12-2004

Consistency Results for the ROC Curves of Fused Classifiers

Kristopher S. Bjerkaas

AFIT/GAM/ENC/05-01

CONSISTENCY RESULTS FOR THE ROC CURVES OF FUSED CLASSIFIERS

THESIS

Kristopher S. Bjerkaas, S.B.

AFIT/GAM/ENC/05-01

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

Disclaimer

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, the Department of Defense, or the United States Government.

AFIT/GAM/ENC/05-01

CONSISTENCY RESULTS FOR THE ROC CURVES OF FUSED CLASSIFIERS

THESIS

Presented to the Faculty of the Graduate School of Engineering and Management

of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the

Requirements for the Degree of

Master of Science in Applied Mathematics

Kristopher S. Bjerkaas, S.B.

Air Force Institute of Technology

Wright-Patterson AFB, Ohio

December 2004

AFIT/GAM/ENC/05-01

CONSISTENCY RESULTS FOR THE ROC CURVES OF FUSED CLASSIFIERS

Kristopher S. Bjerkaas, S.B.

Approved:

Dr. Mark E. Oxley Committee Chairman	Date
Dr. Kenneth W. Bauer, Jr. Committee member	Date

Table of Contents

Page

Table of Contents iv

List of Figures vii

List of Tables ix

Abstract x

I. Introduction 1

1.1 General Discussion 1

1.2 Problem Description 1

1.2.1 Background 1

1.2.2 Problem Statement 4

1.3 Research Objectives 4

II. Literature Review 5

2.1 Overview 5

2.2 ROC Curves 5

2.2.1 Description 5

2.2.2 ROC Curve Construction 6

2.2.3 Mathematical Description of ROC Curves 8

2.3 ROC Curve Comparison 10

2.3.1 ROC Metrics 10

2.3.2 ROC Convergence 12

Page

2.4 Classifier and ROC Fusion . . . 12

2.4.1 Classifier Fusion Rules . . . 12

2.4.2 Within and Across Fusion . . . 14

2.4.3 Independence . . . 17

2.4.4 ROC Fusion . . . 18

III. Derivation and Methodology . . . 20

3.1 Introduction . . . 20

3.2 Framework and Notation . . . 20

3.3 Convergence for Continuous Substitution Transformation . . . 20

3.4 Convergence Framework Application . . . 26

IV. Analysis and Findings . . . 29

4.1 Overview . . . 29

4.2 Experimental Design . . . 30

4.2.1 Classifiers . . . 30

4.2.2 Feature Data . . . 30

4.2.3 ROC Estimates . . . 30

4.2.4 Empirical ROC Curve Construction . . . 32

4.2.5 Test Metric . . . 32

4.2.6 Within Fusion . . . 33

4.2.7 Across Fusion . . . 34

4.3 Results . . . 36

Page

4.3.1 Within-OR (Non-parametric) 36

4.3.2 Across-OR (Non-parametric) 37

4.3.3 Within-OR (Parametric) 39

4.3.4 Convergence as a Function of Sample Size 41

V. Summary and Recommendations 47

5.1 Summary of Contributions 47

5.2 Recommendations for Future Research 47

Bibliography 49

Vita 51

List of Figures

Page

1.1 Single classifier system. 2

1.2 Multiple classifier system with label fusion. 3

2.1 A typical ROC curve. 6

2.2 Target and non-target pdfs for a two-class system. 7

2.3 ROC trajectory and ROC curve projection. 8

2.4 Within-MCS. 15

2.5 Across-MCS. 15

2.6 Compact notation for MCS. 16

4.1 Classifier A_θ within fusion case. 33

4.2 Classifier B_ϕ within fusion case. 34

4.3 Classifier B_ϕ across fusion case. 35

4.4 Average metric distances for the within-OR (non-parametric case) for varying n. 38

4.5 Average metric distances for the across-OR (non-parametric case) for varying n. 40

4.6 Average metric distances for the within-OR (parametric case) for varying n. 42

4.7 Average metric distance as a function of sample size - within-OR (non-parametric case) 44

Figure Page

4.8 Average metric distance as a function of sample size - within-OR (parametric case) 45

4.9 Average metric distance as a function of sample size - across-OR (non-parametric case) 46

List of Tables

Page

2.1 Classification outcomes. 5

2.2 Fusion categories. 13

2.3 Boolean fusion rules. 14

4.1 Within-OR (non-parametric): average metric distances for $\hat{f}_{\mathbb{A}}^{(n)}$. 37

4.2 Within-OR (non-parametric): average metric distances for $\hat{f}_{\mathbb{B}}^{(n)}$. 37

4.3 Within-OR (non-parametric): average metric distances for $\hat{f}_{\mathbb{C}}^{(n)}$. 37

4.4 Across-OR (non-parametric): average metric distances for $\hat{f}_{\mathbb{A}}^{(n)}$. 39

4.5 Across-OR (non-parametric): average metric distances for $\hat{f}_{\mathbb{B}}^{(n)}$. 39

4.6 Across-OR (non-parametric): average metric distances for $\hat{f}_{\mathbb{C}}^{(n)}$. 39

4.7 Parametric vs. non-parametric: average metric distances for $\hat{f}_{\mathbb{A}}^{(n)}$. 41

4.8 Parametric vs. non-parametric: average metric distances for $\hat{f}_{\mathbb{B}}^{(n)}$. 41

4.9 Parametric vs. non-parametric: average metric distances for $\hat{f}_{\mathbb{C}}^{(n)}$. 41

Abstract

The U.S. Air Force is researching the fusion of multiple sensors and classifiers. Given a finite collection of classifiers to be fused one seeks a new classifier with improved performance. An established performance quantifier is the Receiver Operating Characteristic (ROC) curve, which allows one to view the probability of detection versus the probability of false alarm in one graph. Previous research shows that one does not have to perform tests to determine the ROC curve of this new fused classifier. If the ROC curve for each individual classifier has been determined, then formulas for the ROC curve of the fused classifier exist for certain fusion rules. This will be an enormous saving in time and money since the performance of many fused classifiers can be determined without having to perform tests on each one.

In reality only finite data is available so only an estimated ROC curve can be constructed. It has been proven that estimated ROC curves will converge to the true ROC curve in probability. This research examines if convergence is preserved when these estimated ROC curves are fused. It provides a general result for fusion rules that are governed by a Lipschitz continuous ROC fusion function and establishes a metric that can be used to prove this convergence. This framework is then applied to the OR fusion rule, as well as an example study. The study examines two ROC curves, estimated both parametrically and non-parametrically, fused with the OR rule.

CONSISTENCY RESULTS FOR THE ROC CURVES OF FUSED CLASSIFIERS

I. Introduction

1.1 General Discussion

Combat identification (CID), the ability to detect and classify friend versus foe, has moved to the forefront of technological challenges facing the U.S. military today. Even while the military superiority of coalition forces has been overwhelming in recent conflicts, and losses have been relatively low, an increasing percentage of losses has been due to friendly fire. The U.S. military has invested a large amount of resources to solve this problem. Reduced, ideally zero, fratricide rate and increased lethality are the goals. To that end, new sensors are being developed and fielded that can more effectively exploit target signature data to more confidently make the determination of hostile or friendly. Building hardware capable of extracting this signature information is one part of the solution, the other part is properly classifying it.

Many different classification techniques have been developed over the years. One popular technique is Fisher's linear discriminant, where multi-dimensional data from two classes are projected onto a one-dimensional space making for easy discrimination [7]. Another approach is to utilize a neural network that learns how to classify data [24]. Along the same lines, Support Vector Machines, borrowing a page from statistical learning theory, develop a classifier by minimizing the training set error [21]. With the availability of many sensors and many classification techniques, military CID systems generally do not employ just a single classifier system, but seek to optimize performance by fusing the classification decisions of multiple systems.

1.2 Problem Description

1.2.1 Background.

A classifier system in its most elementary form consists of a sensor, a processor, and a classifier. Figure 1.1 depicts this notional classifier system. The sensor starts by collecting raw data on an event it observes. This could be a thermal sensor collecting temperature data or a radar collecting radio frequency returns. The processor then extracts a feature from this raw data that is salient to

discrimination. This could be a temperature profile or a radar cross section. Finally the classifier applies its decision logic to the feature set, classifying the observed event. The system depicted in Figure 1.1 is a two-state classifier, where the decision label is either target or non-target.

Multiple classifier systems (MCSs) seek to increase the performance of individual classifiers by intelligently fusing their outputs [17]. Figure 1.2 depicts a notional MCS. In the depicted MCS, two single classifier systems each classify an observed event and their decision labels are fused using some fusion rule. In general the classifiers are assumed to be independent, although a significant amount of research has been done for the case where classifiers are correlated [15].

There are many advantages to MCSs. For example, each individual classifier system can focus on a different type of feature data, such as length or temperature gradient. They can also be specifically trained to classify different target types; one could specialize in identifying trucks while another could specialize in tanks. With this in mind, it is important that fusion strategies optimize the strengths and minimize the weaknesses of the individual classifiers.

Researchers can compare classifier performance by constructing receiver operating characteristic (ROC) curves from experimental data. To generate the ROC curve for an MCS would typically require additional experiments beyond those used to produce the ROCs for the individual classifiers. Oxley and Bauer, however, demonstrated that this is not always required and that for certain fusion rules, the ROC curve for an MCS can be determined solely from the ROC curves of the individual classifiers [16]. Hill further contributed to this area by devising a matrix-based approach for fusing ROC curves with any Boolean rule [13].

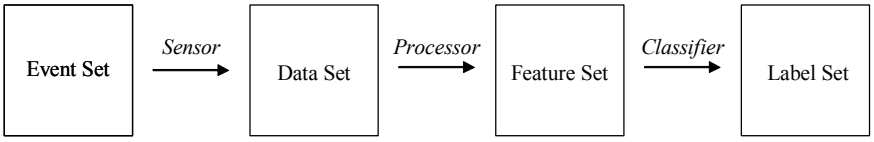

Figure 1.1. Single classifier system.

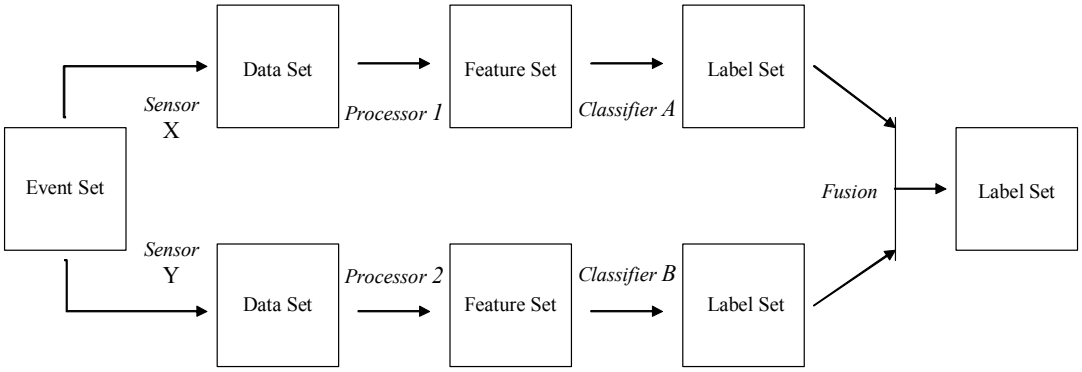

Figure 1.2. Multiple classifier system with label fusion.

1.2.2 Problem Statement.

In real world applications, the true ROC curve is seldom, if ever, known. ROC curves must be estimated, based on sample data. This fact is generally disregarded, and an estimated ROC curve is treated as if it were the true ROC curve representing the true performance of the classifier. Alsing examined this assumption and proved that estimated ROC curves do converge to the true ROC curve in the probability sense [1]. This thesis extends Alsing's convergence work to the ROC curves of fused classifiers.

1.3 Research Objectives

The goal of this research is to provide consistency results for the ROC curves of fused classifiers. The approach requires a mathematical framework to be developed that can be used to prove that fused empirical ROC curves do converge to the true fused ROC curve. A general class of fusion rules will be investigated using this framework, in addition to a specific examination of the OR fusion rule. For this research, independence of the classifiers is assumed.

II. Literature Review

2.1 Overview

This chapter reviews the literature pertinent to receiver operating characteristic (ROC) curves and their use in evaluating classifier performance. Section 2.2 provides a description of ROC curves, demonstrates how their graphical representation enables a quick visual comparison of classifier performance, and provides a mathematical description of their construction. Section 2.3 discusses the notion of ROC convergence and introduces the metrics needed to compare ROC curves. Much of the work done with ROC curves assumes that a limiting or true ROC curve exists, so this is an important contribution. Finally Section 2.4 reviews different methods of fusing classifiers and presents an important result that allows two ROC curves to be fused analytically.

2.2 ROC Curves

2.2.1 Description.

Consider a two-class classifier system that attempts to classify targets of interest (*tar* class) and non-targets (*non* class). There are four possible outcomes when this system attempts to classify an object. When a *tar* class object is observed, it can be correctly labeled a *tar* or incorrectly labeled a *non*. Likewise, when a *non* class object is observed, it can be correctly labeled a *non* or incorrectly labeled a *tar*. Table 2.1 summarizes these outcomes with their associated terminologies and conditional probabilities.

Table 2.1. Classification outcomes.

Outcome	Terminology	Conditional Probability
tar labeled *tar*	True Positive (TP)	$P_{TP} = \Pr(\text{labeled } tar \mid tar \text{ present})$
non labeled *tar*	False Positive (FP)	$P_{FP} = \Pr(\text{labeled } tar \mid non \text{ present})$
tar labeled *non*	False Negative (FN)	$P_{FN} = \Pr(\text{labeled } non \mid tar \text{ present}).$
non labeled *non*	True Negative (TN)	$P_{TN} = \Pr(\text{labeled } non \mid non \text{ present}).$

The ROC curve provides a visual description of classifier performance by graphically representing the trade-off between P_{TP} and P_{FP} (Figure 2.1). Commonly, the probability of true positive

is referred to as the hit rate or the probability of detection, while the probability of false positive is often referred to as the false alarm rate or probability of false alarm (as in the figure). The ROC curve is constructed by varying the decision thresholds internal to the classifier system and measuring the observed hit and false alarm rate. A very conservative decision threshold will yield a low hit rate and low false alarm rate. An aggressive decision threshold will yield a high hit rate, but generally at the expense of a high false alarm rate.

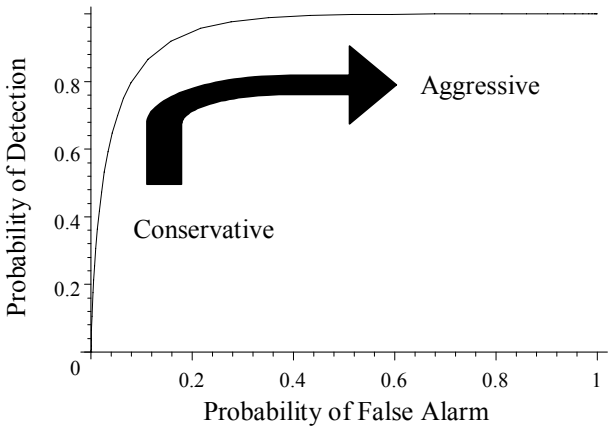

Figure 2.1. A typical ROC curve.

When considering the total probability for each condition (i.e., the condition when a *tar* is present and the condition when a *non* is present), the following relationships are observed,

$$\begin{aligned} P_{TP} + P_{FN} &= 1 \\ P_{FP} + P_{TN} &= 1. \end{aligned}$$

So not only does the ROC curve graphically relate P_{TP} and P_{FP}, but, by virtue of their complements, provides information about P_{TN} and P_{FN} as well.

2.2.2 ROC Curve Construction.

Consider a simple two-class classifier system, like the one from Section 2.2.1, that evaluates a single feature $x \in \mathbb{R}$. Figure 2.2 shows the distribution of this feature for both the *tar* and *non*

classes. Let X be a real-valued random variable and $p(x)$ represent its probability density function (pdf). The distribution of x for the *tar* class is the conditional pdf $p(x|tar)$ and the distribution of x for the *non* class is the conditional pdf $p(x|non)$.

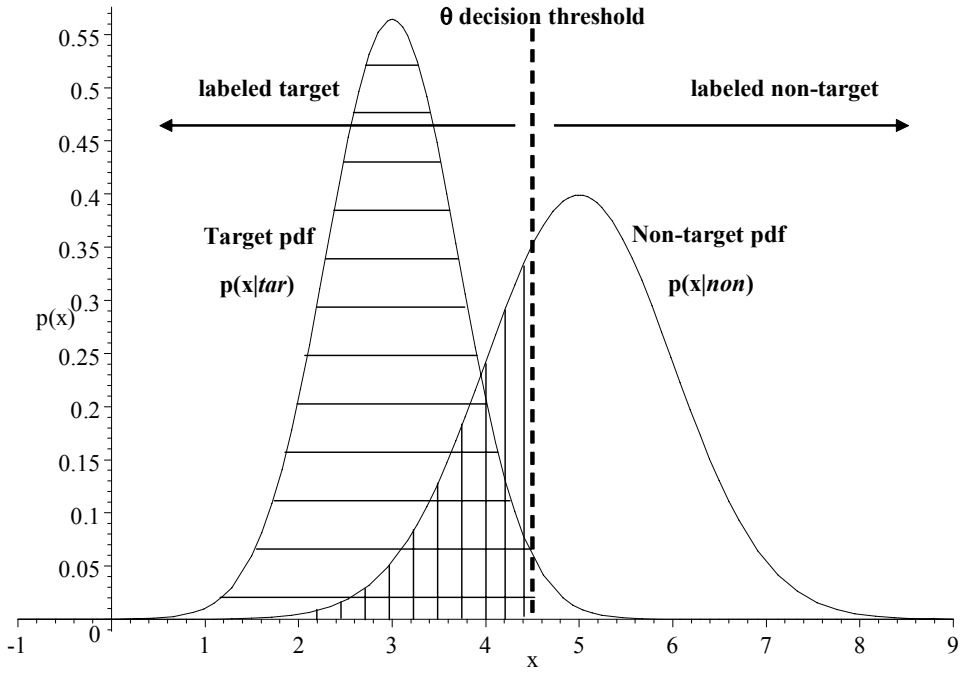

Figure 2.2. Target and non-target pdfs for a two-class system.

In this feature space, lower values of x are considered a stronger indication of a target. So the decision labels are

$$\begin{array}{ll} tar & \text{if} \quad x < \theta, \\ non & \text{if} \quad x \geq \theta, \end{array}$$

where θ is the decision threshold. The horizontal shaded area represents the probability of true positive at θ, while the vertical shaded area represents the probability of false positive at θ. The ROC curve simply is the plot of each probability pair (P_{FP}, P_{TP}) for all θ values.

2.2.3 Mathematical Description of ROC Curves.

Since P_{FP} and P_{TP} are functions of θ, the ROC curve is actually the projection of a three-dimensional trajectory in (θ, P_{TP}, P_{FP})-space onto the two-dimensional (P_{TP}, P_{FP})-plane. This is depicted in Figure 2.3. The 3-dimensional trajectory is known as the ROC trajectory. Alsing defines the ROC trajectory as

$$t = \{(\theta, P_{FP}(\theta), P_{TP}(\theta)) : \theta \in \Theta\},$$

where Θ is the admissible threshold set [1]. The admissible threshold set for the random variable X is the set $\Theta = (a, b) \subset \mathbb{R}$ such that

$$\begin{aligned} \lim_{\theta \to a^+} P_{FP}(\theta) &= 0 \text{ and } \lim_{\theta \to a^+} P_{TP}(\theta) = 0, \\ \lim_{\theta \to b^-} P_{FP}(\theta) &= 1 \text{ and } \lim_{\theta \to b^-} P_{TP}(\theta) = 1. \end{aligned}$$

The ROC curve f is defined as the projection of t onto the (P_{TP}, P_{FP})-plane,

$$f = \{(P_{FP}(\theta), P_{TP}(\theta)) : \theta \in \Theta\}. \tag{2.1}$$

Some other properties of the ROC curve are that f is a non-decreasing and upper semi-continuous function of θ.

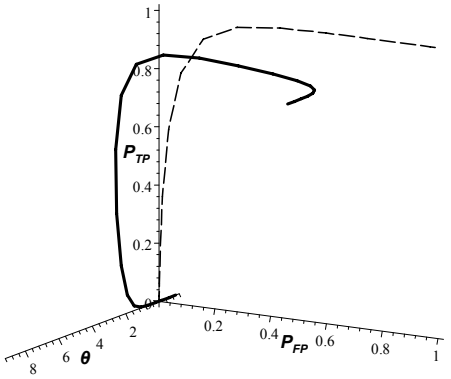

Figure 2.3. ROC trajectory and ROC curve projection.

In applications, the probabilities P_{TP} and P_{FP} cannot be determined exactly since only finite data is available. Hence, the true ROC curve cannot be generated, but must be estimated. These estimated ROC curves are constructed empirically. There are three methods commonly used to determine the estimates $\hat{P}_{TP}$ and $\hat{P}_{FP}$:

1. Assume the family of the underlying distribution (binomial or normal, for instance) [10] is known.
2. Estimate the unknown distribution based on the sample data [12].
3. Calculate the observed true positive and false positive frequencies for varying θ.

Methods 1 and 2 are referred to as parametric methods, since they require key parameters of the statistical distribution to be known or estimated. Method 3 is referred to as a non-parametric method, since it does not specifically require knowledge of the underlying distribution. This last method is the one most commonly used in practice, although this research will investigate both parametric and non-parametric methods.

The estimated probabilities, $\hat{P}_{TP}$ and $\hat{P}_{FP}$, are random variables since they depend on the data used to estimate them. Let ω be the specific instantiation of this data drawn from all possible sets of observed data Ω (also known as the uber-event set or population set). The observed data itself can be represented as a set of feature vectors $\mathbf{x}_i, i = 1$ to $2n$, where there are n observations from both classes. The classifier evaluates the feature vectors for a specific decision threshold $\theta \in \Theta$. Therefore, each probability estimate can be expressed as a function of ω and θ, for n observations:

$$\hat{P}_{TP}^{(n)}(\omega, \theta) \text{ and } \hat{P}_{FP}^{(n)}(\omega, \theta).$$

The estimated, or empirical, ROC curve then is given by

$$\hat{f}^{(n)}(\omega) = \{(\hat{P}_{TP}^{(n)}(\omega, \theta), \hat{P}_{FP}^{(n)}(\omega, \theta)) : \theta \in \Theta\}.$$

2.3 ROC Curve Comparison

2.3.1 ROC Metrics.

In the automatic target recognition community, it is a commonly held assumption that in the case of unlimited data, a limiting ROC curve exists [2]. As n becomes large, $\hat{f}^{(n)}(\omega)$ should converge to this limiting ROC curve f. Alsing provided a proof for ROC convergence and in doing so developed some very valuable metrics for ROC comparison [1].

Fristedt and Gray, with a slight difference in notation, provide the following definition of a metric and metric space [8].

Definition 2.1 (Metric Space.) A metric space (S, d) consists of a function $d : S \times S \to \mathbb{R}$ defined on set S that satisfies the following properties:

1. $d(x, y) \geq 0$ for all $x, y \in S$;
2. $d(x, y) = 0$ if and only if $y = x$;
3. $d(x, y) = d(y, x)$ for all $x, y \in S$;
4. $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y \in S$.

The function d is called the metric.

As an example, consider the following family of metrics defined on $S = \mathbb{R}^2 = \{\mathbf{x} = (x_1, x_2) | x_i \in \mathbb{R}\}$,

$$\rho_q(\mathbf{x}, \mathbf{y}) = (|x_1 - y_1|^q + |x_2 - y_2|^q)^{\frac{1}{q}} \tag{2.2}$$

for each $q \in [1, \infty)$. This family of metrics will be used extensively in this research. A function that adheres to all but property 2 is known as a *pseudo-metric* and is associated with a *pseudo-metric space*. As an example, area under the curve (AUC), is commonly used to compare ROC curves. The difference in AUC for two curves, f and g, is a pseudo-metric since the difference could equal zero for some $f \neq g$.

Definition 2.2 (Equivalent Metrics.) Let d_α and d_β be two metrics defined on S. The metrics d_α and d_β are equivalent if there exists constants $k, K > 0$ such that

$$k d_\beta(x, y) \leq d_\alpha(x, y) \leq K d_\beta(x, y) \text{ for all } x, y \in S.$$

Furthermore, it can be shown that all metrics on $\mathbb{R}^2$ are equivalent [1].

Consider two ROC curves f and g with the same threshold set Θ and represented by

$$\begin{aligned} f &= \{(P_{TP}^{(f)}(\theta), P_{FP}^{(f)}(\theta)) : \theta \in \Theta\} \\ &= \{\mathbf{P}^{(f)}(\theta) : \theta \in \Theta\}, \\ g &= \{(P_{TP}^{(g)}(\theta), P_{FP}^{(g)}(\theta)) : \theta \in \Theta\} \\ &= \{\mathbf{P}^{(g)}(\theta) : \theta \in \Theta\}. \end{aligned}$$

Alsing proposed the following metric on ROC curves for $1 \leq r < \infty$, where ρ is any metric on $\mathbb{R}^2$,

$$d_{\rho,r}(f,g) = \left(\int_\Theta \rho \left(\mathbf{P}^{(f)}(\theta), \mathbf{P}^{(g)}(\theta) \right)^r d\theta \right)^{\frac{1}{r}}. \tag{2.3}$$

One metric from this family,

$$d_{\rho_1,1}(f,g) = \int_\Theta \rho_1 \left(\mathbf{P}^{(f)}(\theta), \mathbf{P}^{(g)}(\theta) \right) d\theta \tag{2.4}$$

provides the total metric distance between f and g. This metric can be normalized to a value between 0 and 1 by dividing by the measure of Θ, $\mu(\Theta)$. This is known as the average metric distance. Like the difference in AUC pseudo-metric, average metric distance takes on values between 0 and 1. For both metrics, smaller values indicate similar overall performance between two classifier systems, and larger values indicate differing overall performance. The important distinction though is that an average metric distance of 0 implies that f and g are the same curve; while, as stated previously, a difference in AUC of 0 is ambiguous as to the difference. This last fact is why $d_{\rho_1,1}$ is utilized in proving convergence. For empirical ROC curves, the discrete form is often employed,

$$\text{average metric distance} \approx \frac{\sum_{i=1}^{m} \rho_1 \left(\mathbf{P}^{(f)}(\theta_i), \mathbf{P}^{(g)}(\theta_i) \right)}{m}, \tag{2.5}$$

where m is the total number of evenly spaced discrete θ_i values in Θ.

2.3.2 ROC Convergence.

If empirical ROC curves do not converge to a limiting ROC curve, then there is no guarantee of consistency in ROC comparisons [1]. To that end, Alsing developed the ROC Convergence Theorem to prove that empirical ROC curves do, indeed, converge provided certain conditions. The most important condition is that as more feature data is collected, it is collected in such a way that it adequately fills the feature space.

Let $S \subset \mathbb{R}^v$, where v is the number of elements in the feature vector $\mathbf{x}$, and S is the total feature set. Let $\mathcal{D}^{(n)} \subset S$ be the set of n feature vectors collected per class, i.e., $\mathcal{D}^{(n)} = \{\mathbf{x}_i \in S : i = 1, ..., 2n\}$ when there are two classes. To ensure that as more feature data is collected it spans across the entire feature space and not just a single subset associated with one label, Alsing requires the sequence of sets, $\{\mathcal{D}^{(n)}\}$, to converge to S in the Hausdorff metric [4]. Under this condition, the sequence of empirical ROC curves, $\{\hat{f}^{(n)}\}$, converges to a limiting ROC curve f.

Theorem 2.3 ***(ROC Convergence [1].)*** *If $\{\mathcal{D}^{(n)}\}$ converges to S in the Hausdorff sense, i.e., given $\varepsilon > 0$, there exists N such that for all $n > N$, $d_H(\mathcal{D}^{(n)}, S) < \varepsilon$, then $\{\hat{f}^{(n)}(\omega)\}$ converges to f in probability, i.e., given $\varepsilon > 0$, there exists N such that for all $n > N$,*

$$\Pr\left(\left\{\omega \in \Omega : d_{p,r}\left(\hat{f}^{(n)}(\omega), f\right) \geq \varepsilon\right\}\right) < \varepsilon.$$

The proof is rather lengthy, so only an outline will be provided. The proof involves four steps:

1. Prove that $\hat{P}_{TP}^{(n)}(\omega, \theta)$ is a consistent estimator for $P_{TP}(\theta)$ and that $\hat{P}_{FP}^{(n)}(\omega, \theta)$ is a consistent estimator for $P_{FP}(\theta)$.
2. Prove pointwise convergence for the estimated probability pair $\left(\hat{P}_{TP}^{(n)}(\omega, \theta), \hat{P}_{FP}^{(n)}(\omega, \theta)\right)$.
3. Prove that the integral of a real-valued random variable converges.
4. Prove that the sequence of ROC curves $\{\hat{f}^{(n)}(\omega)\}$ converges.

2.4 Classifier and ROC Fusion

2.4.1 Classifier Fusion Rules.

There is often a performance advantage to be gained by fusing the results of individual classifiers. Consider Bayesian classifiers that estimate the posterior probability of a particular observation belonging to a certain class. Since this estimate is based on observed data, there is inherent

sample variance associated with these estimates. When these estimates are averaged over several Bayesian classifiers, this variance is reduced [18]. Also, certain classifiers may have better performance against specific targets or in certain situations [20]. In this case, a fusion strategy could be devised that weights a specific classifier's decision more heavily under certain conditions.

According to Dasarathy, classifier decision fusion is simply a subset of sensor fusion [6]. He notes that fusion can be accomplished at the data, feature, or decision level with the following inputs and outputs:

1. Data in $\longmapsto$ Data out,
2. Data in $\longmapsto$ Feature out,
3. Feature in $\longmapsto$ Feature out,
4. Feature in $\longmapsto$ Decision out,
5. Decision in $\longmapsto$ Decision out.

Thorsen describes data fusion using category theory [19]. In category theory, objects are represented by boxes and mappings are represented by arrows. Referencing Figure 1.2; data, feature, and label fusion then is the fusion of objects; while sensor, processor, and classifier fusion is the fusion of mappings. It is interesting to note that with the exception of sensor fusion, Dasarathy's and Thorsen's fusion categories parallel each other (Table 2.2). Sensor fusion would be the equivalent of "Event in $\rightarrow$ Data out" fusion, which Dasarathy does not categorize.

Table 2.2. Fusion categories.

Dasarathy's I/O Category	Thorsen's Object/Mapping Fusion
Data/Data	Data Fusion
Data/Feature	Processor Fusion
Feature/Feature	Feature Fusion
Feature/Decision	Classifier Fusion
Decision/Decision	Label Fusion

Decision, or label, fusion is the focus of this research. There are many advantages to this type of fusion. There is no concern about the compatibility of the data, nor the specific architecture used to design individual classifiers when the fuser only considers discrete decision values. A practical consideration may also be the amount of bandwidth or processing capability the fusion center has

[22]. Once again, a decision fuser is not burdened by having to transmit or process a large amount of raw or feature data.

The simplest method for fusing data is to use Boolean rules. Consider the case where there are two classifiers, A and B. A fusion approach that wants to maximize the number of true positives, at the possible expense of an increased false positive rate, will make a target declaration if either A *or* B indicates a target. This is known as the OR fusion rule and will be denoted as $A \vee B$. A more conservative approach that tries to maintain a low false positive rate will make a target declaration only if both A *and* B indicate a target. This is known as the AND fusion rule and will be denoted as $A \wedge B$. Fusion can be extended to include a third classifier C. In this case, the majority vote fusion rule can be used. The majority vote, in the three classifier case, requires two of the three individual classifiers to agree. The formula for this rule is $(A \wedge B) \vee (A \wedge C) \vee (B \wedge C)$. Table 2.3 summarizes these fusion rules. Of course, these rules can be extended to any number of classifiers.

Table 2.3. Boolean fusion rules.

Fusion Rule	Notation
2-classifier OR	$A \vee B$
2-classifier AND	$A \wedge B$
3-classifier Majority Vote	$(A \wedge B) \vee (A \wedge C) \vee (B \wedge C)$

2.4.2 Within and Across Fusion.

Hill discusses two types of fusion on MCSs: *across* and *within* [13]. Figures 2.4 and 2.5 show the block diagrams that Hill developed for both ROC fusion types. In the within-MCS, the event set E is partitioned into two event subsets: targets of interest (E_{tar}) and non-targets (E_{non}). Each individual classifier system seeks to classify the *same* type of target. The decision from each classifier is then sent to the fusion center. In the across-MCS, the event set E is partitioned into three subsets: target type 1 ($E_{tar,1}$), target type 2 ($E_{tar,2}$), and non-targets (E_{non}). Each individual classifier system seeks *different* types of targets (e.g., one is trained to detect tanks, while another is trained to detect troop carriers). The decision from each classifier is sent to the fusion center. Note that for ease of notation, the processor is assumed to be part of the sensor in these MCS representations.

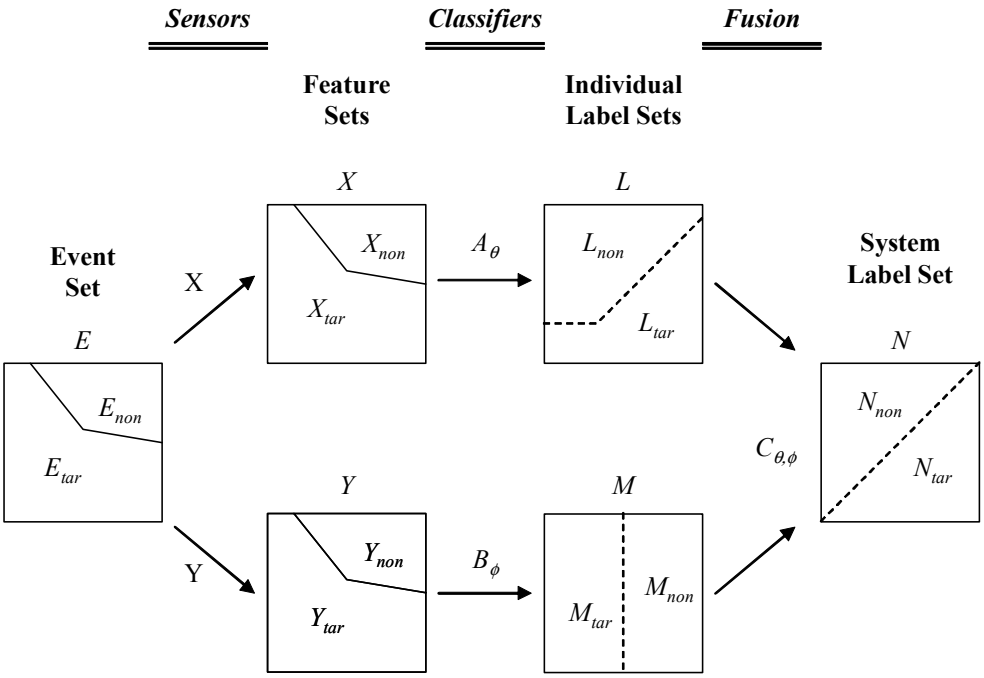

Figure 2.4. Within-MCS.

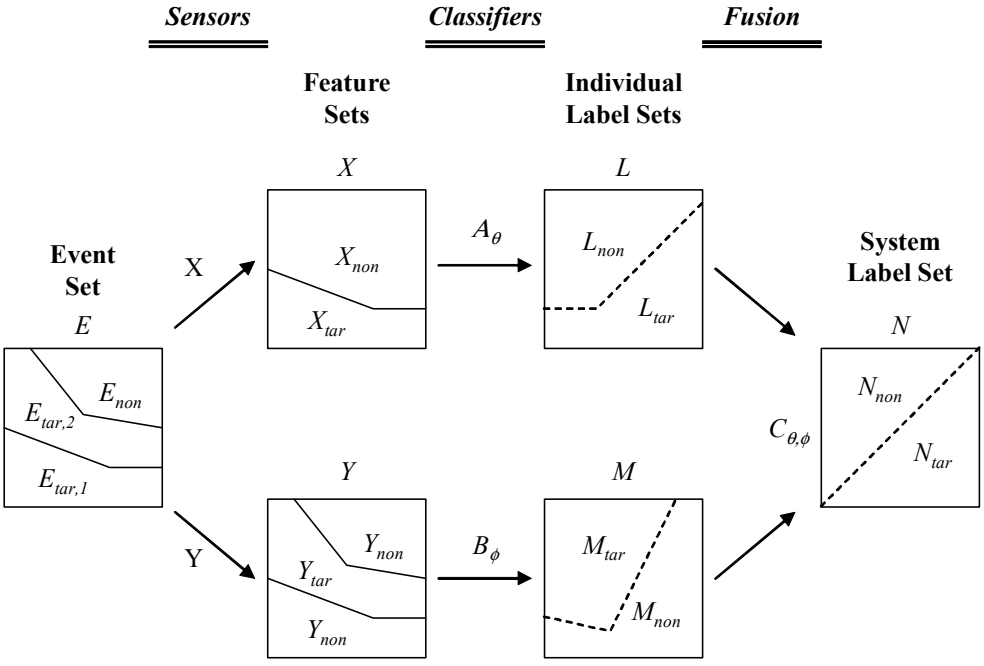

Figure 2.5. Across-MCS.

For both types of MCSs, a compact notation can be defined. Figure 2.6 shows the block diagram. Classifier A_θ maps feature set X to label set L and classifier B_ϕ maps feature set Y to label set M. More formally $A_\theta : X \to L$ for $\theta \in \Theta$, where Θ is the admissible threshold set for A_θ, and $B_\phi : Y \to M$ for $\phi \in \Phi$, where Φ is the admissible threshold set for B_ϕ. The fused classifier $C_{\theta,\phi}$ is the result of the fusion rule $\mathbf{r}$ acting on the labels or

$$C_{\theta,\phi}(x, y) = \mathbf{r}(A_\theta(x), B_\phi(y)),$$

where $(x, y) \in X \times Y$ and $(\theta, \phi) \in \Theta \times \Phi$. This can also be stated in terms of the label fusion as $\mathbf{r}$: $L \times M \to N$.

One final item on notation is presented. When referring to the families of classifiers for A_θ, B_ϕ, and $C_{\theta,\phi}$, the following notation is used

$$\begin{aligned} \mathbb{A} &= \{A_\theta : \theta \in \Theta\}, \\ \mathbb{B} &= \{B_\phi : \phi \in \Phi\}, \text{ and} \\ \mathbb{C} &= \{C_{\theta,\phi} : \theta \in \Theta, \phi \in \Phi\}. \end{aligned}$$

This is relevant when referring to the ROC curves for these classifiers, since the ROC curve represents the performance of the entire classifier family, not just the classifier for a given threshold.

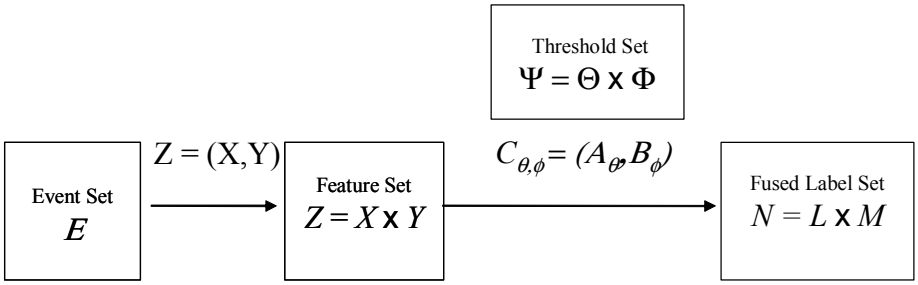

Figure 2.6. Compact notation for MCS.

2.4.3 Independence.

When analyzing multiple classifier systems (MCSs) it is often assumed that the classifiers are statistically independent [14]. More accurately, it is assumed that the measurements of the classifiers are independent. Two events, E_1 and E_2, are statistically independent if the following condition is met,

$$\Pr(E_1 \cap E_2) = \Pr(E_1) \cdot \Pr(E_2).$$

To extend this idea to the independence of classifiers, consider classifiers $A, B : X \to L$. Classifiers A and B are said to be independent if for every subset of labels $L_1, L_2 \subset L$,

$$\Pr(A^{-1}[L_1] \cap B^{-1}[L_2]) = \Pr(A^{-1}[L_1]) \cdot \Pr(B^{-1}[L_2])$$

where $A^{-1}[L_2]$ and $B^{-1}[L_2]$ are the inverse images of the classifiers. That is,

$$A^{-1}[L_1] = \{x \in X : A(x) \in L_1\} \text{ and } B^{-1}[L_2] = \{x \in X : B(x) \in L_2\}.$$

With this assumption, analysts can more easily compute the joint probabilities of an MCS by not having to consider the correlation between the classifiers. Any in-depth analysis of a specific MCS should examine whether this assumption is valid.

In some situations it is appropriate to assume that two classifiers are conditionally independent. Two events, E_1 and E_2, are said to be conditionally independent if they are independent with respect to the same condition G. In other words,

$$\Pr(E_1 \cap E_2|G) = \Pr(E_1|G) \cdot \Pr(E_2|G).$$

Similar to statistical independence this can be very helpful when calculating joint conditional probabilities, but once again, the validity of this assumption should be examined for any specific analysis.

Some researchers are less concerned with the classifiers being independent and more concerned with the classifiers making independent errors. For classifiers to make independent errors, they

must satisfy the following conditions,

$$P_{FP}(A^{-1}[L_{tar}] \cap B^{-1}[L_{tar}] | x \in X_{non}) = P_{FP}(A^{-1}[L_{tar}] | x \in X_{non}) \cdot P_{FP}(B^{-1}[L_{tar}] | x \in X_{non}), \text{ and}$$

$$P_{FN}(A^{-1}[L_{non}] \cap B^{-1}[L_{non}] | x \in X_{tar}) = P_{FN}(A^{-1}[L_{non}] | x \in X_{tar}) \cdot P_{FN}(B^{-1}[L_{non}] | x \in X_{tar}).$$

A number of researchers have used this assumption [9], [11]. Kuncheva contends however, that negatively correlated errors can actually improve the performance of an MCS [15].

2.4.4 ROC Fusion.

Oxley and Bauer demonstrated that one can analytically construct a ROC curve for an MCS simply by using the ROC curves of the individual classifiers [16]. They begin by providing the following definition for ROC curves. The ROC curve $f_{\mathbb{A}}$ is defined as

$$f_{\mathbb{A}}(p) = \max\{P_{TP}(A_\theta) : \theta \in \Theta \text{ and } P_{FP}(A_\theta) = p\} \tag{2.6}$$

for each $p \in [0, 1]$. In other words, if there are multiple θ values for which $P_{FP}(A_\theta) = p$, then the highest associated $P_{TP}(A_\theta)$ value is chosen to construct the ROC curve.

Similarly, the ROC curve $f_{\mathbb{B}}$ is defined as

$$f_{\mathbb{B}}(q) = \max\{P_{TP}(B_\phi) : \phi \in \Phi \text{ and } P_{FP}(B_\phi) = q\} \tag{2.7}$$

for each $q \in [0, 1]$.

Finally, the ROC curve $f_{\mathbb{C}}$ is defined to be

$$f_{\mathbb{C}}(r) = \max\{P_{TP}(C_{\theta,\phi}) : (\theta, \phi) \in \Theta \times \Phi \text{ and } P_{FP}(C_{\theta,\phi}) = r\} \tag{2.8}$$

for each $r \in [0, 1]$.

For the across-OR, Oxley and Bauer showed that

$$P_{FP}(C_{\theta,\phi}) = P_{FP}(A_\theta) + P_{FP}(B_\phi) - P_{FP}(A_\theta)P_{FP}(B_\phi), \tag{2.9}$$

and

$$\begin{aligned} P_{TP}(C_{\theta,\phi}) &= \frac{(1-\alpha)\beta}{\alpha+\beta-\alpha\beta}P_{FP}(A_\theta) + \frac{\alpha}{\alpha+\beta-\alpha\beta}P_{TP}(A_\theta) + \frac{(1-\alpha)\beta}{\alpha+\beta-\alpha\beta}P_{FP}(B_\phi) \\ &\quad + \frac{\beta}{\alpha+\beta-\alpha\beta}P_{TP}(B_\phi) - \frac{(1-\alpha)\beta}{\alpha+\beta-\alpha\beta}P_{FP}(A_\theta)P_{TP}(B_\phi) \\ &\quad - \frac{(1-\alpha)\beta}{\alpha+\beta-\alpha\beta}P_{TP}(A_\theta)P_{FP}(B_\phi) - \frac{\alpha\beta}{\alpha+\beta-\alpha\beta}P_{TP}(A_\theta)P_{TP}(B_\phi), \end{aligned}$$

where $\alpha = \Pr(X_{tar})$ and $\beta = \Pr(Y_{tar})$ are the *a priori* probabilities.

Considering Equation (2.9) and using the notation from the ROC curve definitions, it is seen that

$$r = p + q - pq$$

for a choice of p and q. For a given $r \in [0,1]$ then, p or q will be constrained such that if $q \in [0,1]$, then $p \in [0,r]$. In particular, for each r,

$$q = Q(p,r) = \left\{ \begin{array}{ccccc} \frac{r-p}{1-p} & \text{for} & 0 \le p < r & \text{when } r < 1 \\ 1 & \text{for} & p = r & \text{when } r = 1 \end{array} \right. .$$

Using this relationship and Equation (2.8), it was derived in [16] that the formula for the fused across-OR ROC curve is

$$\begin{aligned} f_{\mathbb{C}}(r) &= \frac{1}{\gamma} - \frac{(1-\gamma)r}{\gamma} - \\ &\quad \frac{1}{\gamma}\min_{0\le p\le r}\left\{[1-(\alpha f_{\mathbb{A}}(p)+(1-\alpha)p)]\cdot\left[1-\left(\beta f_{\mathbb{B}}\left(\frac{r-p}{1-p}\right)+(1-\beta)\left(\frac{r-p}{1-p}\right)\right)\right]\right\} \end{aligned} \tag{2.10}$$

where $\gamma = \alpha + \beta - \alpha\beta$.

In similar fashion, it was derived in [16] that the fused within-OR ROC curve is

$$f_{\mathbb{C}}(r) = \max_{0\le p\le r}\left\{f_{\mathbb{A}}(p) + f_{\mathbb{B}}\left(\frac{r-p}{1-p}\right) - f_{\mathbb{A}}(p)f_{\mathbb{B}}\left(\frac{r-p}{1-p}\right)\right\}. \tag{2.11}$$

Again, both of these formulas assume that the classifiers A_θ and B_ϕ are independent.

III. Derivation and Methodology

3.1 Introduction

This chapter extends Alsing's ROC convergence theorem by demonstrating that convergence is preserved when two empirical ROC curves are fused. Section 3.2 introduces the framework and notation that will be used in demonstrating this convergence. Section 3.3 provides the general result that convergence of a fused ROC curve is guaranteed when the ROC fusion is accomplished by a Lipschitz continuous transformation. Section 3.4 applies this result to the within-OR and across-OR fusion.

3.2 Framework and Notation

A classifier fusion rule, $\mathbf{r}$, establishes how classifiers A and B are combined into a fused classifier C, so that $C = \mathbf{r}(A, B)$. As Oxley and Bauer demonstrated, for certain fusion rules there may be a mapping or transformation, $\mathbf{T}$, that relates the fused ROC curve, $f_{\mathbb{C}}$, to the ROC curves of the individual classifier families, $f_{\mathbb{A}}$ and $f_{\mathbb{B}}$, so that $f_{\mathbb{C}} = f_{\mathbf{r}(\mathbb{A},\mathbb{B})} = \mathbf{T}(f_{\mathbb{A}}, f_{\mathbb{B}})$. Furthermore, this transformation may have a function g associated with it that allows for a pointwise evaluation of the fused ROC curve. Considering the individual ROC curves as functions of $p \in [0, 1]$, the following relations hold true

$$f_{\mathbb{C}}(p) = f_{\mathbf{r}(\mathbb{A},\mathbb{B})}(p) = \mathbf{T}(f_{\mathbb{A}}, f_{\mathbb{B}})(p) = g(f_{\mathbb{A}}(p), f_{\mathbb{B}}(p)).$$

Thus, the transformation $\mathbf{T}$ is a substitution transformation.

3.3 Convergence for Continuous Substitution Transformation

This section looks at a class of substitution transformations, $\mathbf{S}$, that satisfy the Lipschitz condition. If a particular fusion rule, $\mathbf{r}$, yields a Lipschitz continuous substitution transformation, $\mathbf{S}$, then convergence of the fused ROC curve for that rule is guaranteed by this result. A number of definitions and theorems will be required to establish the framework to prove this result.

The ROC curve f has been discussed in some detail to this point, but a few more remarks should be made before proceeding with this analysis. The ROC curve f has been referred to

primarily as a relation. Apostol defines a relation as any set of ordered pairs [3]. This is consistent with Equation (2.1). For this analysis, however, it will be more useful to think of f as a function, such that for $(x, y) \in f$, there exists a unique value $y = f(x)$ for a given x. With this in mind, the following ROC curve definition is stated.

Definition 3.1 (ROC Curve.) The function $f : [0, 1] \to [0, 1]$ is said to be a ROC curve if the domain of f is $[0, 1]$ and f is non-decreasing and upper semi-continuous. Let $\Re$ denote the collection of all ROC curves, that is

$$\Re = \{f : [0, 1] \to [0, 1] \mid f \text{ is non-decreasing and upper semi-continuous}\}.$$

For $f, g \in \Re$, the following mapping is introduced,

$$d_{\rho_1,1}(f, g) = \int_0^1 \rho_1(f(p), g(p))dp, \tag{3.1}$$

where ρ_1 is the standard metric on $\mathbb{R}$. Note this is the same mapping defined by Equation (2.4), but is generalized from θ-space to p-space. If $P_{FP}^{(f)}(\theta) = P_{FP}^{(g)}(\theta)$ for all $\theta \in \Theta$, then Equation (2.4) and Equation (3.1) are equivalent. This would seem to imply that Equation (3.1) is a special case of Equation (2.4), however it actually has more general applicability. Since the $d_{\rho_1,1}$ mapping from Equation (3.1) is defined over the projected ROC curves as a function of p, it does not require specific knowledge of the $\theta \in \Theta$ used to construct the ROC curves. This is very important when the threshold sets for $\mathbb{A}$ and $\mathbb{B}$ are not equal (i.e., $\Phi \neq \Theta$). The following theorem proves that the $d_{\rho_1,1}$ mapping is a metric on $\Re$.

Theorem 3.2 *$(\Re, d_{\rho_1,1})$ is a metric space.*

Proof. For $f_{\mathbb{A}}, f_{\mathbb{B}} \in \Re$, it must be shown that $d_{\rho_1,1}(f_{\mathbb{A}}, f_{\mathbb{B}})$ exists and satisfies the four required properties of a metric (Definition 2.1). Since $f_{\mathbb{A}}, f_{\mathbb{B}} \in \Re$, then $f_{\mathbb{A}} - f_{\mathbb{B}}$ has bounded variation. Hence $|f_{\mathbb{A}} - f_{\mathbb{B}}|$ has bounded variation and is therefore Riemann integrable. Thus,

$$d_{\rho_1,1}(f_{\mathbb{A}}, f_{\mathbb{B}}) = \int_0^1 |f_{\mathbb{A}}(p) - f_{\mathbb{B}}(p)|dp$$

exists for every $f_{\mathbb{A}}, f_{\mathbb{B}} \in \Re$.

1. *Nonnegativity.* Since $|f_{\mathbb{A}}(p) - f_{\mathbb{B}}(p)| \geq 0$. This implies $\int_0^1 \rho_1(f_{\mathbb{A}}(p), f_{\mathbb{B}}(p))dp \geq 0$. Therefore $d_{\rho_1,1}(f_{\mathbb{A}}, f_{\mathbb{B}}) \geq 0$ for all $f_{\mathbb{A}}$, $f_{\mathbb{B}}$ and so $d_{\rho_1,1}$ satisfies nonnegativity.

2. *Definiteness.* Assume $d_{\rho_1,1}(f_{\mathbb{A}}, f_{\mathbb{B}}) = 0$. Then $\int_0^1 |f_{\mathbb{A}}(p) - f_{\mathbb{B}}(p)|dp = 0$. This implies $|f_{\mathbb{A}}(p) - f_{\mathbb{B}}(p)| = 0$ for almost every $p \in [0,1]$, but $|f_{\mathbb{A}}(p) - f_{\mathbb{B}}(p)|$ of bounded variation on $[0,1]$ implies there are only a finite number of discontinuities. Since $f_{\mathbb{A}}, f_{\mathbb{B}}$ are upper semi-continuous, then $f_{\mathbb{A}} = f_{\mathbb{B}}$.

   Now assume $f_{\mathbb{A}} = f_{\mathbb{B}}$, then $|f_{\mathbb{A}}(p) - f_{\mathbb{B}}(p)| = 0$ for all $p \in [0,1]$ and implies $\int_0^1 |(f_{\mathbb{A}}(p) - f_{\mathbb{B}}(p)|dp = 0$. So, $d_{\rho_1,1}(f_{\mathbb{A}}, f_{\mathbb{B}}) = 0$. Therefore $d_{\rho_1,1}$ satisfies the definiteness property.

3. *Symmetry.* Since ρ_1 is a metric on $\mathbb{R}$, then $\rho_1(x,y) = \rho_1(y,x)$ for all $x, y \in \mathbb{R}$. This implies $\rho_1(f_{\mathbb{A}}(p), f_{\mathbb{B}}(p)) = \rho_1(f_{\mathbb{B}}(p), f_{\mathbb{A}}(p))$ for all $p = [0,1]$; hence,

$$d_{\rho_1,1}(f_{\mathbb{A}}, f_{\mathbb{B}}) = \int_0^1 \rho_1(f_{\mathbb{A}}(p), f_{\mathbb{B}}(p))dp = \int_0^1 \rho_1(f_{\mathbb{B}}(p), f_{\mathbb{A}}(p))dp = d_{\rho_1,1}(f_{\mathbb{B}}, f_{\mathbb{A}}).$$

   Therefore $d_{\rho_1,1}$ satisfies the symmetry property.

4. *Triangle inequality.* Let $f_{\mathbb{A}}, f_{\mathbb{B}}, f_{\mathbb{D}} \in \Re$. Since ρ_1 is a metric on $\mathbb{R}$, then $\rho_1(x,y) \leq \rho_1(x,z) + \rho_1(z,y)$ for all $x, y, z \in \mathbb{R}$. Therefore $\rho_1(f_{\mathbb{A}}(p), f_{\mathbb{B}}(p)) \leq \rho_1(f_{\mathbb{A}}(p), f_{\mathbb{D}}(p)) + \rho_1(f_{\mathbb{D}}(p), f_{\mathbb{B}}(p))$ for all $p = [0,1]$; hence,

$$\begin{aligned} d_{\rho_1,1}(f_{\mathbb{A}}, f_{\mathbb{B}}) &= \int_0^1 \rho_1(f_{\mathbb{A}}(p), f_{\mathbb{B}}(p))dp \\ &\leq \int_0^1 \rho_1(f_{\mathbb{A}}(p), f_{\mathbb{D}}(p))dp + \int_0^1 \rho_1(f_{\mathbb{D}}(p), f_{\mathbb{B}}(p))dp \\ &= d_{\rho_1,1}(f_{\mathbb{A}}, f_{\mathbb{D}}) + d_{\rho_1,1}(f_{\mathbb{D}}, f_{\mathbb{B}}). \end{aligned}$$

   So $d_{\rho_1,1}$ satisfies the triangle inequality property.

Definition 3.3 (Average metric distance.) For $f_{\mathbb{A}}, f_{\mathbb{B}} \in \Re$, the mapping

$$d_{\rho_1,1}(f_{\mathbb{A}}, f_{\mathbb{B}}) = \int_0^1 \rho_1(f_{\mathbb{A}}(p), f_{\mathbb{B}}(p))dp \tag{3.2}$$

is defined to be the average metric distance.

Note that when this metric was defined over θ (See Equation (2.4).), $d_{\rho_1,1}(f_{\mathbb{A}}, f_{\mathbb{B}})$ was referred to as the total metric distance. Since the measure of the interval $[0,1]$ is 1, then for Equation (3.2) the total metric distance and average metric distance are equivalent, and the $d_{\rho_1,1}(f_{\mathbb{A}}, f_{\mathbb{B}})$ notation will be used to refer to both.

Two metrics in $\mathbb{R}^2$ will also be used. The first is the Manhattan metric on a 2-element vector. For $S = [0,1]^2 = [0,1] \times [0,1] = \{\mathbf{x} = (\xi, \eta) \in \mathbb{R}^2 \mid 0 \leq \xi \leq 1, 0 \leq \eta \leq 1\}$, the following metric is defined

$$\rho_1^{(2)}(\mathbf{x}_1, \mathbf{x}_2) = |\xi_1 - \xi_2| + |\eta_1 - \eta_2|.$$

The metric space $\left([0,1]^2, \rho_1^{(2)}\right)$ is commonly used and is stated without proof. Similarly for $S = \Re^2 = \Re \times \Re$, the following metric is defined

$$d^{(2)}_{\rho_1,1}[(f_{\mathbb{A}}, f_{\mathbb{B}}), (f_{\mathbb{A}'}, f_{\mathbb{B}'})] = d_{\rho_1,1}(f_{\mathbb{A}}, f_{\mathbb{A}'}) + d_{\rho_1,1}(f_{\mathbb{B}}, f_{\mathbb{B}'}).$$

The proof that $(\Re^2, d^{(2)}_{\rho_1,1})$ is a metric space is a simple extension of Theorem 3.2. These metrics will be used in conjunction with the following definitions.

Definition 3.4 (Lipschitz continuous mapping.) Let (S, d_S) and (T, d_T) be metric spaces. The mapping $\sigma : S \to T$ is said to be a Lipschitz continuous mapping if there exists some $L > 0$ such that

$$d_T(\sigma(s_1), \sigma(s_2)) \leq L d_S(s_1, s_2) \text{ for all } s_1, s_2 \in S.$$

$\mathcal{L}ip(S, T)$ is used to denote the collection of all Lipschitz continuous mappings from (S, d_S) into (T, d_T).

Definition 3.5 (ROC fusion function.) The function, $g : [0,1]^2 \to [0,1]$ is said to be a ROC fusion function if the domain of g is $[0,1] \times [0,1]$ and g is non-decreasing in both variables. Let $\mathcal{G}$ denote the collection of all ROC fusion functions, that is

$$\mathcal{G} = \{g : [0,1]^2 \to [0,1] \mid g \text{ is non-decreasing in both variables}\}.$$

Notice that if there exists an $L > 0$ such that

$$\rho_1(g(\mathbf{x}_1), g(\mathbf{x}_2)) \leq L\rho_1^{(2)}(\mathbf{x}_1, \mathbf{x}_2)$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in [0,1]^2$, then $g : [0,1]^2 \to [0,1]$ is a Lipschitz continuous mapping from $\left([0,1]^2, \rho_1^{(2)}\right)$ into $([0,1], \rho_1)$, or $g \in \mathcal{L}ip([0,1]^2, [0,1])$.

Definition 3.6 (Substitution transformation.) For $g : [0,1]^2 \to [0,1]$, the substitution transformation, $\mathbb{S}_g : \Re^2 \to \Re$, is defined as

$$\mathbb{S}_g(f_{\mathbb{A}}, f_{\mathbb{B}})(p) = g(f_{\mathbb{A}}(p), f_{\mathbb{B}}(p)) \text{ for all } p \in [0,1].$$

Notice that if there exists an $L > 0$ such that

$$d_{\rho_1,1}[\mathbb{S}_g(f_{\mathbb{A}}, f_{\mathbb{B}}), \mathbb{S}_g(f_{\mathbb{A}'}, f_{\mathbb{B}'})] \leq L d^{(2)}_{\rho_1,1}[(f_{\mathbb{A}}, f_{\mathbb{B}}), (f_{\mathbb{A}'}, f_{\mathbb{B}'})]$$

for all $(f_{\mathbb{A}}, f_{\mathbb{B}}), (f_{\mathbb{A}'}, f_{\mathbb{B}'}) \in \Re^2$, then $\mathbb{S}_g : \Re^2 \to \Re$ is a Lipschitz continuous mapping from $\left(\Re^2, d^{(2)}_{\rho_1,1}\right)$ into $\left(\Re, d_{\rho_1,1}\right)$, or $\mathbb{S}_g \in \mathcal{L}ip(\Re^2, \Re)$.

With this framework developed, the main result can now be proven. The approach will be to demonstrate that if a ROC fusion function is Lipschitz continuous, then the associated substitution transformation is also Lipschitz continuous. When this Lipschitz substitution transformation is used to fuse two individual ROC curves that converge in probability, it is proven that the fused ROC curve will converge in probability as well.

Theorem 3.7 *Let $g \in \mathcal{L}ip([0,1]^2,[0,1]) \cap \mathcal{G}$, then $\mathbb{S}_g \in \mathcal{L}ip(\Re^2,\Re)$.*

Proof. Let $(f_{\mathbb{A}}, f_{\mathbb{B}}), (f_{\mathbb{A}'}, f_{\mathbb{B}'}) \in \Re^2$ and

$$d_{\rho_1,1}[\mathbb{S}_g(f_{\mathbb{A}}, f_{\mathbb{B}}), \mathbb{S}_g(f_{\mathbb{A}'}, f_{\mathbb{B}'})] = \int_0^1 |\mathbb{S}_g(f_{\mathbb{A}}, f_{\mathbb{B}})(p) - \mathbb{S}_g(f_{\mathbb{A}'}, f_{\mathbb{B}'})(p)|dp.$$

Show that the Lipschitz condition is met.

$$\begin{aligned}
d_{\rho_1,1}[\mathbb{S}_g(f_{\mathbb{A}}, f_{\mathbb{B}}), \mathbb{S}_g(f_{\mathbb{A}'}, f_{\mathbb{B}'})] &= \int_0^1 |\mathbb{S}_g(f_{\mathbb{A}}, f_{\mathbb{B}})(p) - \mathbb{S}_g(f_{\mathbb{A}'}, f_{\mathbb{B}'})(p)|dp \\
&= \int_0^1 |g(f_{\mathbb{A}}(p), f_{\mathbb{B}}(p)) - g(f_{\mathbb{A}'}(p), f_{\mathbb{B}'}(p))|dp \\
&\leq \int_0^1 L(|f_{\mathbb{A}}(p) - f_{\mathbb{A}'}(p)| + |f_{\mathbb{B}}(p) - f_{\mathbb{B}'}(p)|)dp \\
&= L\int_0^1 |f_{\mathbb{A}}(p) - f_{\mathbb{A}'}(p)|dp + L\int_0^1 |f_{\mathbb{B}}(p) - f_{\mathbb{B}'}(p)|dp \\
&= L[d_{\rho_1,1}(f_{\mathbb{A}}, f_{\mathbb{A}'}) + d_{\rho_1,1}(f_{\mathbb{B}}, f_{\mathbb{B}'})] \\
&= L d^{(2)}_{\rho_1,1}[(f_{\mathbb{A}}, f_{\mathbb{B}}), (f_{\mathbb{A}'}, f_{\mathbb{B}'})]
\end{aligned}$$

So as desired, the Lipschitz condition on $\mathbb{S}_g$ is met

$$d_{\rho_1,1}[\mathbb{S}_g(f_{\mathbb{A}}, f_{\mathbb{B}}), \mathbb{S}_g(f_{\mathbb{A}'}, f_{\mathbb{B}'})] \leq L d^{(2)}_{\rho_1,1}[(f_{\mathbb{A}}, f_{\mathbb{B}}), (f_{\mathbb{A}'}, f_{\mathbb{B}'})],$$

and $\mathbb{S}_g \in \mathcal{L}ip(\Re^2,\Re)$.

Theorem 3.8 *Assume $\mathbb{T} : \left(\Re^2, d^{(2)}_{\rho_1,1}\right) \to \left(\Re, d_{\rho_1,1}\right)$ is a Lipschitz continuous mapping. Assume $\{\hat{f}^{(n)}_{\mathbb{A}}\}, \{\hat{f}^{(m)}_{\mathbb{B}}\} \subset \Re$ converge in probability to $f_{\mathbb{A}}$ and $f_{\mathbb{B}}$, respectively, then $\hat{f}^{(n,m)}_{\mathbb{C}} = \mathbb{T}(\hat{f}^{(n)}_{\mathbb{A}}, \hat{f}^{(m)}_{\mathbb{B}})$ converges in probability to $f_{\mathbb{C}} = \mathbb{T}(f_{\mathbb{A}}, f_{\mathbb{B}})$.*

Proof. Let $L > 0$ be the Lipschitz constant for $\mathbb{T}$. For $\varepsilon > 0$, there exists an N_ε such that for all $n, m > N_\varepsilon$,

$$Pr\left\{\omega_{\mathbb{A}} \in \Omega_{\mathbb{A}} : d_{\rho_1,1}(\hat{f}_{\mathbb{A}}^{(n)}(\omega_{\mathbb{A}}), f_{\mathbb{A}}) \geq \frac{\varepsilon}{2L}\right\} < \frac{\varepsilon}{2}$$

and

$$Pr\left\{\omega_{\mathbb{B}} \in \Omega_{\mathbb{B}} : d_{\rho_1,1}(\hat{f}_{\mathbb{B}}^{(m)}(\omega_{\mathbb{B}}), f_{\mathbb{B}}) \geq \frac{\varepsilon}{2L}\right\} < \frac{\varepsilon}{2}.$$

To prove $\hat{f}_{\mathbb{C}}^{(n,m)}$converges in probability to $f_{\mathbb{C}}$, the following must be shown

$$Pr\left\{(\omega_{\mathbb{A}}, \omega_{\mathbb{B}}) \in \Omega_{\mathbb{A}} \times \Omega_{\mathbb{B}} : d_{\rho_1,1}(\hat{f}_{\mathbb{C}}^{(m,n)}(\omega_{\mathbb{A}}, \omega_{\mathbb{B}}), f_{\mathbb{C}}) \geq \varepsilon\right\} < \varepsilon.$$

Consider the set

$$\left\{(\omega_{\mathbb{A}}, \omega_{\mathbb{B}}) \in \Omega_{\mathbb{A}} \times \Omega_{\mathbb{B}} : d_{\rho_1,1}(\hat{f}_{\mathbb{C}}^{(m,n)}(\omega_{\mathbb{A}}, \omega_{\mathbb{B}}), f_{\mathbb{C}}) \geq \varepsilon\right\}$$

and the relationship from Theorem 3.7 yields

$$d_{\rho_1,1}(\hat{f}_{\mathbb{C}}^{(m.n)}(\omega_{\mathbb{C}}), f_{\mathbb{C}}) \leq L d_{\rho_1,1}(\hat{f}_{\mathbb{A}}^{(n)}(\omega_{\mathbb{A}}), f_{\mathbb{A}}) + L d_{\rho_1,1}(\hat{f}_{\mathbb{B}}^{(m)}(\omega_{\mathbb{B}}), f_{\mathbb{B}}).$$

It can be seen that

$$\begin{aligned}
&\left\{(\omega_{\mathbb{A}}, \omega_{\mathbb{B}}) \in \Omega_{\mathbb{A}} \times \Omega_{\mathbb{B}} : d_{\rho_1,1}(\hat{f}_{\mathbb{C}}^{(m,n)}(\omega_{\mathbb{A}}, \omega_{\mathbb{B}}), f_{\mathbb{C}}) \geq \varepsilon\right\} \\
\subset\ &\left\{(\omega_{\mathbb{A}}, \omega_{\mathbb{B}}) \in \Omega_{\mathbb{A}} \times \Omega_{\mathbb{B}} : L d_{\rho_1,1}(\hat{f}_{\mathbb{A}}^{(n)}(\omega_{\mathbb{A}}), f_{\mathbb{A}}) + L d_{\rho_1,1}(\hat{f}_{\mathbb{B}}^{(m)}(\omega_{\mathbb{B}}), f_{\mathbb{B}}) \geq \varepsilon\right\} \\
\subset\ &\left\{(\omega_{\mathbb{A}}, \omega_{\mathbb{B}}) \in \Omega_{\mathbb{A}} \times \Omega_{\mathbb{B}} : d_{\rho_1,1}(\hat{f}_{\mathbb{A}}^{(n)}(\omega_{\mathbb{A}}), f_{\mathbb{A}}) \geq \frac{\varepsilon}{2L}\right\} \\
&\cup \left\{(\omega_{\mathbb{A}}, \omega_{\mathbb{B}}) \in \Omega_{\mathbb{A}} \times \Omega_{\mathbb{B}} : d_{\rho_1,1}(\hat{f}_{\mathbb{B}}^{(m)}(\omega_{\mathbb{B}}), f_{\mathbb{B}}) \geq \frac{\varepsilon}{2L}\right\}.
\end{aligned}$$

Now considering the probability measure of this event, gives

$$\begin{aligned}
&Pr\left\{(\omega_{\mathbb{A}}, \omega_{\mathbb{B}}) \in \Omega_{\mathbb{A}} \times \Omega_{\mathbb{B}} : d_{\rho_1,1}(\hat{f}_{\mathbb{C}}^{(m,n)}(\omega_{\mathbb{A}}, \omega_{\mathbb{B}}), f_{\mathbb{C}}) \geq \varepsilon\right\} \\
\leq\ &Pr\left\{(\omega_{\mathbb{A}}, \omega_{\mathbb{B}}) \in \Omega_{\mathbb{A}} \times \Omega_{\mathbb{B}} : d_{\rho_1,1}(\hat{f}_{\mathbb{A}}^{(n)}(\omega_{\mathbb{A}}), f_{\mathbb{A}}) \geq \frac{\varepsilon}{2L}\right\} \\
&+Pr\left\{(\omega_{\mathbb{A}}, \omega_{\mathbb{B}}) \in \Omega_{\mathbb{A}} \times \Omega_{\mathbb{B}} : d_{\rho_1,1}(\hat{f}_{\mathbb{B}}^{(m)}(\omega_{\mathbb{B}}), f_{\mathbb{B}}) \geq \frac{\varepsilon}{2L}\right\}.
\end{aligned}$$

By the independence assumption then,

$$Pr\left\{(\omega_{\mathbb{A}},\omega_{\mathbb{B}})\in\Omega_{\mathbb{A}}\times\Omega_{\mathbb{B}}:d_{\rho_1,1}(\hat{f}_{\mathbb{A}}^{(n)}(\omega_{\mathbb{A}}),f_{\mathbb{A}})\geq\frac{\varepsilon}{2L}\right\}=Pr\left\{\omega_{\mathbb{A}}\in\Omega_{\mathbb{A}}:d_{\rho_1,1}(\hat{f}_{\mathbb{A}}^{(n)}(\omega_{\mathbb{A}}),f_{\mathbb{A}})\geq\frac{\varepsilon}{2L}\right\}$$

and

$$Pr\left\{(\omega_{\mathbb{A}},\omega_{\mathbb{B}})\in\Omega_{\mathbb{A}}\times\Omega_{\mathbb{B}}:d_{\rho_1,1}(\hat{f}_{\mathbb{B}}^{(m)}(\omega_{\mathbb{B}}),f_{\mathbb{B}})\geq\frac{\varepsilon}{2L}\right\}=Pr\left\{\omega_{\mathbb{B}}\in\Omega_{\mathbb{B}}:d_{\rho_1,1}(\hat{f}_{\mathbb{B}}^{(m)}(\omega_{\mathbb{B}}),f_{\mathbb{B}})\geq\frac{\varepsilon}{2L}\right\}.$$

Therefore,

$$\begin{aligned}
& Pr\left\{(\omega_{\mathbb{A}},\omega_{\mathbb{B}})\in\Omega_{\mathbb{A}}\times\Omega_{\mathbb{B}}:d_{\rho_1,1}(\hat{f}_{\mathbb{C}}^{(m,n)}(\omega_{\mathbb{A}},\omega_{\mathbb{B}}),f_{\mathbb{C}})\geq\varepsilon\right\}\\
\leq\ & Pr\left\{\omega_{\mathbb{A}}\in\Omega_{\mathbb{A}}:d_{\rho_1,1}(\hat{f}_{\mathbb{A}}^{(n)}(\omega_{\mathbb{A}}),f_{\mathbb{A}})\geq\frac{\varepsilon}{2L}\right\}+Pr\left\{\omega_{\mathbb{B}}\in\Omega_{\mathbb{B}}:d_{\rho_1,1}(\hat{f}_{\mathbb{B}}^{(m)}(\omega_{\mathbb{B}}),f_{\mathbb{B}})\geq\frac{\varepsilon}{2L}\right\}\\
<\ & \frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon
\end{aligned}$$

as desired, and $\hat{f}_{\mathbb{C}}^{(n,m)}$converges in probability to $f_{\mathbb{C}}$.

This is an important result that covers a large class of mappings. Just as important is the framework that was constructed. This framework provides a valuable tool for the comparison of ROC curves and classifiers.

3.4 Convergence Framework Application

Since the transformation for the within-OR ROC fusion (Equation 2.11) contains a maximization and the transformation for the across-OR ROC fusion (Equation 2.10) contains a minimization, it will be difficult to show that these are Lipschitz continuous mappings.

Consider first a simpler case, the ROC fusion function $g(\zeta,\eta)=\zeta+\eta-\zeta\eta$ so that

$$f_{\mathbb{C}}(p)=g(f_{\mathbb{A}}(p),f_{\mathbb{B}}(p))=f_{\mathbb{A}}(p)+f_{\mathbb{B}}(p)-f_{\mathbb{A}}(p)f_{\mathbb{B}}(p).$$

Since $f_{\mathbb{A}}$ and $f_{\mathbb{B}}$ are, by definition, non-decreasing functions of p, then for $f_{\mathbb{C}}$ evaluated at r

$$f_{\mathbb{C}}(r)=f_{\mathbb{A}}(r)+f_{\mathbb{B}}(r)-f_{\mathbb{A}}(r)f_{\mathbb{B}}(r)\geq\max_{0\leq p\leq r}\left\{f_{\mathbb{A}}(p)+f_{\mathbb{B}}\left(\frac{r-p}{1-p}\right)-f_{\mathbb{A}}(p)f_{\mathbb{B}}\left(\frac{r-p}{1-p}\right)\right\}.$$

Therefore, $f_{\mathbb{C}}$ is actually an upper bound for the fused OR curve. This is consistent with Clutz's research. Clutz showed for the within-OR fusion case, that a point on the fused ROC curve is given

by

$$(r, f_{\mathbb{C}}(r)) = (p + q - pq, f_{\mathbb{A}}(p) + f_{\mathbb{B}}(q) - f_{\mathbb{A}}(p) f_{\mathbb{B}}(q)),$$

assuming that the classifiers are independent in their measurements [5]. Clutz states this is a weak upper bound for the fused ROC curve.

It can be shown that $f_{\mathbb{C}}$ is a Lipschitz continuous mapping. The strategy will be to demonstrate that for some $L > 0$,

$$d_{\rho_1,1}(f_{\mathbb{C}}, f_{\mathbb{C}'}) \leq L[d_{\rho_1,1}(f_{\mathbb{A}}, f_{\mathbb{A}'}) + d_{\rho_1,1}(f_{\mathbb{B}}, f_{\mathbb{B}'})].$$

Consider the following,

$$\begin{aligned}
& d_{\rho_1,1}(f_{\mathbb{C}}, f_{\mathbb{C}'}) \\
= & \int_0^1 |f_{\mathbb{C}}(p) - f_{\mathbb{C}'}(p)|\, dp \\
= & \int_0^1 |f_{\mathbb{A}}(p) + f_{\mathbb{B}}(p) - f_{\mathbb{A}}(p) f_{\mathbb{B}}(p) - [f_{\mathbb{A}'}(p) + f_{\mathbb{B}'}(p) - f_{\mathbb{A}'}(p) f_{\mathbb{B}'}(p)]\, | dp \\
= & \int_0^1 |f_{\mathbb{A}}(p) - f_{\mathbb{A}}(p) + f_{\mathbb{B}}(p) - f_{\mathbb{B}'}(p) + f_{\mathbb{A}'}(p) f_{\mathbb{B}'}(p) - f_{\mathbb{A}}(p) f_{\mathbb{B}}(p)| dp \\
= & \int_0^1 |f_{\mathbb{A}}(p) - f_{\mathbb{A}'}(p) + f_{\mathbb{B}}(p) - f_{\mathbb{B}'}(p) + f_{\mathbb{A}'}(p) f_{\mathbb{B}'}(p) - f_{\mathbb{A}}(p) f_{\mathbb{B}}(p) + f_{\mathbb{A}}(p) f_{\mathbb{B}'}(p) - f_{\mathbb{A}}(p) f_{\mathbb{B}'}(p)| dp \\
= & \int_0^1 |f_{\mathbb{A}}(p) - f_{\mathbb{A}'}(p) + f_{\mathbb{B}'}(p)\, [f_{\mathbb{A}'}(p) - f_{\mathbb{A}}(p)] + f_{\mathbb{B}}(p) - f_{\mathbb{B}'}(p) + f_{\mathbb{A}}(p)\, [f_{\mathbb{B}'}(p) - f_{\mathbb{B}}(p)]\, | dp \\
= & \int_0^1 |\, [1 - f_{\mathbb{B}'}(p)] \cdot [f_{\mathbb{A}}(p) - f_{\mathbb{A}'}(p)] + [1 - f_{\mathbb{A}}(p)] \cdot [f_{\mathbb{B}}(p) - f_{\mathbb{B}'}(p)]\, | dp \\
\leq & \int_0^1 |\, \max [1 - f_{\mathbb{B}'}(p)] \cdot [f_{\mathbb{A}}(p) - f_{\mathbb{A}'}(p)] + \max [1 - f_{\mathbb{A}}(p)] \cdot [f_{\mathbb{B}}(p) - f_{\mathbb{B}'}(p)]\, | dp \\
= & \int_0^1 1 \cdot |(f_{\mathbb{A}}(p) - f_{\mathbb{A}'}(p)| dp + \int_0^1 1 \cdot |f_{\mathbb{B}}(p) - f_{\mathbb{B}'}(p)|) dp
\end{aligned}$$

So,

$$d_{\rho_1,1}(f_{\mathbb{C}}, f_{\mathbb{C}'}) \leq d_{\rho_1,1}(f_{\mathbb{A}}, f_{\mathbb{A}'}) + d_{\rho_1,1}(f_{\mathbb{B}}, f_{\mathbb{B}'})$$

with $L = 1$, and as desired the transformation is Lipschitz continuous. Therefore, by Theorem 3.8, if $\hat{f}_{\mathbb{A}}^{(n)}$ converges in probability to $f_{\mathbb{A}}$ and $\hat{f}_{\mathbb{B}}^{(n)}$ converges in probability to $f_{\mathbb{B}}$, then $\hat{f}_{\mathbb{C}}^{(n,m)}$converges in probability to $f_{\mathbb{C}}$.

This example demonstrates how the convergence framework can be applied to a ROC fusion mapping to show it is Lipschitz continuous and therefore convergent. This framework can be applied to experimental data as well. The conjecture that within-OR and across-OR fusion are Lipschitz continuous will be examined experimentally in the next chapter.

IV. Analysis and Findings

4.1 Overview

This chapter provides an experimental analysis of the convergence of a sequence of empirical ROC curves. As the test sample size n is increased, the convergence is measured using the numerical approximation of the average metric distance. The average metric distances for the single-classifier ROC curves as well as the fused-classifier ROC curves are calculated. The relationship between the two will be examined.

In addition to varying sample size, the number of interpolation points, m, used to approximate the ROC curve is also varied. In application, interpolation is used to construct a continuous ROC curve from discrete sample data. As more interpolation points are used, the better this curve approximates the actual empirical ROC curve. So just as increasing n should make $\hat{f}^{(n)}$ a better estimate of f, increasing m should make the interpolated curve a better approximate of the actual curve. This approximation is sometimes overlooked when doing ROC analysis.

The empirical ROC curves in this experiment are constructed both parametrically and non-parametrically. In the non-parametric case, the frequency of false positives and true positives is measured at discrete threshold values, and the resultant ordered pairs are used to construct the ROC curve. Put more simply, the number of false positives and true positives will be counted at each n for each threshold. In the parametric case, the underlying statistical distribution of the data is assumed, or in the case of this experiment, known. The target and non-target sample mean and standard deviation will be calculated for each n, and the ROC curve will be constructed by evaluating the cumulative distribution function at each threshold. This process is repeated for five trials.

4.2 Experimental Design

4.2.1 Classifiers.

This experiment examines the OR rule applied to two classifiers for both the within and across fusion cases. Each classifier is a simple two-label classifier, declaring either target or non-target. Classifier A_θ is defined on the feature set $X = \mathbb{R}$ for each $\theta \in \Theta = \mathbb{R}$ as:

$$A_\theta(x) = \begin{cases} tar & \text{if} \quad x < \theta \\ non & \text{if} \quad x \geq \theta \end{cases}$$

Classifier B_ϕ is defined on the feature set $Y = \mathbb{R}$ for each $\phi \in \Phi = (0, \infty)$ as follows:

$$B_\phi(x) = \begin{cases} tar & \text{if} \quad -1 - \phi < x < -1 + \phi \\ non & \text{if} \quad \text{otherwise} \end{cases}$$

4.2.2 Feature Data.

Normally distributed feature data for both the target and non-target classes is generated using the MATLAB **normrnd** function. One thousand samples are generated for each class. The first n samples from each class are selected to constitute the sample sets $X_{tar}^{(n)}$ and $X_{non}^{(n)}$, with $X^{(n)} = X_{tar}^{(n)} \cup X_{non}^{(n)}$. This is done for $n = 10, 100, 200, 400$, and 1000. By choosing the sets in this manner, it ensures they are nested, that is,

$$X^{(10)} \subset X^{(100)} \subset X^{(200)} \subset X^{(400)} \subset X^{(1000)}.$$

Although this nesting is not strictly necessary in a statistical sense, this is consistent with Alsing's work [1].

4.2.3 ROC Estimates.

Classifiers A_θ and B_ϕ are applied to each $X^{(n)}$ to generate $\hat{P}_{TP}^{(n)}$ and $\hat{P}_{FP}^{(n)}$ at each threshold, θ_i and ϕ_j, respectively. For the non-parametric case, the number of true positives and false positives is counted at each threshold and divided by n. For classifier A_θ for each $\theta \in \Theta$,

$$\hat{P}_{TP}^{(n)}(\theta) = \frac{\#\text{TP}}{n} \text{ and } \hat{P}_{FP}^{(n)}(\theta) = \frac{\#\text{FP}}{n}.$$

More formally this can be expressed as

$$\hat{P}_{TP}^{(n)}(\theta) = \frac{card\{x_i < \theta | x_i \in X_{tar}^{(n)}\}}{n} \text{ for } i = 1, ..., n,$$

$$\hat{P}_{FP}^{(n)}(\theta) = \frac{card\{x_i < \theta | x_i \in X_{non}^{(n)}\}}{n} \text{ for } i = 1, ..., n.$$

where *card* is the cardinality of the set. For classifier B_ϕ for each $\phi \in \Phi$,

$$\hat{P}_{TP}^{(n)}(\phi) = \frac{\#\text{TP}}{n} \text{ and } \hat{P}_{FP}^{(n)}(\phi) = \frac{\#\text{FP}}{n}, \text{ for given } \phi.$$

More formally this can be expressed as

$$\hat{P}_{TP}^{(n)}(\phi) = \frac{card\{-1 - \phi < x_i < -1 + \phi | x_i \in X_{tar}^{(n)}\}}{n} \text{ for } i = 1, ..., n$$

$$\hat{P}_{FP}^{(n)}(\phi) = \frac{card\{-1 - \phi < x_i < -1 + \phi | x_i \in X_{non}^{(n)}\}}{n} \text{ for } i = 1, ..., n$$

For the parametric case, the target sample mean $\widehat{\mu}_{tar}^{(n)}$ and standard deviation $\widehat{\sigma}_{tar}^{(n)}$ are calculated for each $X^{(n)}$ as well as the non-target sample mean $\widehat{\mu}_{non}^{(n)}$ and standard deviation $\widehat{\sigma}_{non}^{(n)}$. These are used in the evaluation of the MATLAB **normcdf** function at each threshold. For classifier A_θ for each θ,

$$\begin{aligned} \hat{P}_{TP}^{(n)}(\theta) &= \mathbf{normcdf}(\theta, \widehat{\mu}_{tar}^{(n)}, \widehat{\sigma}_{tar}^{(n)}), \\ \hat{P}_{FP}^{(n)}(\theta) &= \mathbf{normcdf}(\theta, \widehat{\mu}_{non}^{(n)}, \widehat{\sigma}_{non}^{(n)}). \end{aligned}$$

For classifier B_ϕ for each ϕ,

$$\begin{aligned} \hat{P}_{TP}^{(n)}(\phi) &= \mathbf{normcdf}(-1 + \phi, \widehat{\mu}_{tar}^{(n)}, \widehat{\sigma}_{tar}^{(n)}) - \mathbf{normcdf}(-1 - \phi, \widehat{\mu}_{tar}^{(n)}, \widehat{\sigma}_{tar}^{(n)}), \\ \hat{P}_{FP}^{(n)}(\phi) &= \mathbf{normcdf}(-1 + \phi, \widehat{\mu}_{non}^{(n)}, \widehat{\sigma}_{non}^{(n)}) - \mathbf{normcdf}(-1 - \phi, \widehat{\mu}_{non}^{(n)}, \widehat{\sigma}_{non}^{(n)}). \end{aligned}$$

Finally, the true probabilities $P_{FP}(\theta)$ and $P_{TP}(\theta)$ are calculated for each θ and $P_{FP}(\phi)$ and $P_{TP}(\phi)$ are calculated for each ϕ. Similar to the parametric case, these are calculated with the **normcdf** function, this time using the true target and non-target means and standard deviations.

4.2.4 Empirical ROC Curve Construction.

Using the estimates from Section 4.2.2, ROC curves for each classifier for each n can be constructed. Since there are a finite number of threshold values and sample feature data, it is possible that the ROC trajectory for multiple threshold values could project to the same $\hat{P}_{FP}$ value. In this case, the ROC curve is constructed using the highest $\hat{P}_{TP}$ value associated with that $\hat{P}_{FP}$ value. Recall this conforms to Oxley and Bauer's definition of a ROC curve (See equations (2.6) and (2.7).). The empirical ROC curve for classifier family $\mathbb{A}$ then is defined to be

$$\hat{f}_{\mathbb{A}}^{(n)}(p) = \max\{\hat{P}_{TP}^{(n)}(A_\theta) : \theta \in \Theta \text{ and } \hat{P}_{FP}^{(n)}(A_\theta) = p\}.$$

Similarly, the empirical ROC curve for classifier family $\mathbb{B}$ is defined to be

$$\hat{f}_{\mathbb{B}}^{(n)}(q) = \max\{\hat{P}_{TP}^{(n)}(B_\phi) : \phi \in \Phi \text{ and } \hat{P}_{FP}^{(n)}(B_\phi) = q\}.$$

Finally, the empirical fused-OR ROC curve for classifier family $\mathbb{C}$ is defined to be

$$\hat{f}_{\mathbb{C}}^{(n)}(r) = \max\{\hat{P}_{TP}^{(n)}(C_{\theta,\phi}) : (\theta,\phi) \in \Theta \times \Phi \text{ and } \hat{P}_{FP}^{(n)}(C_{\theta,\phi}) = r\},$$

and is calculated using Oxley and Bauer's ROC fusion formulas (See equations (2.10) and (2.11).).

The $(\hat{P}_{FP}, \hat{P}_{TP})$ values are interpolated with the MATLAB **interp1** function. The MATLAB **interp1** function is a linear interpolation. To measure the effect of the number of interpolation points on the average metric distance, $m = 11, 31, 101, 501$ interpolates are used.

4.2.5 Test Metric.

Recall from Definition 3.3 that average metric distance can be defined over p versus θ. The numerical approximation for average metric distance then is

$$d_{\rho_1,1}(\hat{f}^{(n)}, f) \approx \frac{\sum_{i=1}^{m} \rho_1\left(\hat{f}^{(n)}(p), f(p)\right)}{m}. \tag{4.1}$$

For the purposes of this experiment, the $d_{\rho_1,1}(\hat{f}^{(n)}, f)$ metric is always calculated with $m = 501$. This is done in order to ensure that the average metric distances calculated for varying m can be consistently compared to each other. If this is not done, an $\hat{f}^{(n)}$ approximated with $m = 11$

interpolates would only be summed over 11 points in the average metric distance calculation, while an $\hat{f}^{(n)}$ approximated with $m = 31$ interpolates would be summed over 31 points. To be consistent, an $\hat{f}^{(n)}$ constructed with 11, 31, or 101 interpolates will always be evaluated at 501 points along the curve in the average metric distance calculation. Likewise, the true ROC curve f is constructed with $m = 501$ interpolates. All estimated ROC curves, regardless of the m used to approximate them, are compared to this true ROC curve. The average metric distance for each ROC curve, $d_{\rho_1,1}(\hat{f}_{\mathbb{A}}^{(n)}, f_{\mathbb{A}})$, $d_{\rho_1,1}(\hat{f}_{\mathbb{B}}^{(n)}, f_{\mathbb{B}})$, and $d_{\rho_1,1}(\hat{f}_{\mathbb{C}}^{(n)}, f_{\mathbb{C}})$ is calculated for all n and m.

4.2.6 Within Fusion.

In the within fusion case, classifiers A_θ and B_ϕ act on the same feature set. Target feature data is drawn from a normal distribution with $\mu = -1$ and $\sigma = 1/\sqrt{2}$. Non-target feature data is drawn from a normal distribution with $\mu = 1$ and $\sigma = 1$. Recall the definition of classifier A_θ:

$$A_\theta(x) = \begin{cases} tar & \text{if} \quad x < \theta \\ non & \text{if} \quad x \geq \theta \end{cases}.$$

Figure 4.1 depicts this classifier. The threshold values for A_θ are $\theta \in \{-4, -3.9, -3.8, ..., 4\}$.

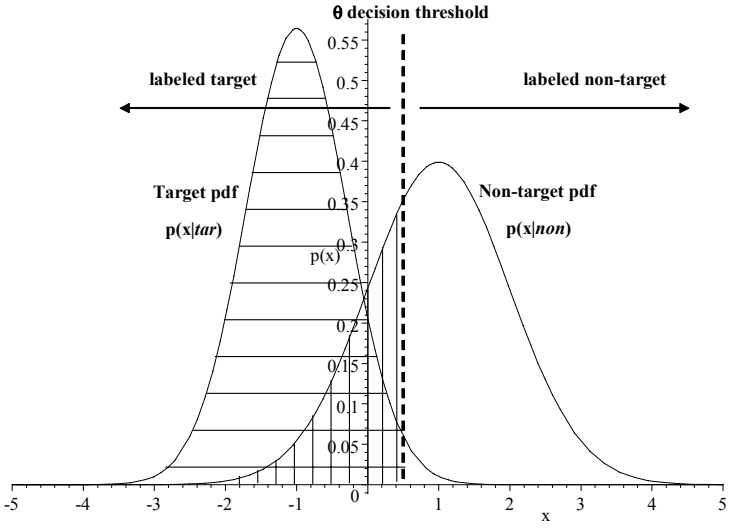

Figure 4.1. Classifier A_θ within fusion case.

Recall classifier B_ϕ:

$$B_\phi(x) = \begin{cases} tar & \text{if} \quad -1-\phi < x < -1+\phi \\ non & \text{if} \qquad \text{otherwise} \end{cases} .$$

Figure 4.2 depicts this classifier. The threshold values for B_ϕ are $\phi \in \{0, 0.05, 0.10, ..., 5\}$. The

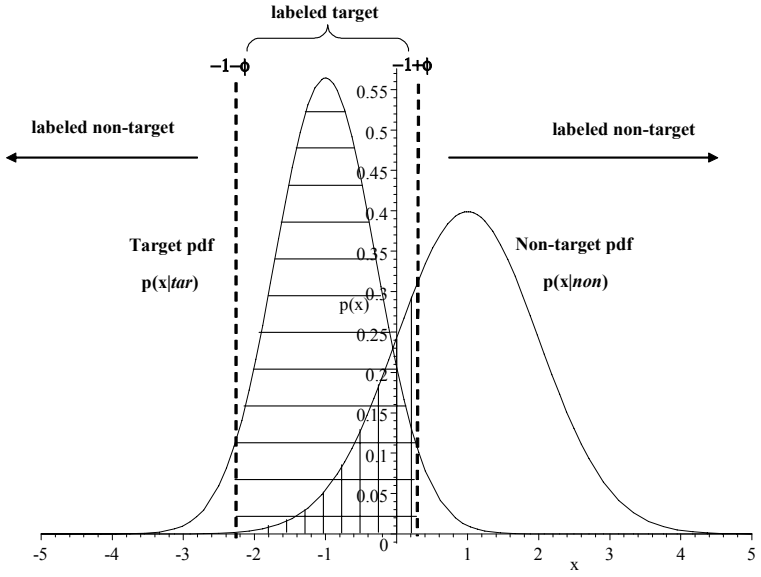

Figure 4.2. Classifier B_ϕ within fusion case.

ROC fusion for $A_\theta \vee B_\phi$ is done numerically using Oxley and Bauer's MATLAB code for within-OR fusion. Both the parametric and non-parametric experiments use the same feature data and threshold values.

4.2.7 Across Fusion.

In across fusion, classifiers A_θ and B_ϕ act on different sets of feature data. For this experiment, classifier A_θ will use the same target and non-target feature data sets from the within case. Classifier B_ϕ acts on new feature set $Y = \mathbb{R}$, with target data drawn from a normal distribution with $\mu = -1$ and $\sigma = 1$ and non-target feature data drawn from a normal distribution with $\mu = 1$ and $\sigma = 2$. Figure 4.3 depicts this. Threshold sets are the same as the within fusion case, specifically $\theta \in \{-4, -3.9, -3.8, ..., 4\}$ and $\phi \in \{0, 0.05, 0.10, ..., 5\}$.

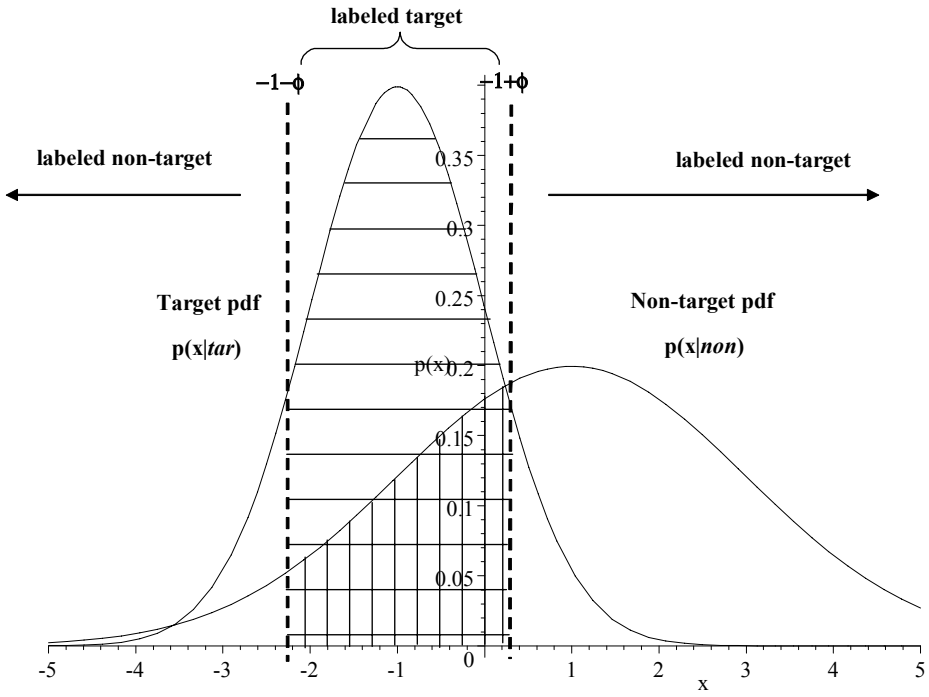

Figure 4.3. Classifier B_ϕ across fusion case.

4.3 Results

4.3.1 Within-OR (Non-parametric).

Tables 4.1, 4.2, and 4.3 provide the results for the $d_{\rho_1,1}(\hat{f}_{\mathbb{A}}^{(n)}, f_{\mathbb{A}})$, $d_{\rho_1,1}(\hat{f}_{\mathbb{B}}^{(n)}, f_{\mathbb{B}})$, and $d_{\rho_1,1}(\hat{f}_{\mathbb{C}}^{(n)}, f_{\mathbb{C}})$ metrics. The tables show the averages over 5 trials. Looking down the columns, as expected, the average metric distance decreases as n increases, with the exception being the $m = 11$ column. This column can be subject to additional variability due to the potentially poor approximation with such a small number of interpolates. The trend down the other columns is fairly stable. Looking across the rows, next, there is a general downward trend in values from the $m = 11$ column to $m = 31$ column. The average metric distance values seem to stabilize for $m = 31, 101, 501$ though. This could be explained by the fact classifiers A_θ and B_ϕ only evaluate at 81 and 101 threshold values, respectively. So the rule of thumb from this example would appear to be to use a number of interpolates similar to the number of threshold values. Anything less may add variability to the results; anything more may yield diminishing returns. Finally, it is interesting to note that the average metric distance for $f_{\mathbb{C}}$ is on the order of that for $f_{\mathbb{A}}$ and $f_{\mathbb{B}}$. In fact, it is even slightly lower.

Figure 4.4 displays the results graphically. The plots displayed are for the $m = 31$ case. The true ROC curve and empirical ROC curve are plotted for $\mathbb{A}$, $\mathbb{B}$, and $\mathbb{C}$. While there is certainly some difference visually at $n = 10$; as quickly as $n = 100$, the empirical ROC curves fall right on top of their respective true ROC curves. For $n = 200, 400, 1000$, they are barely distinguishable.

To put these results into perspective and gain an intuitive feel for these values, consider first that the numerical approximation for the average metric distance (See Equation (4.1).) is very close to a Midpoint quadrature rule for numerical integration. So the average metric distance can also be thought of as the area of the difference between $\hat{f}^{(n)}$ and f. Now consider that an average metric distance of 1 is the maximum possible difference between two ROC curves and a value of 0 means the two ROC curves are equivalent. Furthermore, the average metric distance between the chance line and the ideal ROC curve is 0.5. So in the $n = 1000$ case, for all three ROC curves, there is

approximately a 0.5% difference in area. Even at $n = 100$, the difference is less than 2%, suggesting that one could construct fairly tight confidence intervals around the curve for relatively small sample sizes.

Table 4.1. Within-OR (non-parametric): average metric distances for $\hat{f}_{\mathbb{A}}^{(n)}$.

$n \diagdown m$	11	31	101	501
10	0.0461	0.0461	0.0461	0.0461
100	0.0201	0.0179	0.0180	0.0180
200	0.0219	0.0122	0.0120	0.0120
400	0.0231	0.0095	0.0080	0.0082
1000	0.0200	0.0057	0.0042	0.0044

Table 4.2. Within-OR (non-parametric): average metric distances for $\hat{f}_{\mathbb{B}}^{(n)}$.

$n \diagdown m$	11	31	101	501
10	0.0604	0.0604	0.0604	0.0604
100	0.0147	0.0173	0.0180	0.0180
200	0.0144	0.0169	0.0171	0.0170
400	0.0153	0.0121	0.0124	0.0126
1000	0.0111	0.0064	0.0066	0.0067

Table 4.3. Within-OR (non-parametric): average metric distances for $\hat{f}_{\mathbb{C}}^{(n)}$.

$n \diagdown m$	11	31	101	501
10	0.0353	0.0353	0.0353	0.0353
100	0.0158	0.0156	0.0165	0.0165
200	0.0182	0.0107	0.0115	0.0115
400	0.0229	0.0086	0.0070	0.0071
1000	0.0197	0.0046	0.0035	0.0036

4.3.2 Across-OR (Non-parametric).

Tables 4.4, 4.5, and 4.6 provide the results for the $d_{\rho_1,1}(\hat{f}_{\mathbb{A}}^{(n)}, f_{\mathbb{A}})$, $d_{\rho_1,1}(\hat{f}_{\mathbb{B}}^{(n)}, f_{\mathbb{B}})$, and $d_{\rho_1,1}(\hat{f}_{\mathbb{C}}^{(n)}, f_{\mathbb{C}})$ metric. The across-OR experiment yields generally similar results to the within-OR case. The average metric distance values are on par with the within case, and the trends are similar with respect to increasing n and m. Notice one interesting difference in Table 4.5 for the $m = 11$ column though. The approximation at $m = 11$ is better than the approximation at $m = 31, 101, 501$. So the rule discussed previously does not appear to be hard and fast, although it still probably is a

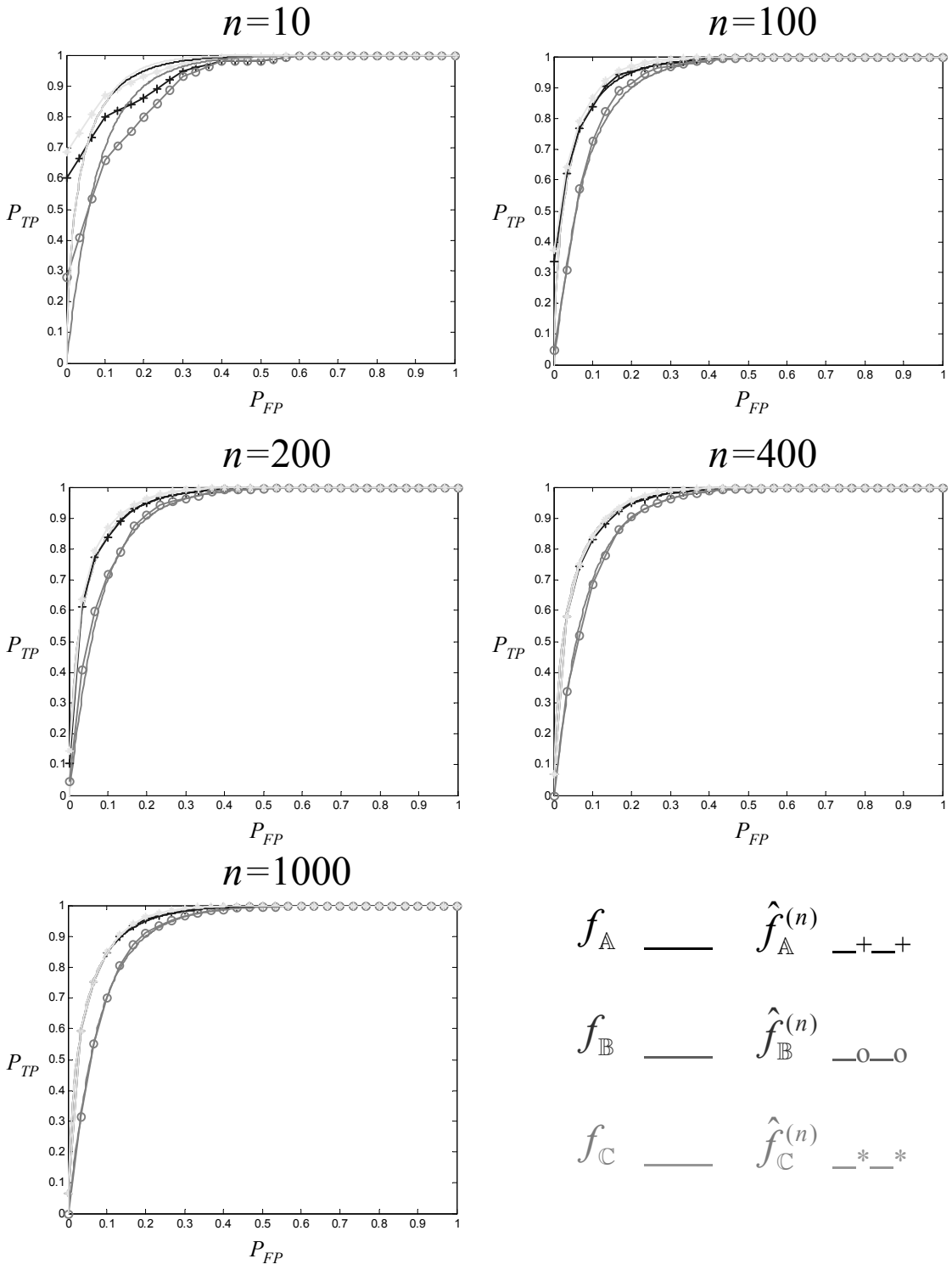

Figure 4.4. Average metric distances for the within-OR (non-parametric case) for varying n.

best practice to choose an optimal m based on the number of θ's. Another difference is that the average metric distance for $f_{\mathbb{C}}$ is not smaller than that for $f_{\mathbb{A}}$ and $f_{\mathbb{B}}$, however, it is still on the same order. Figure 4.5 displays this graphically. Once again the convergence visually appears to be very fast.

Table 4.4. Across-OR (non-parametric): average metric distances for $\hat{f}_{\mathbb{A}}^{(n)}$.

$n \diagdown m$	11	31	101	501
10	0.0461	0.0461	0.0461	0.0461
100	0.0201	0.0179	0.0180	0.0180
200	0.0219	0.0122	0.0120	0.0120
400	0.0231	0.0095	0.0080	0.0082
1000	0.0200	0.0057	0.0042	0.0044

Table 4.5. Across-OR (non-parametric): average metric distances for $\hat{f}_{\mathbb{B}}^{(n)}$.

$n \diagdown m$	11	31	101	501
10	0.1226	0.1226	0.1226	0.1226
100	0.0220	0.0272	0.0282	0.0282
200	0.0162	0.0167	0.0174	0.0175
400	0.0174	0.0165	0.0164	0.0166
1000	0.0087	0.0093	0.0092	0.0093

Table 4.6. Across-OR (non-parametric): average metric distances for $\hat{f}_{\mathbb{C}}^{(n)}$.

$n \diagdown m$	11	31	101	501
10	0.0660	0.0665	0.0669	0.0671
100	0.0168	0.0194	0.0207	0.0208
200	0.0193	0.0140	0.0144	0.0146
400	0.0210	0.0120	0.0110	0.0113
1000	0.0162	0.0062	0.0054	0.0055

4.3.3 Within-OR (Parametric).

Tables 4.7, 4.8, and 4.9 provide comparisons of the $d_{\rho_1,1}(\hat{f}_{\mathbb{A}}^{(n)}, f_{\mathbb{A}})$, $d_{\rho_1,1}(\hat{f}_{\mathbb{B}}^{(n)}, f_{\mathbb{B}})$, and $d_{\rho_1,1}(\hat{f}_{\mathbb{C}}^{(n)}, f_{\mathbb{C}})$ metrics for the parametric and non-parametric cases. The tables show the averages over 5 trials. For this case only $m = 31$ interpolates was calculated. The convergence results for the parametric case are comparable to the results for the non-parametric case. Figure 4.6 displays this graphically. Aesthetically the parametric ROC curves look much better, but from this data one

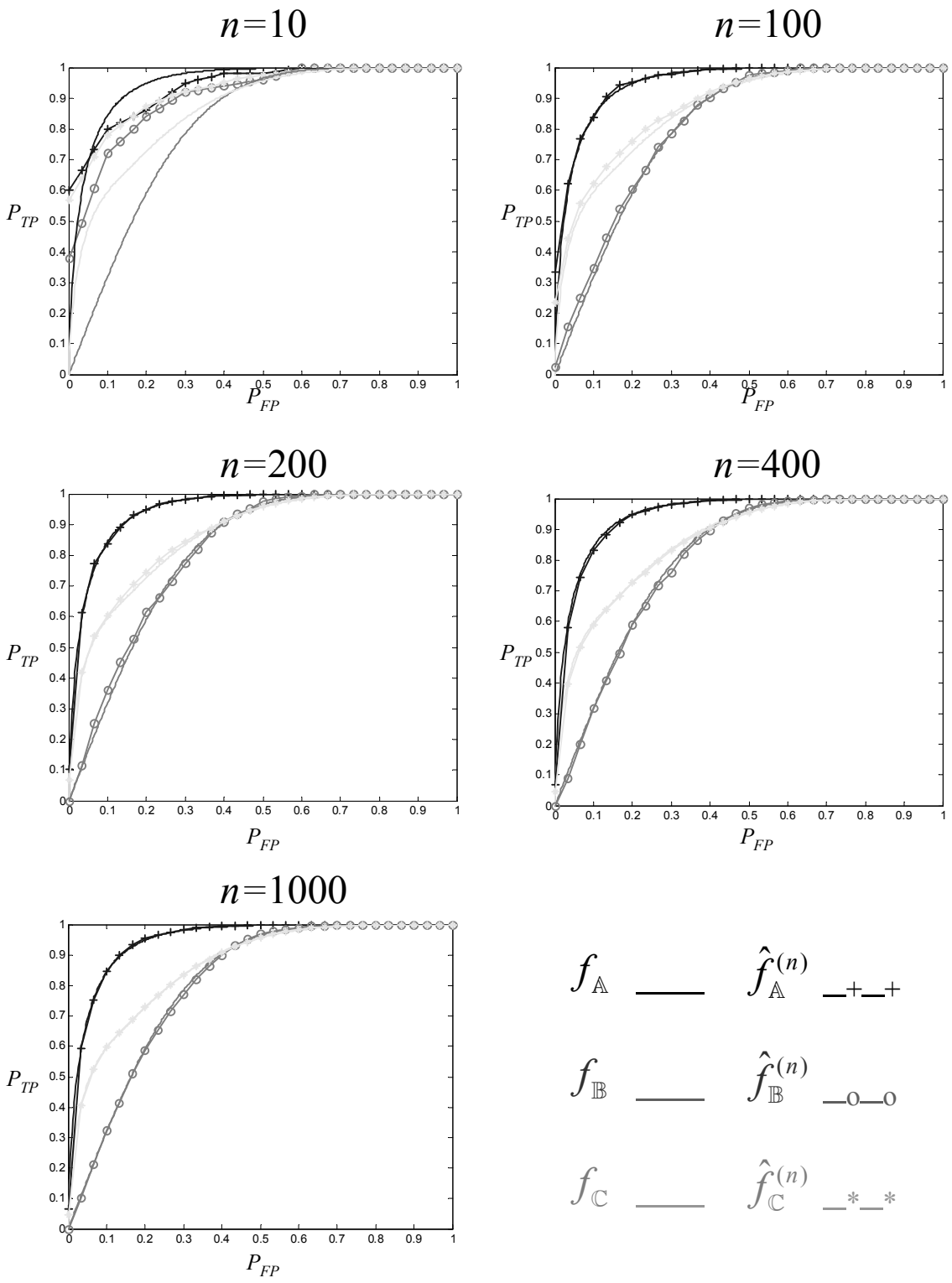

Figure 4.5. Average metric distances for the across-OR (non-parametric case) for varying n.

cannot conclude there is a performance advantage over the non-parametric case. The parametric method is certainly easier to compute, so that may be a consideration for an application where computing power is at a premium.

Table 4.7. Parametric vs. non-parametric: average metric distances for $\hat{f}_{\mathbb{A}}^{(n)}$.

	parametric	non-parametric
$n \diagdown m$	31	31
10	0.0483	0.0461
100	0.0129	0.0179
200	0.0099	0.0122
400	0.0080	0.0095
1000	0.0061	0.0057

Table 4.8. Parametric vs. non-parametric: average metric distances for $\hat{f}_{\mathbb{B}}^{(n)}$.

	parametric	non-parametric
$n \diagdown m$	31	31
10	0.0483	0.0604
100	0.0113	0.0173
200	0.0098	0.0169
400	0.0070	0.0121
1000	0.0046	0.0064

Table 4.9. Parametric vs. non-parametric: average metric distances for $\hat{f}_{\mathbb{C}}^{(n)}$.

	parametric	non-parametric
$n \diagdown m$	31	31
10	0.0426	0.0353
100	0.0114	0.0156
200	0.0095	0.0107
400	0.0072	0.0086
1000	0.0060	0.0046

4.3.4 Convergence as a Function of Sample Size.

Figures 4.7, 4.8, and 4.9 display the convergence as a function of sample size for the within-OR fusion (non-parametric and parametric) and the across-OR fusion cases. The convergence rate appears to be of the order $\frac{1}{\sqrt{n}}$. This is consistent with Tchebysheff's Theorem [23]. Applying

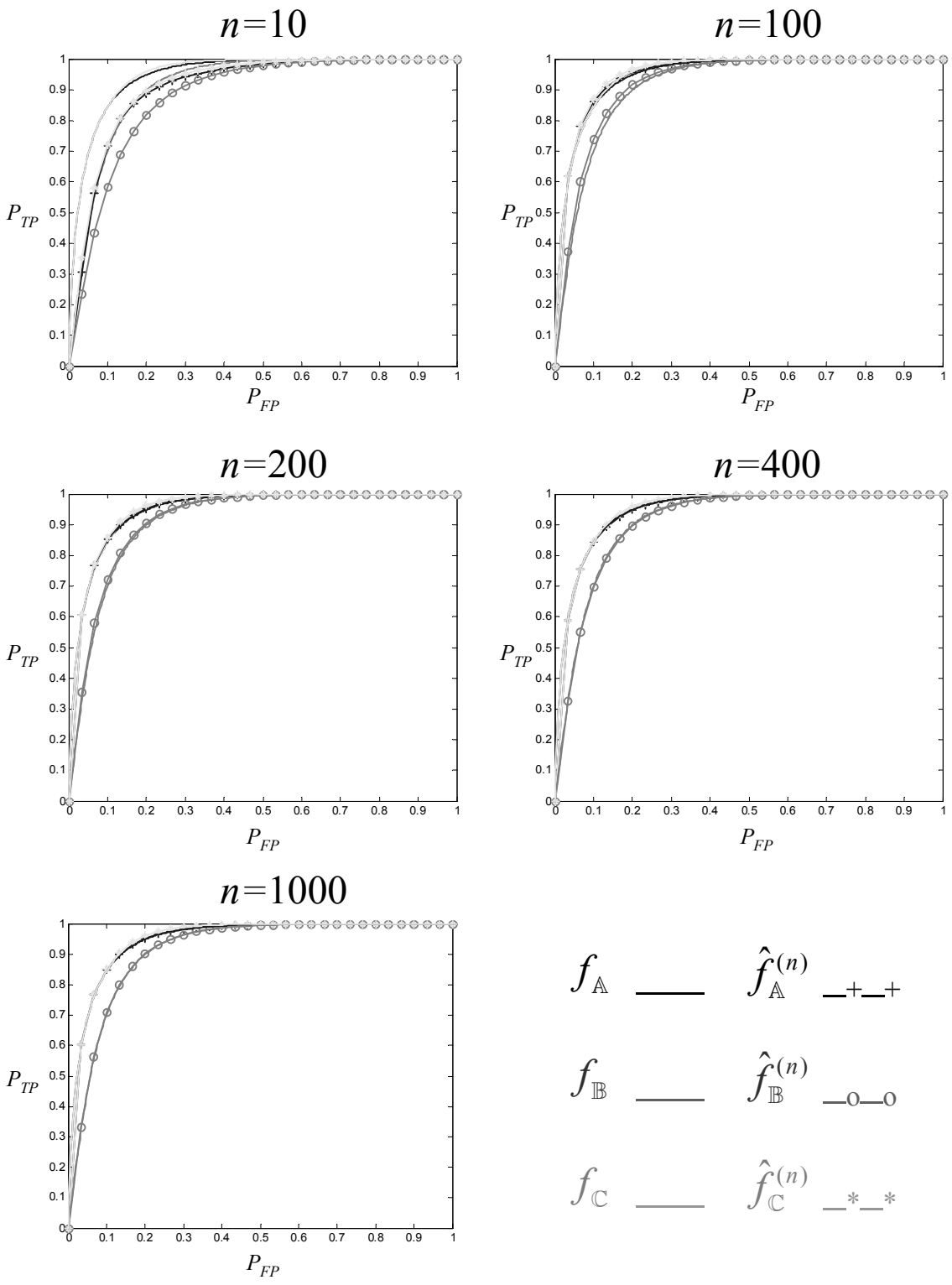

Figure 4.6. Average metric distances for the within-OR (parametric case) for varying n.

Tchebysheff's Theorem to this analysis, for positive constant k at given p, that is

$$\Pr\left(|\hat{f}^{(n)}(p) - f(p)| \geq k\frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{k^2}$$

where σ is the standard deviation of the random variable $\hat{f}^{(n)}(p)$. So Tchebysheff's Theorem says at a given confidence level, $\hat{f}^{(n)}(p)$ should become a better estimate of $f(p)$ as a function of $\frac{1}{\sqrt{n}}$, consistent with the findings of the experiment. The figures also provide more visual evidence that the average metric distances for the individual ROC curves and the fused ROC curves are of the same order. This is a good result, since this implies there is no penalty to the rate of convergence for fused ROC curves. Finally, observe that this experiment supports the conjecture that the fused empirical ROC curve converges for within-OR and across-OR fusion. Furthermore, for this experiment, the observed Lipschitz constant is less than one, and in fact is very close to 0.5.

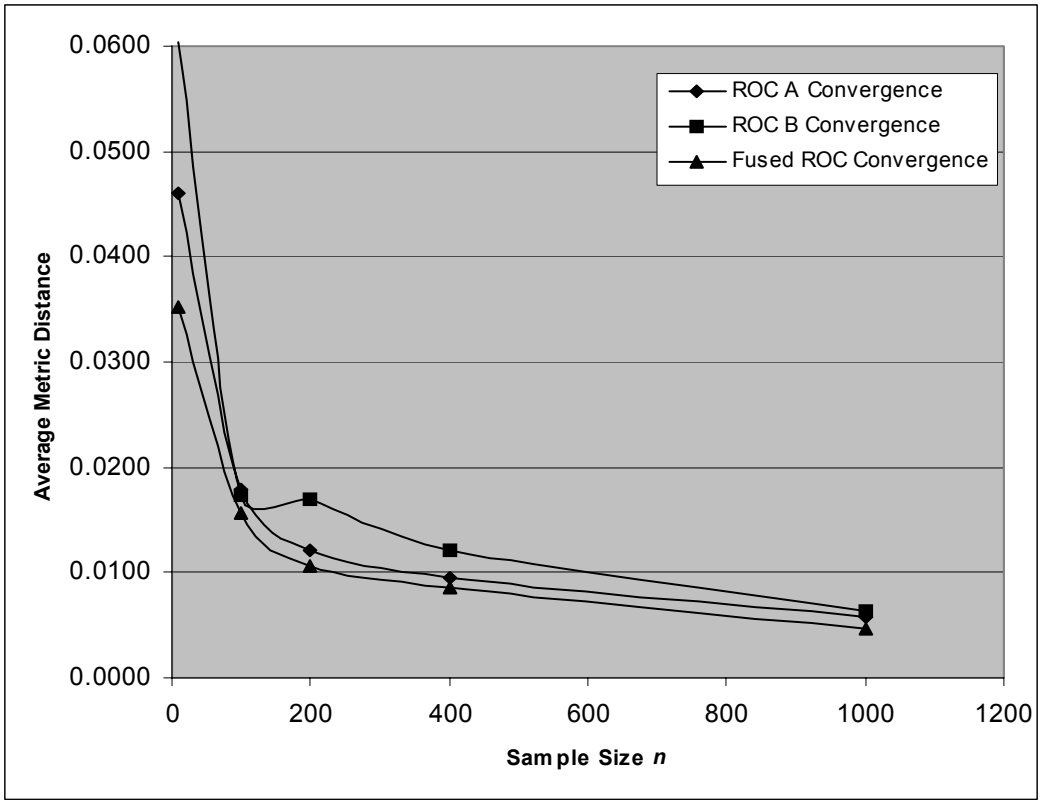

Figure 4.7. Average metric distance as a function of sample size - within-OR (non-parametric case)

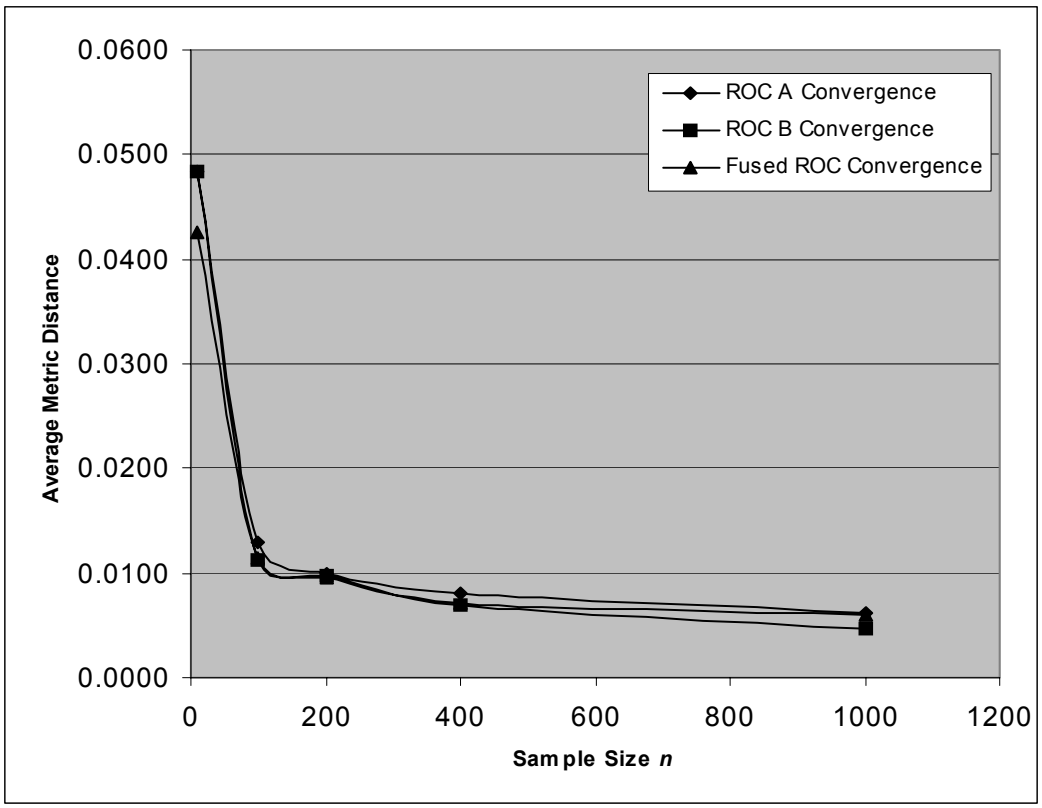

Figure 4.8. Average metric distance as a function of sample size - within-OR (parametric case)

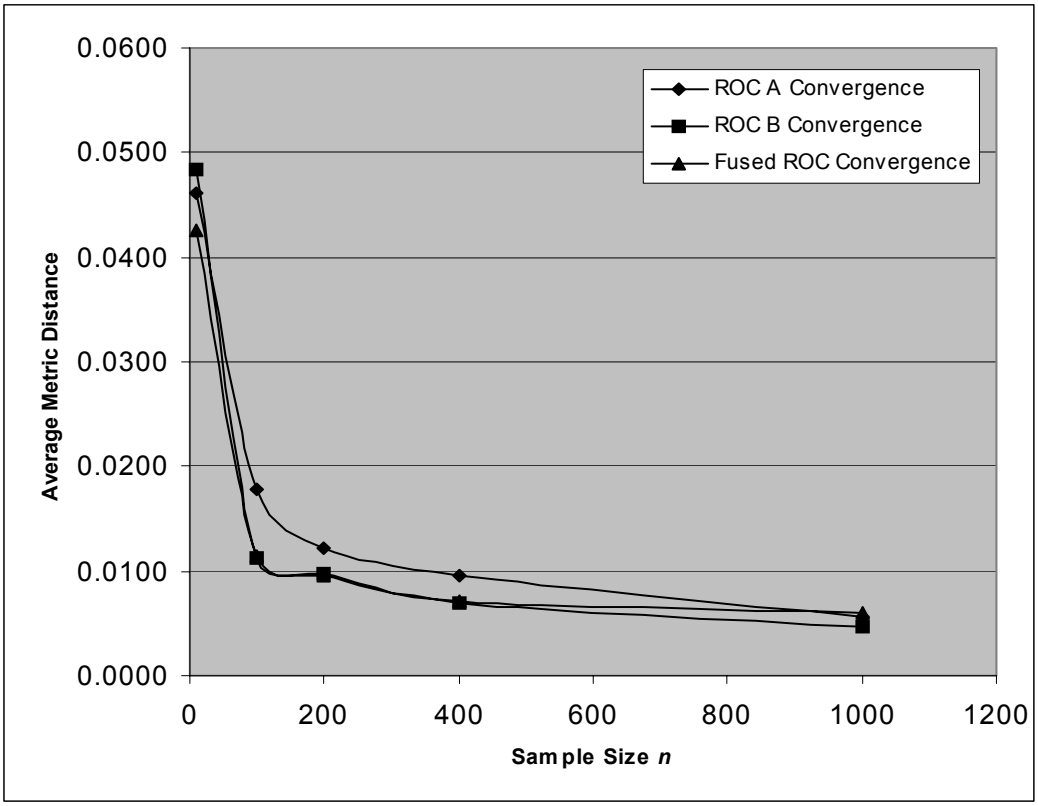

Figure 4.9. Average metric distance as a function of sample size - across-OR (non-parametric case)

V. Summary and Recommendations

5.1 Summary of Contributions

The primary contribution of this thesis is a proof of convergence for fused empirical ROC curves. It proves that if the individual empirical ROC curves converge and the ROC fusion is accomplished with a Lipschitz continuous transformation, a large class of transformations, that the convergence of the fused ROC curve is guaranteed. This is an important contribution since it establishes that fused empirical ROC curves are consistent estimators of the true fused ROC curve and therefore comparisons with these curves are valid. This thesis also provides a mathematical framework to prove this convergence. This framework is applied to a generic ROC fusion transformation, as well as the OR rule fusion transformation.

The experiment using the OR rule provides valuable insight into how this framework can be applied. The results of the experiment indicate that the convergence of the fused ROC curve is on the same order as the individual ROC curves, a nice result. The utility of this framework also extends beyond establishing convergence and could be used for comparing any two ROC curves. Finally, the interpolation used to construct the empirical ROC curve should be a consideration when doing ROC analysis.

5.2 Recommendations for Future Research

There are several candidates for continuing research in this area. There are potentially many other ROC fusion transformations to which this framework could be applied. In particular, the convergence under AND fusion, for which a formula is available, could be proven with these results. This could also be extended to the majority vote, and potentially other fusers, such as artificial neural nets. Although it was conjectured, and supported experimentally, that OR fusion is Lipschitz continuous; a formal convergence theorem could be developed. Finally, this work assumed statistical independence of classifiers, which may not be a good assumption for many multiple clas-

sifier systems. There has been much effort in studying the effects of correlated classifiers on ROC analysis. Convergence of fused correlated classifiers would be a rich area for study.

Bibliography

[1] Alsing, S. G. *The Evaluation of Competing Classifiers*. PhD dissertation, Air Force Institute of Technology, Wright Patterson AFB OH, March 2000.

[2] Alsing, S. G., Bauer, K. W., and Miller, J. O. "A Multinomial Selection Procedure for Evaluating Pattern Recognition Systems," *Pattern Recognition*, *35*:2397–2412 (2002).

[3] Apostol, T. *Mathematical Analysis*. MA: Addison-Wesley Publishing Company, 1974.

[4] Barnsley, M. *Fractals Everywhere*. New York: Academic Press, 1988.

[5] Clutz, T. C. *A Framework for Prognostics Reasoning*. PhD dissertation, Air Force Institute of Technology, Wright Patterson AFB OH, January 2003.

[6] Dasarathy, B. V. *Decision Fusion*. Los Alamitos, CA: IEEE Computer Society Press, 1994.

[7] Fisher, R. A. "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, *7*:179–188 (1936).

[8] Fristedt, B. and Gray, L. *A Modern Approach to Probability Theory*. Cambridge MA: Birkhauser Boston, 1997.

[9] Giacinto, G., Roli, F., and Fumera, G. "Design of Effective Multiple Classifier Systems by Clustering of Classifiers." *Proceedings Multiple Classifier Systems, First International Workshop, MCS 2000*. 160–163.

[10] Hanley, J. A. and McNeil, B. J. "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, *143*:29–36 (1982).

[11] Hansen, L. K. and Salamon, P. "Neural Network Ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(10):993–1001 (October 1990).

[12] Harrup, G. K. *ROC Analysis of IR Segmentation Techniques*. MS thesis, Air Force Institute of Technology, Wright Patterson AFB OH, December 1994.

[13] Hill, J. M. *Evaluating the Performance of Multiple Classifier Systems: A Matrix Algebra Representation of Boolean Fusion Rules*. MS thesis, Air Force Institute of Technology, Wright Patterson AFB OH, March 2003.

[14] Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. "On Combining Classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(3):226–239 (March 1998).

[15] Kuncheva, L. I., Whitaker, C. J., Shipp, C. A., and Duin, R. P. W. "Is Independence Good for Combining Classifiers?." *Proc. 15th International Conference on Pattern Recognition2*. 168–171. 2000.

[16] Oxley, M. E. and Bauer, K. W. *Classifier Fusion for Improved System Performance*. Technical Report, Air Force Institute of Technology, January 2002.

[17] Roli, F., "Fusion of Multiple Classifiers." Short Course Notes, July 2002.

[18] Saranli, A. and Demirekler, M. "A Statistical Unified Framework for Rank-Based Multiple

Classifier Decision Combination," *Pattern Recognition*, *34*:865–884 (2001).

[19] Thorsen, S. N. and Oxley, M. E. "Describing Data Fusion Using Category Theory." *Proceedings 6th International Conference on Information Fusion*. 1202–1208. July 2003.

[20] Tumer, K. and Ghosh, J. "Analysis of Decision Boundaries in Linearly Combined Neural Classifiers," *Pattern Recognition*, *29*(2):341–348 (1996).

[21] Vapnik, V. *Statistical Learning Theory*. NY: Wiley and Sons, 1998.

[22] Varshney, P. K. *Distributed Detection and Data Fusion*. New York: Springer-Verlag, 1997.

[23] Wackerly, D. D., Mendenhall III, W., and Scheaffer, R. L. *Mathematical Statistics with Applications*. Pacific Grove CA: Duxbury, 2002.

[24] Wasserman, P. D. *Advanced Methods in Neural Computing*. NY: Van Nostrand Reinhold, 1993.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
14-12-2004	**Master's Thesis**	Mar 2003 – Dec 2004

4. TITLE AND SUBTITLE

Consistency Results for the ROC Curves of Fused Classifiers

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Bjerkaas, Kristopher S., Civilian

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GAM/ENC/05-01

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFRL/SNAT
Attn: Devert Wicker
2241 Avionics Circle
WPAFB OH 45433-7321 DSN: 785-1115 x4250

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The U.S. Air Force is researching the fusion of multiple classifiers. Given a finite collection of classifiers one seeks a new fused classifier with improved performance. An established performance quantifier is the Receiver Operating Characteristic (ROC) curve, which allows one to view the probability of detection versus the probability of false alarm in one graph. Previous research shows that one does not have to perform tests to determine the ROC curve of this new fused classifier. If the ROC curve for each individual classifier has been determined, then formulas for the ROC curve of the fused classifier exist for certain fusion rules. This will be an enormous saving in time and money since the performance of many fused classifiers can be determined analytically.

In reality only finite data is available so only an estimated ROC curve can be constructed. It has been proven that estimated ROC curves will converge to the true ROC curve in probability. This research examines if convergence is preserved when these estimated ROC curves are fused. It provides a general result for fusion rules that are governed by a Lipschitz continuous ROC fusion function and establishes a metric that can be used to prove this convergence.

15. SUBJECT TERMS
Sensor fusion, classification, evaluation, receiver operating characteristic curve (ROC), consistent estimator

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
REPORT	**ABSTRACT**	**c. THIS PAGE**			Dr. Mark E. Oxley, AFIT/ENC
U	**U**	**U**	**UU**	63	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-3636, x4515; e-mail: Mark.Oxley@afit.edu

Standard Form 298 (Rev: 8-98)
Prescribed by ANSI Std. Z39-18