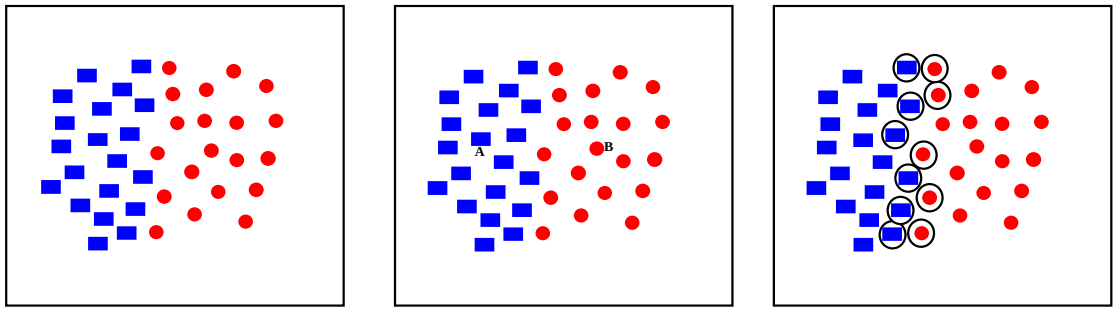
If we fast-forward the clock by five decades since the initial formulation of Pattern Recognition (PR) as a research field, the informed reader will also be aware of the development of efficient classification methods in which the schemes achieve their task based on a *subset* of the training patterns. These are commonly referred to as "Prototype Reduction Schemes" (PRS)[7, 21]. For the sake of our work, a PRS will be considered to be a generic method for reducing the number of training vectors, while simultaneously attempting to guarantee that the classifier built on the reduced design set performs as well, or nearly as well, as the classifier built on the original design set [12]. Thus, instead of considering all the training patterns for the classification, a subset of the whole set is selected based on certain criteria. The learning (or training) is then performed on this reduced training set, which is also called the "Reference" set. This Reference set not only contains the patterns which are closer to the true discriminant's boundary, but also the patterns from the other regions of the space that can adequately represent the entire training set.

Border Identification (BI) algorithms, which are a subset of PRSs, work with a Reference set which only contains "border" points. To enable the reader to perceive the difference between a traditional PRS and a BI algorithm, we present some typical data points in Figure 1. Consider Figure 1a in which the circles belong to ω_1 and rectangles belong to ω_2. A PRS would attempt to determine the relevant samples in both the classes which are capable of achieving near-optimal classification. Observe that some samples which fall strictly *within* the collection of points in each

class, such as A and B in Figure 1b, could be prototypes, because testing samples that fall close to them will be correctly classified. As opposed to this, in a BI algorithm, one uses *only* those samples which lie close to the *boundaries* of the two classes, as shown in Figure 1c.

Recent research [14] has shown that for overseeing the task of achieving the classification, the samples extracted by a BI scheme, and which lie *close to the boundaries* of the discriminant function, have significant information when it concerns the classification ability of the classifier.

Although this is quite amazing, *the formal analytical reason for this is yet unproven.*



(a) Training Set (b) Prototypes (c) Border Patterns

Figure 1: Border patterns vs Prototypes

In this paper, we present some pioneering results which bridge this conceptual gap. First of all, we shall formally and analytically show that operating in a totally anti-Bayesian perspective can, in spite of the diametrically opposite philosophy, still lead to an *optimal* Bayesian classification. More precisely, we shall show that by computing the appropriate distances/norms to certain (in many cases, as few as *two*) points that are distant from the means, we can obtain classification *identical* to the Bayesian scheme – as long as the corresponding comparisons are achieved using the appropriate data points that characterize the distributions. Indeed, the points that we refer to here will be shown to be the expected values of the moments of Order Statistics (OS) of the respective distributions. These representative OS points will model the above-mentioned representative prototypes derived by means of a BI algorithm, thus closing the conceptual gap.

The question of how we can compute the BI points which match these criteria from the training data is presently being investigated.

1.1 Problem Formulation

The objective of a PRS is to reduce the cardinality of the training set to be as small as possible by selecting some training patterns based on various criteria, as long as the reduction does not affect the performance. Specializing this criterion, the current-day BI algorithms, designed by Duch, Foody, and Li *et al.*, and which are briefly explained in Section 2.2, attempt to select a Reference set which contains border patterns derived, in turn, from the set of training patterns. Observe that,

in effect, these algorithms also yield reduced training sets. Once the Reference set is obtained, all of these traditional methods perform the classification by invoking some sort of classifier, like the SVM, a neural network etc. Recent research in the field of PR have claimed that the points which are closer to the class borders are more informative, and that they can play a significant role in the classification. Contrary to a Bayesian intuition, these border patterns have the ability to accurately classify the testing patterns, as we shall presently demonstrate. The prime objective of this paper is to explain the theoretical reason for why those points are more informative and important.

Our main hypothesis is that the classification could just as well be attempted in the OS space as in the original space itself. Our work will show that the OS points themselves are not necessarily central to the distribution.

1.2 Contributions of this Paper

The novel contributions of this paper are the following:

- We propose an "anti-Bayesian" paradigm for the classification of patterns within the parametric mode of computation, where the distance computations are not with regard to the "mean" but with regard to some samples "distant" from the mean. These points, which is sometimes as few as *two*, are the moments of OS of the distributions;
- We provide a theoretical framework for adequately responding to the question of why the border points are more informative for the task of classification;
- To justify these claims, we submit a formal analysis and the results of various experiments which have been performed for many distributions within the exponential family, and the results are clearly conclusive.

We conclude by mentioning that our results probably represent the state-of-the-art in BI.

1.3 Paper Organization

The rest of the paper is organized as follows. First of all, in Section 2 we present a brief overview of the PRSs and the present-day BI algorithms, and also include a brief introduction to the concept of Order Statistics (OS) of a distribution, and of the *moments* of the OS. Section 3 includes the main results of the paper and gives a clear understanding of the optimal classification that can be achieved by invoking the properties of the moments of the OS for various distributions. In each case, we include the analytic proofs followed by the experimental results obtained by rigorous testing. In many cases, the results have been derived and simulated for uni-dimensional and multi-dimensional distributions. Section 4, which concludes the paper, also includes a suite of problems which can be studied from the perspective of OS criteria.

2 Relevant Background Areas

2.1 Prototype Reduction Schemes

Zillions of PRS [13] techniques have developed over the years, and it is clearly impossible to survey all of these here. These include the Condensed Nearest Neighbor (CNN) rule [10], the Reduced Nearest Neighbor (RNN) rule [8], the Prototypes for Nearest Neighbor (PNN) classifiers [2], the Selective Nearest Neighbor (SNN) rule [18], the Edited Nearest Neighbor (ENN) rule [3], Vector Quantization (VQ) etc., some of which are briefly explained below[1].

While some of the above techniques merely *select* a subset of the existing patterns as prototypes, other techniques *create* new prototypes so as to represent *all* the existing patterns in the best manner. Of the above-listed PRS techniques, the CNN, RNN, SNN and ENN merely *select* prototypes from the existing patterns, while the PNN and VQ create new prototypes that collectively represent the entire training set. The review below is necessarily brief since we have merely listed two schemes which *select* prototypes and one which *creates* them. Comprehensive survey of the state-of-the-art in PRSs can be found in [7, 11, 21]. The formal algorithms[2] are also found in [19].

2.1.1 Condensed Nearest Neighbor (CNN)

The CNN has been suggested as a rule that retains the basic approach of a Nearest Neighbor paradigm to determine a consistent subset of the original sample set. However, this technique, in general, will not lead to a minimal consistent sample set, which is a set that contains a minimum number of samples able to correctly classify all the remaining samples in the given set.

Initially, the first pattern of the original training set T is copied to T_{CNN}. Then, the second pattern of T is classified by considering T_{CNN} as the Reference set. If that pattern is correctly classified, it is moved to R, which is the set of patterns to be removed. Otherwise, it is moved to the Reference set. This procedure is repeated for all the patterns of T. Once all the patterns have been considered for such a verification phase, the same procedure is repeated for the set R, which contains the patterns to be removed. This phase will be repeated until either the set R becomes empty (i.e., the Reference set is equivalent to the original set), or no more patterns are left in R which have any effect on the classification.

Once this pre-processing has been achieved, T_{CNN} will be the Reference set for the NN rule. The patterns that are moved to R will be discarded.

[1] This section can be removed or abridged if requested by the referees.

[2] A copy of the PhD proposal can be found at `http://people.scs.carleton.ca/~athomas1/Proposal.pdf`.

2.1.2 Reduced Nearest Neighbor (RNN)

Gates proposed the RNN as an extension of the CNN, that attempts to further reduce the original training set, from what was suggested by the CNN. The RNN algortihm first invokes the CNN algorithm to obtain T_{CNN}, the reduced training set derived by the CNN rule. It then tries to discard those patterns which do not have any influence in the classification process. To accomplish this, the RNN algorithm removes one pattern per iteration, by classifying T using the set T_{RNN} as the Reference set. If at least one pattern is not correctly classified, it is obvious that the removed pattern has some influence on the classification. Consequently, the removed pattern is again included into T_{RNN}, and the procedure is continued with the next pattern of T_{RNN}.

2.1.3 Prototypes for Nearest Neighbor (PNN)

Another PRS scheme, the PNN algorithm [2], can be described as follows: Given a training set T, the algorithm starts with every point in T as a prototype. We now define two auxiliary sets A and B. Initially, set A is empty and set B is equal to T, where every prototype (data sample) has an associated weight of unity. The algorithm selects an arbitrary point in B and initially assigns it to A. After this, the two closest prototypes $\mathbf{p}$ in A and $\mathbf{q}$ in B of the same class are merged, successively, into a new prototype, $\mathbf{p}^*$. This is done only if the merging will not degrade the classification of the patterns in T, where $\mathbf{p}^*$ is the weighted average of $\mathbf{p}$ and $\mathbf{q}$. After determining the new value of $\mathbf{p}^*$, $\mathbf{p}$ from A and $\mathbf{q}$ from B are deleted, and $\mathbf{p}^*$ is included into A. Thereafter, the procedure is repeated until a static condition attains.

If either $\mathbf{p}$ and $\mathbf{q}$ are not of the same class, or if merging is unsuccessful, $\mathbf{q}$ is moved from B to A, and the procedure is repeated. When B becomes empty, the entire procedure is repeated by setting B to be the final A obtained from the previous cycle, and by resetting A to be the empty set, until no new merged prototypes are obtained. The final prototypes in A are then used as the Reference set in a NN classifier. The bottom-up nature of this method is crucial to its convergence.

2.2 Border Identification Algorithms

Border Identification (BI) algorithms form a distinct subset of PRSs. Since our aim is to also formalize the rationale for BI methods, they are briefly surveyed here.

2.2.1 Duch's Algorithms

Duch, in a pioneering endeavor, developed algorithms to obtain the Reference set based on a border analysis of every training pattern. He designed two techniques which serve to select the most effective reference vectors *near* the class borders. The first method, referred to as `Duch1`, starts with an empty Reference set. For every training pattern $\mathbf{x}$, the scheme identifies k nearest patterns,

and those patterns which are from the class *other than* the class of $\mathbf{x}$ are added to the Reference set. In this way, the algorithm, in effect, attempts to add patterns which are closer to the class boundary, to the Reference set [5].

The whole procedure is repeated a number of times, from a maximum value of k, denoted by K_2, to a minimum value of k, denoted by K_1, with $K_1 < K_2$.

The alternate method proposed by Duch to select the Reference set starts with the entire training set, T. For every pattern $\mathbf{x}$ in T, k nearest patterns are identified. If all the k nearest patterns are from the same class as that of $\mathbf{x}$, the pattern $\mathbf{x}$ is removed from the Reference set, since all the removed patterns are, possibly, farther from the class borders. Thus, the Reference set that is retained contains only the patterns which are closer to the class borders.

In order to apply Duch's approaches, the value of k should be considered as a user-defined parameter.

2.2.2 Foody's Algorithm

Another important development in this area was proposed by Foody [6]. According to his approach, the training set is divided into two sets - the first comprising of the set of border patterns, and the second being the set of non-border patterns. A border training set should contain patterns from different classes, but which are close together in the feature space and which are thus in the proximity of the true classification boundary. A pattern which is almost as close to its actual class as it is to the other class can be considered as a border pattern. In order to decide whether a training pattern is a border pattern or not, a scale of "borderness" is defined. Foody expressed "borderness" as the difference between the two smallest Mahalanobi's distances measured for each training pattern. It is obvious that even though the algorithm is aimed for obtaining border patterns, it yields near and far borders.

2.2.3 Border Identification in Two Stages ($\mathbf{BI_2}$)

The "traditional" BI algorithms proposed by Duch and Foody are based on the concept of similarity distance metrics. However, Li claimed that those algorithms are unable to learn the so-called "Full" border points, which include the "Far" border points. This being the case, more recent research had proposed the concept of finding the border patterns in two stages (i.e., the $\mathrm{BI_2}$ algorithm) [14]. The border patterns obtained by the traditional approaches are considered to be the "Near" borders, and using the latter, the "Far" borders are identified from the remaining data points. It turns out that the final border points computed in this manner are more accurate than the initially identified "Near" borders. The "Near" and the "Far" borders collectively constitute the so-called Full border set for the training data.

The first stage of the BI_2 algorithm, proposed by Li *et al.*[14], used an approach similar to Duch's first algorithm to get the initial Reference set. After obtaining the Near borders, the algorithm then identifies the Far borders from the Remaining set. To achieve this, the scheme removes the redundant data patterns from the Remaining set, and this process is repeated until the Remaining set becomes empty. This scheme has been further augmented by the inclusion of a phase referred to as Progressive Sampling (PS). Li *et al.* have recommended that PS be applied to BI_2 so as to detect when the scheme has converged to, thereby, yield a more optimal border set. The new algorithm, the PBS, proposed by Li *et al.*, progressively learns sufficient borders that can classify the whole training set. In other words, it is used to yield a criterion by which one can get an iterative process for finding new border patterns.

2.3 Order Statistics

Let $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ be a univariate random sample of size n that follows a continuous distribution function Φ, where the probability density function (pdf) is $\varphi(\cdot)$. Let $\mathbf{x}_{1,n}$, $\mathbf{x}_{2,n}$, ..., $\mathbf{x}_{n,n}$ be the corresponding Order Statistics (OS). The r^{th} OS, $\mathbf{x}_{r,n}$, of the set is the r^{th} smallest value among the given random variables. The pdf of $\mathbf{y} = \mathbf{x}_{r,n}$ is given by:

$$f_{\mathbf{y}}(y) = \frac{n!}{(r-1)!(n-r)!} \{\Phi(y)\}^{r-1} \{1-\Phi(y)\}^{n-r} \varphi(y),$$

where $r = 1, 2, ..., n$. The reasoning for the above expression is straightforward. If the r^{th} OS appears at a location given by $\mathbf{y} = \mathbf{x}_{r,n}$, it implies that the $r-1$ smaller elements of the set are drawn independently from a Binomial distribution with a probability $\Phi(y)$, and the other $n-r$ samples are drawn using the probability $1-\Phi(y)$. The factorial terms result from the fact that the $(r-1)$ elements can be independently chosen from the set of n elements.

Although the distribution $f_{\mathbf{y}}(y)$ contains all the information resident in $\mathbf{y}$, the literature characterizes the OS in terms of quantities which are of paramount importance, namely its moments [20]. To better appreciate the results presented later in this paper, an understanding of the moments of the OS is necessary. This is briefly presented below.

Using the distribution $f_{\mathbf{y}}(y)$, one can see that the k^{th} moment of $\mathbf{x}_{r,n}$ can be formulated as:

$$E[\mathbf{x}_{r,n}^k] = \frac{n!}{(r-1)!(n-r)!} \int_{-\infty}^{+\infty} y^k \Phi(y)^{k-1} (1-\Phi(y))^{n-r} \varphi(y) dy,$$

provided that both sides of the equality exist [1, 16].

The fundamental theorem concerning the OS that we invoke is found in many papers [15, 16, 20]. The result is merely cited below inasmuch as the details of the proof are irrelevant and outside the scope of this study. The theorem can be summarized as follows.

Let $n \geq r \geq k+1 \geq 2$ be integers. Then, since Φ is a nondecreasing and right-continuous function from $\mathbb{R} \rightarrow \mathbb{R}$, $\Phi(\mathbf{x}_{r,n})$ is uniform in [0,1]. If we now take the k^{th} moment of $\Phi(\mathbf{x}_{r,n})$, it has the form [15]:

$$E[\Phi^k(\mathbf{x}_{r,n})] = \frac{B(r+k, n-r+1)}{B(r, n-r+1)} = \frac{n!\ (r+k-1)!}{(n+k)!\ (r-1)!}, \tag{1}$$

where $B(a,b)$ denotes the *Beta* function, and $B(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}$ since its parameters are integers.

The above fundamental result can also be used for characterization purposes as follows [15]. Let $n \geq r \geq k+1 \geq 2$ be integers, with Φ being nondecreasing and right-continuous. Let G be *any* nondecreasing and right-continuous function from $\mathbb{R} \rightarrow \mathbb{R}$ on the same support as Φ. The relation

$$E[G^k(\mathbf{x}_{r,n})] = \frac{n!\ (r+k-1)!}{(n+k)!\ (r-1)!} \tag{2}$$

holds if and only if $\forall x, \Phi(x) = G(x)$. In other words, $\Phi(\cdot)$ is the unique function that satisfies Eq. (2), implying that every distribution is characterized by the moments of its OS.

The implications of the above are the following:

1. If $n = 1$, implying that only a *single* sample is drawn from $\mathbf{x}$, from Eq. (1),

$$E[\Phi^1(\mathbf{x}_{1,1})] = \frac{1}{2}, \implies E[\mathbf{x}_{1,1}] = \Phi^{-1}\left(\frac{1}{2}\right). \tag{3}$$

   Informally speaking, the first moment of the 1-order OS would be the value where the cumulative distribution Φ equals $\frac{1}{2}$, which is the Median($\mathbf{x}$).

2. If $n = 2$, implying that only *two* samples are drawn from $\mathbf{x}$, we can deduce from Eq. (1) that:

$$E[\Phi^1(\mathbf{x}_{1,2})] = \frac{1}{3}, \implies E[\mathbf{x}_{1,2}] = \Phi^{-1}\left(\frac{1}{3}\right), \text{and} \tag{4}$$

$$E[\Phi^1(\mathbf{x}_{2,2})] = \frac{2}{3}, \implies E[\mathbf{x}_{2,2}] = \Phi^{-1}\left(\frac{2}{3}\right). \tag{5}$$

   Thus, from a computational perspective, the first moment of the first and second 2-order OS would be the values where the cumulative distribution Φ equal $\frac{1}{3}$ and $\frac{2}{3}$ respectively.

Although the analogous expressions can be derived for the higher order OS, for the rest of this paper we shall merely focus on the 2-order OS, and derive the consequences of using them in classification.

3 Optimal Bayesian Classification using *Two* Order Statistics

3.1 The Generic Classifier

Having characterized the moments of the OS of arbitrary distributions, we shall now consider how they can be used to design a classifier.

Let us assume that we are dealing with the 2-class problem with classes ω_1 and ω_2, where their class-conditional densities are $f_1(x)$ and $f_2(x)$ respectively (i.e, their corresponding distributions are $F_1(x)$ and $F_2(x)$ respectively)[3]. Let ν_1 and ν_2 be the corresponding *medians* of the distributions. Then, classification based on ν_1 and ν_1 would be the strategy that classifies samples based on a *single* OS. We shall show the fairly straightforward result that for all symmetric distributions, the classification accuracy of this classifier attains the Bayes' accuracy.

This result is not too astonishing because the median is centrally located close to (if not exactly) on the mean. The result for higher order OS is actually far more intriguing because the higher order OS are not located centrally (close to the means), but rather distant from the means. Consequently, we shall show that for a large number of distributions, mostly from the exponential family, the classification based on *these* OS again attains the Bayes' bound.

We shall initiate this discussion by examining the Uniform distribution. The reason for this is that even though the distribution itself is rather trivial, the analysis will provide the reader with an insight into the mechanism by which the problem can be tackled, which can then be extended for other distributions.

3.2 The Uniform Distribution

The continuous Uniform distribution is characterized by a constant function $U(a, b)$, where a and b are the minimum and the maximum values that the random variable $\mathbf{x}$ can take. If the class conditional densities of ω_1 and ω_2 are uniformly distributed,

$$f_1(x) = \begin{cases} \frac{1}{b_1 - a_1} & \text{if } a_1 \leq x \leq b_1; \\ 0 & \text{if } x < a_1 \text{ or } x > b_1, \text{ and} \end{cases}$$

$$f_2(x) = \begin{cases} \frac{1}{b_2 - a_2} & \text{if } a_2 \leq x \leq b_2; \\ 0 & \text{if } x < a_2 \text{ or } x > b_2. \end{cases}$$

The reader should observe the following:

- If $a_2 > b_1$, the two distributions are non-overlapping, rendering the classification problem trivial.

[3]Throughout this section, we will assume that the *a priori* probabilities are equal. If they are unequal, the above densities must be weighted with the respective *a priori* probabilities.

- If $a_2 < b_1$, but $b_1 - a_1 \neq b_2 - a_2$, the optimal Bayesian classification is again dependent only on the heights of the distributions. In other words, if $b_2 - a_2 < b_1 - a_1$, the testing sample will be assigned to ω_2 whenever $x > a_2$. This criterion again is not related to the mean of the distributions at all, and is thus un-interesting to our current investigations.

- The meaningful scenario is when $b_1 - a_1$ is exactly equal to $b_2 - a_2$, and if $a_2 < b_1$. In this case, the heights of the two distributions are equal and the distributions are overlapping. This is really the interesting case, and corresponds to the scenario when the two distributions are identical. We shall analyze this in greater detail and demonstrate that the optimal Bayesian classification is also attained by using the OS.

3.2.1 Theoretical Analysis: Uniform Distribution - 2-OS

We shall now derive the formal properties of the classifier that utilizes the OS for the Uniform distribution.

Theorem 1. *For the 2-class problem in which the two class conditional distributions are Uniform and identical, CMOS, the classification using two OS, attains the optimal Bayes' bound.*

Proof. The proof of the result is done in two steps. We shall first show that when the two class conditional distributions are Uniform and identical, the optimal Bayesian classification is achieved by a comparison to the corresponding *means*. The equivalence of this to a comparison to the corresponding OS leads to the final result.

Without loss of generality let the class conditional distributions for ω_1 and ω_2 be $U(0,1)$ and $U(h, 1+h)$, with means $\mu_1 = \frac{1}{2}$ and $\mu_2 = h + \frac{1}{2}$, respectively. In this case, the optimal Bayes' classifier assigns x to ω_1 whenever $x < h$, x to ω_2 whenever $x > 1$, and x to ω_1 and to ω_2 with equal probability when $h < x < 1$. Since:

$$\begin{aligned} D(x,\ \mu_1) < D(x,\ \mu_2) \quad &\Longleftrightarrow \quad x - \frac{1}{2} < h + \frac{1}{2} - x \\ &\Longleftrightarrow \quad 2x < 1 + h \\ &\Longleftrightarrow \quad x < \frac{1+h}{2}, \end{aligned} \tag{6}$$

we see that the optimal Bayesian classifier assigns the sample based on the proximity to the corresponding mean, proving the first assertion.

We now consider the moments of the OS of the distributions. If $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ are n independent univariate random variables that follow the Uniform distribution $U(0,1)$, by virtue of Eq.(1), the expected values of the first moment of the k-order OS can be seen to be $E[\mathbf{x}_{k,n}] = \frac{k}{n+1}$. Thus, for

$U(0,1)$, $E[\mathbf{x}_{1,2}] = \frac{1}{3}$ and $E[\mathbf{x}_{2,2}] = \frac{2}{3}$. Similarly, for the distribution $U(h, 1+h)$, the expected values are $E[\mathbf{x}_{1,2}] = h + \frac{1}{3}$ and $E[\mathbf{x}_{2,2}] = h + \frac{2}{3}$.

The OS-based classification is thus as follows: Whenever a testing sample comes from these distributions, the CMOS will compare the testing sample with $E[\mathbf{x}_{2,2}]$ of the first distribution, i.e., $\frac{2}{3}$, and with $E[\mathbf{x}_{1,2}]$ of the second distribution, i.e., $h+\frac{1}{3}$, and the sample will be labeled with respect to the class which minimizes the corresponding quantity, as shown in Figure 2. Observe that for the above rule to work, we must enforce the ordering of the OS of the two distributions, and this requires that $\frac{2}{3} < h + \frac{1}{3} \implies h > \frac{1}{3}$.

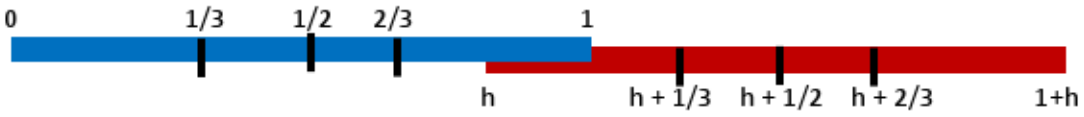


Figure 2: Uniform Distribution

In order to prove that for $h > \frac{1}{3}$ the OS-based classification is identical to the mean-based classification, we have to prove that $D(x, \mu_1) < D(x, \mu_2) \implies D(x, O_1) < D(x, O_2)$, where O_1 is $E[\mathbf{x}_{2,2}]$ of the first distribution and O_2 is $E[\mathbf{x}_{1,2}]$ of the second distribution. By virtue of Eq.(6),

$$D(x, \mu_1) < D(x, \mu_2) \iff x < \frac{h+1}{2}. \tag{7}$$

Similarly,

$$\begin{aligned} D(x, O_1) < D(x, O_2) &\iff D\left(x, \frac{2}{3}\right) < D\left(x, h + \frac{1}{3}\right) \\ &\iff x - \frac{2}{3} < h + \frac{1}{3} - x \\ &\iff x < \frac{h+1}{2}. \end{aligned} \tag{8}$$

The result follows by observing that (7) and (8) are identical comparisons.

For the analogous result for the case when $h < \frac{1}{3}$, the CMOS will compare the testing sample with $E[\mathbf{x}_{1,2}]$ of the first distribution, i.e., $\frac{1}{3}$, and with $E[\mathbf{x}_{2,2}]$ of the second distribution, i.e., $h + \frac{2}{3}$. Again, the sample will be labeled with respect to the class which minimizes the corresponding quantity. The proofs of the equivalence of this to the Bayesian decision follows along the same lines as the case when $h > \frac{1}{3}$, and is omitted to avoid repetition.

Hence the theorem! □

By way of example, consider the distributions $U(0,1)$ and $U(0.8, 1.8)$. Our claim is demonstrated

in Figure 3. In the figure, d_{m1} and d_{m2} are the distances of the testing sample with respect to the means of the first and the second class (i.e., 0.5 and 1.3) respectively, and d_{os1} and d_{os2} are the distances of the testing sample with respect to the moments of the OS for both the classes. The testing sample will be assigned to class ω_1 if $d_{os1} < d_{os2}$ otherwise to class ω_2. The interesting point is that the latter classification is, indeed, the Bayesian conclusion too.

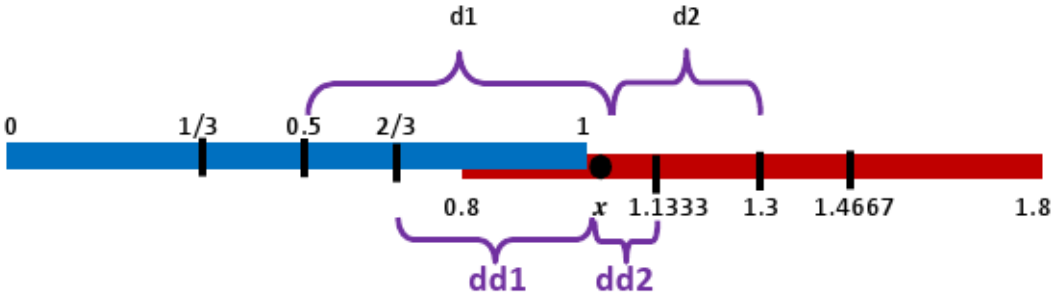


Figure 3: Analysis of Uniform Distribution

3.2.2 Experimental Results: Uniform Distribution - 2-OS

The CMOS method explained in Section 3.2.1 has been rigorously tested for various uniform distributions with 2-OS. In the interest of brevity, a few typical results are given below.

For each of the experiments, we generated 1,000 points for the classes ω_1 and ω_2 characterized by $U(0,1)$ and $U(h,1+h)$ respectively. We then invoked a classification procedure by utilizing the Bayesian and the CMOS strategies. In every case, CMOS was compared with the Bayesian classifier for different values of h, as tabulated in Table 1. The results in Table 1 were obtained by executing each algorithm 50 times using a 10-fold cross-validation scheme.

h	0.95	0.90	0.85	0.80	0.75	0.70
Bayesian	97.58	95.1	92.42	90.23	87.82	85.4
CMOS	97.58	95.1	92.42	90.23	87.82	85.4

Table 1: Classification of Uniformly distributed classes by the CMOS 2-OS method for different values of h.

Observe that in every case, the accuracy of CMOS attained the Bayes' bound.

By way of example, we see that CMOS should obtain the Bayesian bound for the distributions $U(0,1)$ and $U(0.8,1.8)$ whenever $n < \frac{1+0.8}{1-0.8} = 9$. In this case, the expected values of the moments are $\frac{1}{10}$ and $\frac{9}{10}$ respectively. These results justify the claim of Theorem 1.

3.2.3 Theoretical Analysis: Uniform Distribution - k-OS

We have seen from Theorem 1 that the moments of the 2-OS are sufficient for the classification to attain a Bayes' bound. We shall now consider the scenario when we utilize other k-OS. The formal result pertaining to this is given in Theorem 2.

Theorem 2. *For the 2-class problem in which the two class conditional distributions are Uniform and identical as U(0,1) and U(h, 1+h), optimal Bayesian classification can be achieved by using symmetric pairs of the n-OS, i.e., the $n-k$ OS for ω_1 and the k OS for ω_2 if and only if $k > \frac{(n+1)(1-h)}{2}$.*

Proof. We know that for the uniform distribution $U(0,1)$, the expected values of the first moment of the k-order OS have the form $E[\mathbf{x}_{k,n}] = \frac{k}{n+1}$. Our claim is based on the classification in which we can choose any of the symmetric pairs of the n-OS, i.e., the $n-k$ OS for ω_1 and the k OS for ω_2, whose expected values are $\frac{n-k+1}{n+1}$ and $h + \frac{k}{n+1}$ respectively.

Consider the case when $h > 1 - \frac{2k}{n+1}$, the relevance of which will be argued presently. Whenever a testing sample comes, it will be compared with the corresponding k-OS symmetric pairs of the expected values of the n-OS, and the sample will be labeled with respect to the class that minimizes the distance. Observe that for the above rule to work, we must again enforce the ordering of the OS of the two distributions, and this requires that:

$$\frac{n-k+1}{n+1} < h + \frac{k}{n+1} \implies k > \frac{(n+1)(1-h)}{2}. \tag{9}$$

Eq.(9) can be seen to be:

$$k > \frac{(n+1)(1-h)}{2} \implies h > 1 - \frac{2k}{n+1}, \tag{10}$$

which justifies the case under consideration. As we have already proved that the Bayesian bound can be achieved by a comparison to the corresponding means (in Eq.(6)), which in turn simplifies to $x \sim \omega_1 \iff x < \frac{h+1}{2}$, we need to show that to obtain optimal accuracy using these symmetric $n-k$ and k OS, $D(x, O_1) < D(x, O_2) \iff x < \frac{h+1}{2}$. Indeed, the OS-based classification also attains the Bayesian bound because:

$$\begin{aligned} D(x,O_1) < D(x,O_2) &\iff D\left(x,\ \frac{n-k+1}{n+1}\right) < D\left(x,\ h + \frac{k}{n+1}\right) \\ &\iff x - \frac{n-k+1}{n+1} < h + \frac{k}{n+1} - x \\ &\iff x < \frac{h+1}{2}. \end{aligned} \tag{11}$$

For the symmetric argument when $h < 1 - \frac{2k}{n+1}$, the CMOS will compare the testing sample

with $E[\mathbf{x}_{k,n}]$ of the first distribution and $E[\mathbf{x}_{n-k,n}]$ of the second distribution and the classification is obtained based on the class that minimizes *this* distance. The details of the proof are analogous and omitted. Hence the theorem! □

Remark: We can visualize this result from another perspective when we observe that we are concerned about the *ensemble* of symmetric pairs that can be considered to be *effective* for the classification. In order to obtain the maximum accuracy, the expected value of the first moment of the OS for the first class should be less that $\frac{1+h}{2}$, which implies that $\frac{n-k+1}{n+1} < \frac{1+h}{2}$, because if this condition is violated, the testing samples will be misclassified. Thus:

$$\begin{aligned} \frac{n-k+1}{n+1} < \frac{1+h}{2} \quad &\iff \quad 2n-2k+2 < (n+1)(1+h) \\ &\iff \quad k > \frac{(n+1)(1-h)}{2}, \end{aligned} \tag{12}$$

which is again the same condition found in the statement of Theorem 2 and Eq. (9). This, indeed, implies that the optimal Bayesian bound can be obtained with respect to different symmetric pairs of the n-OS, $\frac{n-k+1}{n+1}$ and $h + \frac{k}{n+1}$, if and only if $k > \frac{(n+1)(1-h)}{2}$.

3.2.4 Experimental Results: Uniform Distribution - k-OS

The CMOS method has also been tested for the Uniform distribution for other k OS. In the interest of brevity, we merely cite one example where the distributions for ω_1 and ω_2 were characterized by $U(0,1)$ and $U(0.8,1.8)$ respectively. For each of the experiments, we generated 1,000 points for each class, and the testing samples were classified based on the selected *symmetric* pairs for values k and $n-k$ respectively. The results are displayed in Table 2.

To clarify the table, consider the row given by Trial No. 6 in which the 7-OS were invoked for the classification. Observe that the k-OS are now given by $\frac{n-k+1}{n+1}$ and $\frac{k}{n+1}$ respectively. In this case, the possible symmetric OS pairs could be $\langle 1,6\rangle$, $\langle 2,5\rangle$, and $\langle 3,4\rangle$ respectively. In every single case, the accuracy attained the Bayes' bound, as indicated by the results in the table.

The consequence of violating the condition imposed by Theorem 2 can be seen from the results given in the row denoted by Trial No. 9. In this case, the testing attained the Bayes' accuracy for the symmetric OS pairs $\langle 2,9\rangle$, $\langle 3,8\rangle$, $\langle 4,7\rangle$ and $\langle 5,6\rangle$ respectively. However, the classifier "failed" for the specific 10-OS, when the OS used were $\frac{10}{11}$ and $h + \frac{1}{11}$, as these values did not satisfy the condition $h > 1 - \frac{2k}{n+1}$. Observe that if $h < 1 - \frac{2k}{n+1}$, the symmetric pairs should be reversed, i.e., $\frac{k}{n+1}$ for the first distribution, and $h + \frac{n-k+1}{n+1}$ for the second distribution, to obtain the optimal Bayesian bound. The astonishing facet of this result is that one obtains the Bayes' accuracy even though the classification requires only *two* points distant from the mean, justifying the rationale for BI schemes, and yet operating in an anti-Bayesian manner.

Trial No.	**Order(n)**	**Moments**	OS_1	OS_2	**CMOS**	**Pass/Fail**
1	Two	$\{\frac{i}{3} \mid 1 \leq i \leq 2\}$	$\frac{2}{3}$	$h + \frac{1}{3}$	90.23	Passed
2	Three	$\{\frac{i}{4} \mid 1 \leq i \leq 3\}$	$\frac{3}{4}$	$h + \frac{1}{4}$	90.23	Passed
3	Four	$\{\frac{i}{5} \mid 1 \leq i \leq 4\}$	$\frac{4}{5}$	h + $\frac{1}{5}$	90.23	Passed
4	Five	$\{\frac{i}{6} \mid 1 \leq i \leq 5\}$	$\frac{4}{6}$	$h + \frac{2}{6}$	90.23	Passed
5	Six	$\{\frac{i}{7} \mid 1 \leq i \leq 6\}$	$\frac{4}{7}$	$h + \frac{2}{7}$	90.23	Passed
6	Seven	$\{\frac{i}{8} \mid 1 \leq i \leq 7\}$	$\frac{5}{8}$	$h + \frac{3}{8}$	90.23	Passed
7	Eight	$\{\frac{i}{9} \mid 1 \leq i \leq 8\}$	$\frac{6}{9}$	$h + \frac{3}{9}$	90.23	Passed
8	Nine	$\{\frac{i}{10} \mid 1 \leq i \leq 9\}$	$\frac{7}{10}$	$h + \frac{3}{10}$	90.23	Passed
9	Ten	$\{\frac{i}{11} \mid 1 \leq i \leq 10\}$	$\frac{10}{11}$	$h + \frac{1}{11}$	9.77	Failed
10	Ten	$\{\frac{i}{11} \mid 1 \leq i \leq 10\}$	$\frac{9}{11}$	$h + \frac{2}{11}$	90.23	Passed
11	Ten	$\{\frac{i}{11} \mid 1 \leq i \leq 10\}$	$\frac{7}{11}$	$h + \frac{4}{11}$	90.23	Passed
12	Ten	$\{\frac{i}{11} \mid 1 \leq i \leq 10\}$	$\frac{6}{11}$	$h + \frac{5}{11}$	90.23	Passed

Table 2: Results of the classification obtained by using the symmetric pairs of the OS for different values of n. The value of h was set to be 0.8. Note that in every case, the accuracy attained the Bayes' value whenever the conditions stated in Theorem 2 were satisfied.

Remark: We believe that the CMOS, the classification by the moments of Order Statistics, is also true for multi-dimensional distributions. For a *prima facie* case, we consider two (overlapping) 2-dimensional uniform distributions U_1 and U_2 in which both the features are in $[0, 1]^2$ and $[h, 1+h]^2$ respectively. Consequently, we see that the overlapping region of the distributions forms a square (see Fig. 4). In this case, it is easy to verify that the Bayesian classifier is the diagonal that passes through the intersection points of the distributions. For the classification based on the moments of the 2-OS, because the features are independent for both dimensions, we can show that this is equivalent to utilizing the OS at position $\frac{2}{3}$ of the first distribution for both dimensions, and the OS at the position $h + \frac{1}{3}$ of the second distribution for both dimensions, as shown in Figure 4. In the figure, the points $\mathbf{a} = [a_1, a_2]$ and $\mathbf{c} = [c_1, c_2]$ denote the corresponding locations of the first feature at the positions $\frac{2}{3}$ and $h + \frac{1}{3}$ of the distributions respectively. Similarly, $\mathbf{b} = [b_1, b_2]$ and $\mathbf{d} = [d_1, d_2]$ denote the corresponding locations for the second feature of the given distributions. Observe that the Bayesian classifier passes exactly through the middle of these points.

One can easily observe that a testing point is classified to ω_1 if the value of its first feature is less than a_1 and is classified as ω_2 if the feature value is greater than c_1. The classification of the points that have the feature value as $a_1 < x_1 < c_1$, are equally likely and should thus be considered more carefully. By virtue of the uni-dimensional result derived in Theorem 1, we see that for *all* values

of the second dimension, the classifier for the first dimension in this region lies at the midpoint of **a** and **c**, which is exactly the position determined by the Bayes' classifier. Arguing in the same manner for the second dimension, we see that the values defined by the corresponding OS criteria will again be projected exactly onto the Bayes' classifier. Hence we conclude that the CMOS attains the Bayes' bound.

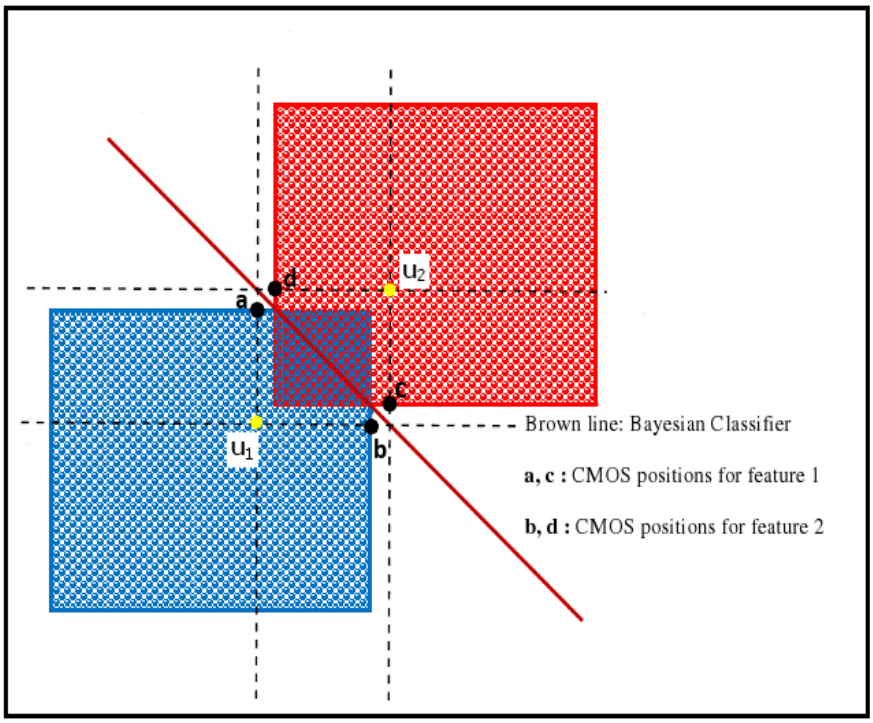


Figure 4: The optimal Bayes' classifier and the 2-OS CMOS for uniformly distributed 2-dimensional features.

The CMOS method for 2-dimensional uniform distributions U_1 (in $[0,1]$ in both dimensions) and U_2 (in $[h, 1+h]$ in both dimensions) has been rigorously tested, and the results are given in Table 3. A formal proof for the case when the second class is distributed in $[h_1, 1+h_1] \times [h_2, 1+h_2]$, and for multi-dimensional features is currently being devised. It will appear in a forthcoming paper.

h	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.60
Bayesian	99.845	99.505	98.875	98.045	97.15	95.555	94.14	91.82
CMOS	99.845	99.505	98.875	98.045	97.15	95.555	94.14	91.82

Table 3: Classification of Uniformly distributed 2-dimensional classes by the CMOS 2-OS method for different values of h. In the last two cases, the OS points of interest are reversed as explained in Section 3.2.4.

We now proceed to consider the CMOS for other distributions in the exponential family.

3.3 The Laplace (or Doubly-Exponential) Distribution

The *Laplace distribution* is a continuous uni-dimensional pdf named after Pierre-Simon Laplace. It is sometimes called the *doubly exponential distribution*, because it can be perceived as being a combination of two exponential distributions, with an additional location parameter, spliced together back-to-back.

If the class conditional densities of ω_1 and ω_2 are doubly exponentially distributed,

$$f_1(x) = \frac{\lambda_1}{2} e^{-\lambda_1 |x - c_1|}, \quad -\infty < x < \infty, \text{and}$$

$$f_2(x) = \frac{\lambda_2}{2} e^{-\lambda_2 |x - c_2|}, \quad -\infty < x < \infty,$$

where c_1 and c_2 are the respective means of the distributions. By elementary integration and straightforward algebraic simplifications, the variances of the distributions can be seen to be $\frac{2}{\lambda_1^2}$ and $\frac{2}{\lambda_2^2}$ respectively.

By way of example, the pdfs of doubly exponential distributions for different values of the parameter λ are given in Figure 5 where the optimal Bayes' classifier will evidently be at the point $\mathbf{x}^*$. Thus, if $\lambda_1 \neq \lambda_2$, the samples can be classified based on the heights of the distributions and their point of intersection. The formal results for the general case are a little more complex. However, to prove the analogous results of Theorem 1 for the Uniform distribution, we shall first consider the case when $\lambda_1 = \lambda_2$. In this scenario, the reader should observe the following:

- Because the distributions have the equal height, i.e. $\lambda_1 = \lambda_2$, the testing sample $\mathbf{x}$ will obviously be assigned to ω_1 if it is less than c_1 and be assigned to ω_2 if it is greater than c_2.
- Further, the crucial case is when $c_1 < \mathbf{x} < c_2$. In this regard, we shall analyze the CMOS classifier and prove that it attains the Bayes' bound even when one uses as few as *only* 2 OSs.

3.3.1 Theoretical Analysis: Doubly-Exponential Distribution - 2-OS

We shall first derive the moments of the 2-OS for the doubly exponential distribution. By virtue of Eq. (4) and (5), the expected values of the first moments of the two OS can be obtained by determining the points where the cumulative distribution function attains the values $\frac{1}{3}$ and $\frac{2}{3}$. Let u_1 be the point for the percentile $\frac{2}{3}$ of the first distribution, and u_2 be the point for the percentile

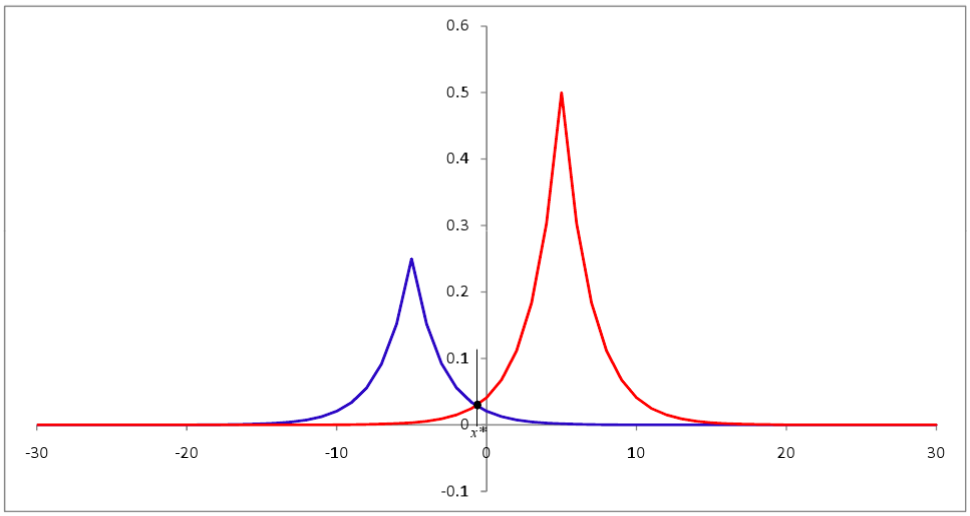


Figure 5: Doubly Exponential Distributions with different values for λ.

$\frac{1}{3}$ of the second distribution. Then:

$$\int_{c_1}^{u_1} \frac{\lambda_1}{2}\, e^{-\lambda_1|x-c_1|}dx \;=\; \frac{2}{3}-\frac{1}{2}=\frac{1}{6}, \text{ and} \tag{13}$$

$$\int_{-\infty}^{u_2} \frac{\lambda_2}{2}\, e^{\lambda_2|x-c_2|}dx \;=\; \frac{1}{3}. \tag{14}$$

The points of interest, i.e., u_1 and u_2, can be obtained by straightforward integrations and simplifications as follows:

$$\begin{aligned}\int_{c_1}^{u_1} \frac{\lambda_1}{2}\, e^{-\lambda_1|x-c_1|}dx \;=\frac{1}{6} \quad &\Longrightarrow \quad \left[-\frac{1}{2}e^{-\lambda_1(x-c_1)}\right]_{c_1}^{u_1}=\frac{1}{6} \\ &\Longrightarrow \quad u_1 = c_1 - \frac{1}{\lambda_1}\, log\left(\frac{2}{3}\right).\end{aligned} \tag{15}$$

Using a similar argument, we can see that:

$$u_2 = c_2 + \frac{1}{\lambda_2}\, log\left(\frac{2}{3}\right). \tag{16}$$

With these points at hand, we shall now demonstrate that, for doubly exponential distributions, the classification based on the expected values of the moments of the 2-OS, CMOS, attains the Bayesian bound.

Theorem 3. *For the 2-class problem in which the two class conditional distributions are Doubly Exponential and identical, CMOS, the classification using two OS, attains the optimal Bayes' bound.*

Proof. The proof can be done in two steps. As in the uniform case, first of all, we shall show that when the class conditional distributions are doubly exponential and identical, the optimal Bayes' bound can be attained by a comparison to the corresponding means, and as the concluding step, this can be shown to be equal to the accuracy of the CMOS, which lead to the proof of the theorem.

Without loss of generality, let the distributions of ω_1 and ω_2 be $D(c_1, \lambda)$ and $D(c_2, \lambda)$, where c_1 and c_2 are the means, and λ is the identical scale parameter. Then, to get the Bayes' classifier, we argue that:

$$\begin{aligned} p(x|\omega_1)p(\omega_1) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} p(x|\omega_2)p(\omega_2) &\implies \frac{\lambda}{2}\, e^{-\lambda|x-c_1|} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{\lambda}{2}\, e^{-\lambda|x-c_2|} \\ &\implies \lambda(x-c_1) \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} \lambda(c_2-x) \\ &\implies x \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} \frac{c_1+c_2}{2}. \end{aligned} \tag{17}$$

We now consider the classification with respect to the expected values of the moments of the 2-OS, u_1 and u_2, where as per Eq. (15) and (16), $u_1 = c_1 - \frac{1}{\lambda}\, log\left(\frac{2}{3}\right)$ and $u_2 = c_2 + \frac{1}{\lambda}\, log\left(\frac{2}{3}\right)$. In order to prove our claim, we need to show that

$$D(x, c_1) < D(x, c_2) \iff D(x, u_1) < D(x, u_2). \tag{18}$$

We first consider the LHS of Eq. (18). Indeed,

$$\begin{aligned} D(x, c_1) < D(x, c_2) &\implies x - c_1 < c_2 - x \\ &\implies 2x < c_1 + c_2 \\ &\implies x < \frac{c_1+c_2}{2}. \end{aligned} \tag{19}$$

What remains to be proven is that the RHS of Eq. (18) also simplifies to the same expression. This is true because:

$$\begin{aligned} D(x, u_1) < D(x, u_2) &\implies D\left(x,\ c_1 - \frac{1}{\lambda}\, log\left(\frac{2}{3}\right)\right) < D\left(x,\ c_2 + \frac{1}{\lambda}\, log\left(\frac{2}{3}\right)\right) \\ &\implies 2x \ < \ c_1 + c_2 \\ &\implies x \ < \ \frac{c_1+c_2}{2}. \end{aligned} \tag{20}$$

The result follows by observing that Eq. (19) and (20) are identical comparisons.

Hence the theorem! □

3.3.2 Data Generation: Doubly-Exponential Distribution

In order to generate data that follow non-uniform distributions, we made use of a Uniform $(0, 1)$ random variate generator. Data values that follow a Doubly Exponential distribution can be generated by using the expression $\mathbf{x} = c \pm \lambda log|2u|$ where c is the mean of the distribution, λ is the scale parameter, and u is uniform in $U(0, 1)$ [4]. For both the classes, 1,000 points were generated with means c_1 and c_2, and with identical values for λ_1 and λ_2.

3.3.3 Experimental Results: Doubly-Exponential Distribution - 2OS

The CMOS classifier was rigorously tested for a number of experiments with various Doubly Exponential distributions having means c_1 and c_2. In every case, the 2-OS CMOS gave exactly the same classification as that of the Bayesian classifier. The method was executed 50 times with the 10-fold cross validation scheme. The test results are depicted in Table 4. From the experimental results

$\mathbf{c_1}$	0	0	0	0	0	0	0	0	0
$\mathbf{c_2}$	10	9	8	7	6	5	4	3	2
Bayesian	99.75	99.65	99.25	99.05	98.9	97.85	96.8	94.05	89.9
CMOS	99.75	99.65	99.25	99.05	98.9	97.85	96.8	94.05	89.9

Table 4: Classification for the Doubly Exponential Distribution by the CMOS.

and the theoretical analysis, we conclude that the expected values of the first moment of the 2-OS of the Doubly Exponential distribution can always be utilized to yield the exact accuracy as that of the Bayes' bound, even though this is a drastically anti-Bayesian operation.

We now proceed to consider the analogous result for the k-OS.

3.3.4 Theoretical Analysis: Doubly-Exponential Distribution - k-OS

We have seen from Theorem 3 that for the Doubly Exponential distribution, the moments of the 2-OS are sufficient for the classification to attain a Bayes' bound. We shall now consider the scenario when we utilize other k-OS. The formal result pertaining to this is given in Theorem 4.

Theorem 4. *For the 2-class problem in which the two class conditional distributions are Doubly Exponential and identical, the optimal Bayesian classification can be achieved by using symmetric pairs of the n-OS, i.e., the $n-k$ OS for ω_1 and the k OS for ω_2 if and only if $log\left(\frac{2k}{n+1}\right) > \frac{c_1-c_2}{2}$.*

Proof. We shall first show that the expected values of the first moment of the k-order OS for the Doubly Exponential distribution have the form $E[\mathbf{x}_{k,n}] = log\left(\frac{2k}{n+1}\right)$. This result is proven by invoking a formal mathematical induction on k and is omitted here for the present.

We have already solved the case when $n = 2$ and can be seen in Section 3.3.1. Now, we shall consider the case when $n = 4$, for which the possible symmetric OS pairs could be $\langle 1, 4\rangle$ and $\langle 2, 3\rangle$ respectively. Considering the OS pair $\langle 1, 4\rangle$, let u_1 be the point for the percentile $\frac{4}{5}$ of the first distribution, and u_2 be the point for the percentile $\frac{1}{5}$ of the second distribution. Then:

$$\int_{c_1}^{u_1} \frac{\lambda_1}{2}\, e^{-\lambda_1|x-c_1|}dx \;=\; \frac{4}{5} - \frac{1}{2} = \frac{3}{5}, \text{ and} \tag{21}$$

$$\int_{-\infty}^{u_2} \frac{\lambda_2}{2}\, e^{\lambda_2|x-c_2|}dx \;=\; \frac{1}{5}. \tag{22}$$

By straightforward integration and simplifications, we obtain:

$$u_1 = c_1 - \frac{1}{\lambda_1}\, log\left(\frac{2}{5}\right), u_2 = c_2 + \frac{1}{\lambda_2}\, log\left(\frac{2}{5}\right). \tag{23}$$

Arguing in the same way, if we consider the symmetric pair $\langle 2, 3\rangle$, we obtain:

$$u_1 = c_1 - \frac{1}{\lambda_1}\, log\left(\frac{4}{5}\right), u_2 = c_2 + \frac{1}{\lambda_2}\, log\left(\frac{4}{5}\right). \tag{24}$$

The CMOS points for different values for n is given in Table 6.

n	**OS percentiles**	$\mathbf{u_1}$	$\mathbf{u_2}$
2	$\left\langle \frac{2}{3}, \frac{1}{3}\right\rangle$	$c_1 - \frac{1}{\lambda}\, log\left(\frac{2}{3}\right)$	$c_2 + \frac{1}{\lambda}\, log\left(\frac{2}{3}\right)$
4	$\left\langle \frac{4}{5}, \frac{1}{5}\right\rangle$	$c_1 - \frac{1}{\lambda}\, log\left(\frac{2}{5}\right)$	$c_2 + \frac{1}{\lambda}\, log\left(\frac{2}{5}\right)$
4	$\left\langle \frac{3}{5}, \frac{2}{5}\right\rangle$	$c_1 - \frac{1}{\lambda}\, log\left(\frac{4}{5}\right)$	$c_2 + \frac{1}{\lambda}\, log\left(\frac{4}{5}\right)$
6	$\left\langle \frac{6}{7}, \frac{1}{7}\right\rangle$	$c_1 - \frac{1}{\lambda}\, log\left(\frac{2}{7}\right)$	$c_2 + \frac{1}{\lambda}\, log\left(\frac{2}{7}\right)$
6	$\left\langle \frac{5}{7}, \frac{2}{7}\right\rangle$	$c_1 - \frac{1}{\lambda}\, log\left(\frac{4}{7}\right)$	$c_2 + \frac{1}{\lambda}\, log\left(\frac{4}{7}\right)$
6	$\left\langle \frac{4}{7}, \frac{3}{7}\right\rangle$	$c_1 - \frac{1}{\lambda}\, log\left(\frac{6}{7}\right)$	$c_2 + \frac{1}{\lambda}\, log\left(\frac{6}{7}\right)$
8	$\left\langle \frac{8}{9}, \frac{1}{9}\right\rangle$	$c_1 - \frac{1}{\lambda}\, log\left(\frac{2}{9}\right)$	$c_2 + \frac{1}{\lambda}\, log\left(\frac{2}{9}\right)$
8	$\left\langle \frac{7}{9}, \frac{2}{9}\right\rangle$	$c_1 - \frac{1}{\lambda}\, log\left(\frac{4}{9}\right)$	$c_2 + \frac{1}{\lambda}\, log\left(\frac{4}{9}\right)$
8	$\left\langle \frac{6}{9}, \frac{3}{9}\right\rangle$	$c_1 - \frac{1}{\lambda}\, log\left(\frac{6}{9}\right)$	$c_2 + \frac{1}{\lambda}\, log\left(\frac{6}{9}\right)$
8	$\left\langle \frac{5}{9}, \frac{4}{9}\right\rangle$	$c_1 - \frac{1}{\lambda}\, log\left(\frac{8}{9}\right)$	$c_2 + \frac{1}{\lambda}\, log\left(\frac{8}{9}\right)$

Table 5: CMOS values for different values of n.

Our present claim is based on the classification in which we can choose any of the symmetric pairs of the n-OS, i.e., the $n-k$ OS for ω_1 and the k OS for ω_2, where these quantities are $c_1 - log\left(\frac{2k}{n+1}\right)$ and $c_2 + log\left(\frac{2k}{n+1}\right)$ respectively.

It is obvious that an OS value can correctly classify a testing point only when the position is somewhere before the intersection of the curves. This point of intersection of the curves can be obtained by equating them as below:

$$\begin{aligned}
\frac{\lambda_1}{2}e^{-\lambda_1(x-c_1)} = \frac{\lambda_2}{2}e^{-\lambda_2(x-c_2)} \quad &\Longrightarrow \quad \frac{\lambda_1}{\lambda_2} = \frac{e^{\lambda_2(x-c_2)}}{e^{-\lambda_1(x-c_1)}} \\
&\Longrightarrow \quad \frac{\lambda_1}{\lambda_2} = e^{\lambda_1(x-c_1)+\lambda_2(x-c_2)} \\
&\Longrightarrow \quad log(\frac{\lambda_1}{\lambda_2}) = x(\lambda_1+\lambda_2) - \lambda_1 c_1 - \lambda_2 c_2 \\
&\Longrightarrow \quad x = \frac{\lambda_1 c_1 + \lambda_2 c_2 + log\left(\frac{\lambda_1}{\lambda_2}\right)}{\lambda_1+\lambda_2}
\end{aligned} \tag{25}$$

Observe that this equality will reduce to $\frac{c_1+c_2}{2}$ when $\lambda_1 = \lambda_2$.

In order to prove the bounds specified in the statement of the theorem, we enforce the ordering of the OS of the distributions as:

$$c_1 - log\left(\frac{2k}{n+1}\right) < \frac{c_1+c_2}{2} < c_2 + log\left(\frac{2k}{n+1}\right) \tag{26}$$

The LHS of Eq. (26) can easily be simplified to:

$$\begin{aligned}
c_1 - log\left(\frac{2k}{n+1}\right) < \frac{c_1+c_2}{2} < c_2 \quad &\Longrightarrow \quad c_1 - \frac{c_1+c_2}{2} < c_2 < log\left(\frac{2k}{n+1}\right) \\
&\Longrightarrow \quad log\left(\frac{2k}{n+1}\right) > \frac{c_1-c_2}{2}.
\end{aligned} \tag{27}$$

The RHS of Eq. (26) can also be simplified to the same expression, for which the algebraic details are omitted.

The fact that the scheme attains the Bayes' accuracy when these bounds are enforced is now demonstrated by observing that:

$$\begin{aligned}
D(x,u_1) < D(x,u_2) \quad &\Longrightarrow \quad D\left(x, c_1 - log\left(\frac{2k}{n+1}\right)\right) < D\left(x, c_2 + log\left(\frac{2k}{n+1}\right)\right) \\
&\Longrightarrow \quad x - \left(c_1 - log\left(\frac{2k}{n+1}\right)\right) < \left(c_2 - log\left(\frac{2k}{n+1}\right)\right) - x \\
&\Longrightarrow \quad x < \frac{c_1+c_2}{2}.
\end{aligned} \tag{28}$$

Hence the theorem! □

3.3.5 Experimental Results: Doubly-Exponential Distribution - k-OS

The CMOS method has been rigorously tested with different possibilities of k-OS and for various values of n, and the test results are given in Table 6.

No.	**Order(n)**	**Moments**	OS_1	OS_2	**CMOS**	**Pass/Fail**
1	Two	$(\frac{2}{3}, \frac{1}{3})$	$c_1 - \frac{1}{\lambda_1} log\left(\frac{2}{3}\right)$	$c_2 + \frac{1}{\lambda_2} log\left(\frac{2}{3}\right)$	95.2	Passed
2	Three	$(\frac{3}{4}, \frac{1}{4})$	$c_1 - \frac{1}{\lambda_1} log\left(\frac{1}{2}\right)$	$c_2 + \frac{1}{\lambda_2} log\left(\frac{1}{2}\right)$	95.2	Passed
3	Four	$\left(\frac{5-i}{5}, \frac{i}{5}\right), 1 \leq i \leq \frac{n}{2}$	$c_1 - \frac{1}{\lambda_1} log\left(\frac{4}{5}\right)$	$c_2 + \frac{1}{\lambda_2} log\left(\frac{4}{5}\right)$	95.2	Passed
4	Five	$\left(\frac{6-i}{6}, \frac{i}{6}\right), 1 \leq i \leq \frac{n}{2}$	$c_1 - \frac{1}{\lambda_1} log\left(\frac{1}{3}\right)$	$c_2 + \frac{1}{\lambda_2} log\left(\frac{1}{3}\right)$	95.2	Passed
5	Six	$\left(\frac{7-i}{7}, \frac{i}{7}\right) 1 \leq i \leq \frac{n}{2}$	$c_1 - \frac{1}{\lambda_1} log\left(\frac{4}{7}\right)$	$c_2 + \frac{1}{\lambda_2} log\left(\frac{4}{7}\right)$	95.2	Passed
6	Seven	$\left(\frac{8-i}{8}, \frac{i}{8}\right) 1 \leq i \leq \frac{n}{2}$	$c_1 - \frac{1}{\lambda_1} log\left(\frac{1}{4}\right)$	$c_2 + \frac{1}{\lambda_2} log\left(\frac{1}{4}\right)$	95.2	Passed
7	Eight	$\left(\frac{9-i}{9}, \frac{i}{9}\right), 1 \leq i \leq \frac{n}{2}$	$c_1 - \frac{1}{\lambda_1} log\left(\frac{2}{9}\right)$	$c_2 + \frac{1}{\lambda_2} log\left(\frac{2}{9}\right)$	4.8	Failed
8	Eight	$\left(\frac{9-i}{9}, \frac{i}{9}\right), 1 \leq i \leq \frac{n}{2}$	$c_1 - \frac{1}{\lambda_1} log\left(\frac{4}{9}\right)$	$c_2 + \frac{1}{\lambda_2} log\left(\frac{4}{9}\right)$	95.2	Passed
9	Nine	$\left(\frac{10-i}{10}, \frac{i}{10}\right), 1 \leq i \leq \frac{n}{2}$	$c_1 - \frac{1}{\lambda_1} log\left(\frac{3}{5}\right)$	$c_2 + \frac{1}{\lambda_2} log\left(\frac{3}{5}\right)$	95.2	Passed

Table 6: Results of the classification obtained by using the symmetric pairs of the OS for different values of n. The value of c_1 and c_2 were set to be 0 and 3. Note that in every case, the accuracy attained the Bayes' value whenever the conditions stated in Theorem 4 were satisfied.

To clarify the table, consider the row given by Trial No. 5 in which the 6-OS were invoked for the classification. In this case, the possible symmetric OS pairs could be $\langle 1, 6\rangle$, $\langle 2, 5\rangle$, and $\langle 3, 4\rangle$ respectively. Observe that the expected values for the first moment of the k-OS has the form $E[\mathbf{x}_{k,n}] = log\left(\frac{2k}{n+1}\right)$. In every single case, the accuracy attained the Bayes' bound, as indicated by the results in the table.

Now, consider the results presented in the row denoted by Trial No. 7. In this case, the testing attained the Bayes' accuracy for the symmetric OS pairs $\langle 2, 7\rangle$, $\langle 3, 6\rangle$ and $\langle 4, 5\rangle$ respectively. However, the classifier "failed" for the specific 8-OS, when the OS used were $c_1 - \frac{1}{\lambda_1} log\left(\frac{2}{9}\right)$ and $c_2 + \frac{1}{\lambda_2} log\left(\frac{2}{9}\right)$, as these values violate the condition $log\left(\frac{2k}{n+1}\right) > \frac{c_1 - c_2}{2}$, imposed by Theorem 4. Observe that if $log\left(\frac{2k}{n+1}\right) < \frac{c_1 - c_2}{2}$, the symmetric pairs should be reversed to obtain the optimal Bayes' bound.

This concludes our discussion on the use of Order Statistics for the PR of features obeying a Doubly Exponential distribution. The multi-dimensional case is currently being investigated and will be published in a forthcoming paper.

3.4 The Gaussian Distribution

The Normal (or Gaussian) distribution is a continuous probability distribution that is often used as a first approximation to describe real-valued random variables that tend to cluster around a single mean value. It is particularly pertinent due to the so-called Central Limit Theorem. The univariate pdf of the distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

The Gaussian curves for two sets of values for μ and σ is given in Figure 6. As is well known, the optimal Bayesian classifier for equiprobable classes is determined by the point of intersection of the curves, i.e., x^*.

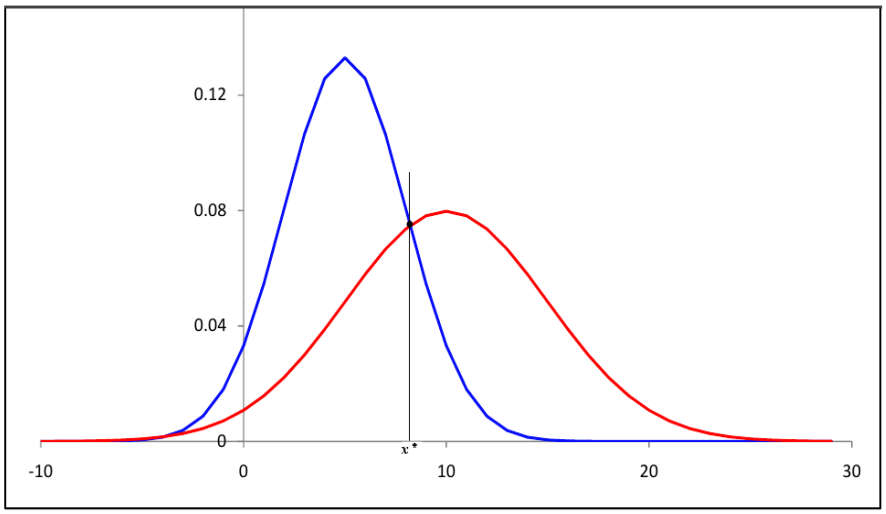


Figure 6: Gaussian Distribution.

We shall now consider the consequence of utilizing CMOS with the 2-OS for classification and again argue the strength of the anti-Bayesian method.

3.4.1 Theoretical Analysis: Gaussian Distribution

Working with the OS of *Normal* distributions is extremely cumbersome because its density function is not integrable in a closed form. One has to resort to tabulated cumulative error functions or to numerical methods to obtain precise percentile values. However, a lot of work has been done in this area for *certain* OS, and can be found in [1, 9, 15, 17], from which we can make some interesting conclusions.

The moments of the OS for the Normal distribution can be determined from the generalized expression:

$$E[\mathbf{x}_{k,n}^r] = \frac{n!}{(k-1)!(n-k)!} \int_{-\infty}^{+\infty} x^r \Phi^{k-1}(x)(1-\Phi(x))^{n-k} \varphi(x) dx,$$

where $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ and $\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt$. From this expression, the expected values of the first moment of the 2-OS can be determined as:

$$E[\mathbf{x}_{1,2}] = \mu - \frac{\sigma}{\sqrt{2\pi}}, \quad \text{and} \tag{29}$$

$$E[\mathbf{x}_{2,2}] = \mu + \frac{\sigma}{\sqrt{2\pi}}, \tag{30}$$

as shown in [1]. Using this result, we now show that for identically distributed classes differing only in the means, the CMOS with 2-OS yields the same Bayesian accuracy, which is the primary thrust of this paper.

Theorem 5. *For the 2-class problem in which the two class conditional distributions are Gaussian and identical, CMOS, the classification using 2-OS, attains the optimal Bayes' bound.*

Proof. As in the previous cases, we shall first show that when the class conditional distributions are Gaussian and identical, the optimal Bayes' bound can be attained by a comparison to the corresponding means, which can then be shown to be equal to the accuracy of the CMOS, whence the theorem is proven.

Without loss of generality, let ω_1 and ω_2 be two classes that follow the Gaussian distribution with means μ_1 and μ_2, and with equal standard deviations, σ. Let u_1 and u_2 be the first moments of the 2-OS, where $u_1 = \mu_1 - \frac{\sigma}{\sqrt{2\pi}}$ and $u_2 = \mu_2 + \frac{\sigma}{\sqrt{2\pi}}$. It is well known that for this scenario the optimal Bayes' classifier can be obtained as:

$$\begin{aligned} p(x|\omega_1)p(\omega_1) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} p(x|\omega_2)p(\omega_2) &\implies \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_1)^2}{2\sigma^2}} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_2)^2}{2\sigma^2}} \\ &\implies x - \mu_1 \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} \mu_2 - x \\ &\implies x \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} \frac{\mu_1 + \mu_2}{2}. \end{aligned} \tag{31}$$

We now prove that:

$$D(x, \mu_1) < D(x, \mu_2) \implies D(x, u_1) < D(x, u_2).$$

We first consider the LHS of the claim. Then,

$$\begin{aligned} D(x, \mu_1) < D(x, \mu_2) &\implies x - \mu_1 < \mu_2 - x \\ &\implies 2x < \mu_1 + \mu_2 \\ &\implies x < \frac{\mu_1 + \mu_2}{2}. \end{aligned} \tag{32}$$

For the result to be proved, we have to also prove that the RHS simplifies to the same expression.

Indeed, this is true because,

$$\begin{aligned} D(x,u_1) < D(x,u_2) &\implies D\left(x,\ u_1\right) < D\left(x,\ u_2\right) \\ &\implies D\left(x,\ \mu_1 - \frac{\sigma}{\sqrt{2\pi}}\right) < D\left(x,\ \mu_2 + \frac{\sigma}{\sqrt{2\pi}}\right) \\ &\implies x - \left(\mu_1 - \frac{\sigma}{\sqrt{2\pi}}\right) < \left(\mu_2 + \frac{\sigma}{\sqrt{2\pi}}\right) - x \\ &\implies 2x < \mu_1 + \mu_2 \\ &\implies x < \frac{\mu_1 + \mu_2}{2}. \end{aligned} \tag{33}$$

The theorem follows! □

3.4.2 Data Generation: Gaussian Distribution

As in the previous cases, we made use of a Uniform $(0, 1)$ random variable generator to generate data values that follow a Gaussian distribution. The expression $\mathbf{z} = \sqrt{-2ln(u_1)}\ cos(2\pi u_2)$, is known to yield data values that follow $N(0, 1)$ [4], from which the data values that follow $N(\mu, \sigma)$ can be generated as $\mathbf{x} = \mu + \mathbf{z}\sigma$, where μ is the mean and σ is the standard deviation of the required distribution. For both the classes, 1,000 points were generated with means μ_1 and μ_2, and with identical values for σ_1 and σ_2.

3.4.3 Experimental Results: Gaussian Distribution

After the data points were generated, the CMOS classifier was rigorously tested for a number of experiments with various Gaussian distributions having means μ_1 and μ_2. In every case, the 2-OS CMOS gave *exactly* the same accuracy as that of the Bayesian classifier. The method was executed 50 times with the 10-fold cross validation scheme. The test results are displayed in Table 7, whence the power of the scheme is clear.

μ_1	0	0	0	0	0	0
μ_2	14	12	10	8	6	4
Bayesian	99.2	96.5	95.1	95	90	85
CMOS	99.2	96.5	95.1	95	90	85

Table 7: Classification of Normally distributed classes by the CMOS 2-OS method for different means.

We believe that the optimal Bayes' bound can also be attained by performing the classification with respect to the k-OS. However, as the density function is not integrable, the expected values of the moments of the k-OS should rather be obtained by invoking a numerical integration. It can be

easily seen that the error function given by:

$$erf(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{\frac{-t^2}{2}} dt, \tag{34}$$

is related to Φ, the cumulative distribution function of the Normal density as per:

$$\Phi(\mathbf{x}) = \frac{1}{2}\left[1 + erf\left(\frac{x}{\sqrt{2}}\right)\right]. \tag{35}$$

The inverse of Φ, named as the *probit* function or the normal quantile function, is expressed as:

$$probit(p) = \sqrt{2}\ erf^{-1}(2p-1). \tag{36}$$

In order to get the exact values of the $\left(\frac{k}{n}\right)^{th}$ percentile, the value computed as per Eq. (34 - 36) should be interpolated.

Alternatively, the $\left(\frac{k}{n}\right)^{th}$ percentile of the normal distribution can be obtained [4] by making use of the inverse of the normal distribution function as:

$$g(u) = \sqrt{-2\ logu} \cdot \frac{A\sqrt{-2\ logu}}{B\sqrt{-2\ logu}}, \tag{37}$$

where $A(x) = \sum_{i=0}^{4} a_i x^i$, $B(x) = \sum_{i=0}^{4} b_i x^i$, and where the coefficients are as shown in Table 8.

i	a_i	b_i
0	-0.3222232431088	0.0993484626060
1	-1.0	0.588581570495
2	-0.342242088547	0.531103462366
3	-0.0204231210245	0.103537752850
4	-0.0000453642210148	0.0038560700634

Table 8: Coefficients for the inverse Normal function.

One can easily see that the $\left(\frac{k}{n}\right)^{th}$ and the $\left(\frac{n-k+1}{n}\right)^{th}$ percentiles of the Normal function obtained in this manner are precisely the CMOS points, which are to be used in the corresponding classification strategy. Using these, the method has been rigorously tested with different possibilities of k-OS and for various values of n, and the test results are given in Table 9. To clarify the table, consider the row given by Trial No. 4 in which the 8-OS were invoked for the classification. In this case, we know that the possible symmetric OS pairs can be $\langle 1, 8\rangle$, $\langle 2, 7\rangle$, $\langle 3, 6\rangle$ and $\langle 4, 5\rangle$ respectively. In every single case, the accuracy attained the Bayes' bound, as indicated by the results in the table.

Now, consider the results presented in the row denoted by Trial No. 5. In this case, the testing attained the Bayes' accuracy for the symmetric OS pairs $\langle 2, 9\rangle$, $\langle 3, 8\rangle$, $\langle 4, 7\rangle$ and $\langle 5, 6\rangle$ respectively.

No.	Order(n)	Moments	CMOS	Pass/Fail
1	Two	$\left(\frac{2}{3}, \frac{1}{3}\right)$	91.865	Passed
2	Four	$\left(\frac{4}{5}, \frac{1}{5}\right)$	91.865	Passed
3	Six	$\left(\frac{6}{7}, \frac{1}{7}\right)$	91.865	Passed
4	Eight	$\left(\frac{8}{9}, \frac{1}{9}\right)$	91.865	Passed
5	Ten	$\left(\frac{10}{11}, \frac{1}{11}\right)$	8.135	Failed
6	Ten	$\left(\frac{9}{11}, \frac{2}{11}\right)$	91.865	Passed
7	Twelve	$\left(\frac{12}{13}, \frac{1}{13}\right)$	8.135	Failed
8	Twelve	$\left(\frac{10}{13}, \frac{3}{13}\right)$	91.865	Passed

Table 9: Results of the classification obtained by using the symmetric pairs of the k-OS for different values of n.

However, the classifier "failed" for the specific 10-OS, when the moments used were $\langle 1, 10 \rangle$. As in the Uniform and Doubly Exponential distributions, if the chosen moments are in the near proximity of the Bayesian classifier, they do not possess sufficient information and capability to classify a testing point inasmuch as both these moments would be almost equidistant from the testing point. However, unlike the Uniform and Doubly Exponential distributions, since the Gaussian pdf is not integrable, it is not possible to derive a closed form expression for this condition. The tabulated cases, however, demonstrate this phenomenon.

A detailed explanation of how this is related to BI issues can be found in [19].

Classification for the multi-dimensional classes is also being investigated, and will be published in a forthcoming paper.

4 Conclusions and Future Work

In this paper, we have shown that the optimal Bayes' bound can be obtained by an "anti-Bayesian" approach named CMOS, Classification by Moments of Order Statistics. We have proved that we can achieve classification by working with a *very few* (sometimes as small as two) points *distant* from the mean. Further, if these points are determined by the *Order Statistics* of the distributions, the optimal Bayes' bound can be attained. The claim has been proved for many uni-dimensional, and some multi-dimensional distributions within the exponential family, and the theoretical results have been verified by rigorous experimental testing.

With regard to future work, we believe that the work that can be done to investigate "anti-Bayesian" and OS-based classification is almost unbounded. First of all, more research and theoret-

ical analysis can be done for the multi-dimensional distributions. Further, even though this paper deals with the 2-class problems, multi-class problems are also currently being investigated. Apart from this, the method can be applied to non-parametric distributions also. Finally, we believe that the CMOS can be applied for clustering, and for pattern recognition in which kernel-based methods are invoked to achieve the classification.

References

[1] M. Ahsanullah and V. B. Nevzorov. *Order Statistics: Examples and Exercises*. Nova Science Publishers, Inc, 2005.

[2] C. L. Chang. Finding Prototypes for Nearest Neighbor Classifiers. In *IEEE Transactions on Computing*, volume 23, pages 1179–1184, 1974.

[3] P. A. Devijver and J. Kittler. On the Edited Nearest Neighbor Rule. In *Fifth International Conference on Pattern Recognition*, pages 72–80, December 1980.

[4] Luc Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, 1986.

[5] W. Duch. Similarity based methods: a general framework for Classification, Approximation and Association. *Control and Cybernetics*, 29(4):937–968, 2000.

[6] G. M. Foody. Issues in Training Set Selection and Refinement for Classification by a Feedforward Neural Network. In *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, pages 409–411, 1998.

[7] S. Garcia, J. Derrac, J. Ramon Cano, and F. Herrera. Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 34(3):417–435, 2012.

[8] G. W. Gates. The Reduced Nearest Neighbor Rule. In *IEEE Transactions on Information Theory*, volume 18, pages 431–433, 1972.

[9] Z. Grudzien and D. Szynal. Characterizations of Distributions by Moments of Order Statistics when the Sample Size is Random. *Applications Mathematicae*, 23:305–318, 1995.

[10] P. E. Hart. The Condensed Nearest Neighbor Rule. In *IEEE Transactions on Information Theory*, volume 14, pages 515–516, 1968.

[11] http://sci2s.ugr.es/pr/.

[12] S. Kim and B. J. Oommen. On Using Prototype Reduction Schemes and Classifier Fusion Strategies to Optimize Kernel-Based Nonlinear Subspace Methods. In *IEEE Transactions on Pattern Analysis and machine Intelligence*, volume 27, pages 455–460, 2005.

[13] L. I. Kuncheva, J. C. Bezdek, and R. P. W. Duin. Decision Templates for Multiple Classifier Fusion: An Experimental Comparison. *Pattern Recognition - The Journal of the Pattern Recognition Society*, 34:299–314, 2001.

[14] G. Li, N. Japkowicz, T. J. Stocki, and R. K. Ungar. Full Border Identification for Reduction of Training Sets. In *Proceedings of the Canadian Society for computational studies of intelligence, 21st conference on Advances in artificial intelligence*, pages 203–215, 2008.

[15] G. D. Lin. Characterizations of Continuous Distributions via Expected values of two functions of Order Statistics. *Sankhya: The Indian Journal of Statistics*, 52:84–90, 1990.

[16] K. W. Morris and D. Szynal. A goodness-of-fit for the Uniform Distribution based on a Characterization. *Journal of Mathematical Science*, 106:2719–2724, 2001.

[17] S. Nadarajah. Explicit Expressions for Moments of Order Statistics. *Statistics and Probability Letters*, 78:196–205, 2008.

[18] G. L. Ritter, H. B. Woodruff, S. R. Lowry, and T. L. Isenhour. An Algorithm for a Selective Nearest Neighbor Rule. In *IEEE Transactions on Information Theory*, volume 21, pages 665–669, 1975.

[19] A. Thomas. *Pattern Classification using Novel Order Statistics and Border Identification Methods*. PhD thesis, School of Computer Science, Carleton University, 2012. (To be Submitted).

[20] Y. Too and G. D. Lin. Characterizations of Uniform and Exponential Distributions. *Academia Sinica*, 7(5):357–359, 1989.

[21] I. Triguero, J. Derrac, S. Garcia, and F. Herrera. A Taxonomy and Experimental Study on Prototype Generation for Nearest Neighbor Classification. *IEEE transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews*, 42:86–100, 2012.