Consistent Binary Classification with Generalized Performance Metrics

Oluwasanmi Koyejo*
Department of Psychology,
Stanford University
sanmi@stanford.edu

Nagarajan Natarajan*
Department of Computer Science,
University of Texas at Austin
naga86@cs.utexas.edu

Pradeep Ravikumar
Department of Computer Science,
University of Texas at Austin
pradeepr@cs.utexas.edu

Inderjit S. Dhillon
Department of Computer Science,
University of Texas at Austin
inderjit@cs.utexas.edu

Abstract

Performance metrics for binary classification are designed to capture tradeoffs between four fundamental population quantities: true positives, false positives, true negatives and false negatives. Despite significant interest from theoretical and applied communities, little is known about either optimal classifiers or consistent algorithms for optimizing binary classification performance metrics beyond a few special cases. We consider a fairly large family of performance metrics given by ratios of linear combinations of the four fundamental population quantities. This family includes many well known binary classification metrics such as classification accuracy, AM measure, F-measure and the Jaccard similarity coefficient as special cases. Our analysis identifies the optimal classifiers as the sign of the thresholded conditional probability of the positive class, with a performance metric-dependent threshold. The optimal threshold can be constructed using simple plug-in estimators when the performance metric is a linear combination of the population quantities, but alternative techniques are required for the general case. We propose two algorithms for estimating the optimal classifiers, and prove their statistical consistency. Both algorithms are straightforward modifications of standard approaches to address the key challenge of optimal threshold selection, thus are simple to implement in practice. The first algorithm combines a plug-in estimate of the conditional probability of the positive class with optimal threshold selection. The second algorithm leverages recent work on calibrated asymmetric surrogate losses to construct candidate classifiers. We present empirical comparisons between these algorithms on benchmark datasets.

1 Introduction

Binary classification performance is often measured using metrics designed to address the shortcomings of classification accuracy. For instance, it is well known that classification accuracy is an inappropriate metric for rare event classification problems such as medical diagnosis, fraud detection, click rate prediction and text retrieval applications [1, 2, 3, 4]. Instead, alternative metrics better tuned to imbalanced classification (such as the F_1 measure) are employed. Similarly, cost-sensitive metrics may useful for addressing asymmetry in real-world costs associated with specific classes. An important theoretical question concerning metrics employed in binary classification is the characteri-

*Equal contribution to the work.

zation of the optimal decision functions. For example, the decision function that maximizes the accuracy metric (or equivalently minimizes the "0-1 loss") is well-known to be $\text{sign}(P(Y = 1|x) - 1/2)$. A similar result holds for cost-sensitive classification [5]. Recently, [6] showed that the optimal decision function for the F_1 measure, can also be characterized as $\text{sign}(P(Y = 1|x) - \delta^*)$ for some $\delta^* \in (0, 1)$. As we show in the paper, it is not a coincidence that the optimal decision function for these different metrics has a similar simple characterization. We make the observation that the different metrics used in practice belong to a fairly general family of performance metrics given by ratios of linear combinations of the four population quantities associated with the confusion matrix.

We consider a family of performance metrics given by ratios of linear combinations of the four population quantities. Measures in this family include classification accuracy, false positive rate, false discovery rate, precision, the AM measure and the F-measure, among others. Our analysis shows that the optimal classifiers for all such metrics can be characterized as the sign of the thresholded conditional probability of the positive class, with a threshold that depends on the specific metric. This result unifies and generalizes known special cases including the AM measure analysis by Menon et al. [7], and the F_β measure analysis by Ye et al. [6]. It is known that minimizing (convex) surrogate losses, such as the hinge and the logistic loss, provably also minimizes the underlying 0-1 loss or equivalently maximizes the classification accuracy [8]. This motivates the next question we address in the paper: can one obtain algorithms that (a) can be used in practice for maximizing metrics from our family, and (b) are *consistent* with respect to the metric? To this end, we propose two algorithms for consistent empirical estimation of decision functions. The first algorithm combines a plug-in estimate of the conditional probability of the positive class with optimal threshold selection. The second leverages the asymmetric surrogate approach of Scott [9] to construct candidate classifiers. Both algorithms are simple modifications of standard approaches that address the key challenge of optimal threshold selection. Our analysis identifies why simple heuristics such as classification using class-weighted loss functions and logistic regression with threshold search are effective practical algorithms for many generalized performance metrics, and furthermore, that when implemented correctly, such apparent heuristics are in fact asymptotically consistent.

Related Work. Binary classification accuracy and its cost-sensitive variants have been studied extensively. Here we highlight a few of the key results. The seminal work of [8] showed that minimizing certain surrogate loss functions enables us to control the probability of misclassification (the expected 0-1 loss). An appealing corollary of the result is that convex loss functions such as the hinge and logistic losses satisfy the surrogacy conditions, which establishes the statistical consistency of the resulting algorithms. Steinwart [10] extended this work to derive surrogates losses for other scenarios including asymmetric classification accuracy. More recently, Scott [9] characterized the optimal decision function for weighted 0-1 loss in cost-sensitive learning and extended the risk bounds of [8] to weighted surrogate loss functions. A similar result regarding the use of a threshold different than 1/2, and appropriately rebalancing the training data in cost-sensitive learning, was shown by [5]. Surrogate regret bounds for proper losses applied to class probability estimation were analyzed by Reid and Williamson [11] for differentiable loss functions. Extensions to the multi-class setting have also been studied (for example, Zhang [12] and Tewari and Bartlett [13]). Analysis of performance metrics beyond classification accuracy is limited. The optimal classifier remains unknown for many binary classification performance metrics of interest, and few results exist for identifying consistent algorithms for optimizing these metrics [7, 6, 14, 15]. Of particular relevance to our work are the AM measure maximization by Menon et al. [7], and the F_β measure maximization by Ye et al. [6].

2 Generalized Performance Metrics

Let $\mathcal{X}$ be either a countable set, or a complete separable metric space equipped with the standard Borel σ-algebra of measurable sets. Let $X \in \mathcal{X}$ and $Y \in \{0, 1\}$ represent input and output random variables respectively. Further, let Θ represent the set of all classifiers $\Theta = \{\theta : \mathcal{X} \mapsto [0, 1]\}$. We assume the existence of a fixed unknown distribution $\mathbb{P}$, and data is generated as iid. samples $(X, Y) \sim \mathbb{P}$. Define the quantities: $\pi = \mathbb{P}(Y = 1)$ and $\gamma(\theta) = \mathbb{P}(\theta = 1)$.

The components of the confusion matrix are the fundamental population quantities for binary classification. They are the true positives (TP), false positives (FP), true negatives (TN) and false negatives

(FN), given by:

$$\begin{aligned} \text{TP}(\theta, \mathbb{P}) &= \mathbb{P}(Y = 1, \theta = 1), & \text{FP}(\theta, \mathbb{P}) &= \mathbb{P}(Y = 0, \theta = 1), \\ \text{FN}(\theta, \mathbb{P}) &= \mathbb{P}(Y = 1, \theta = 0), & \text{TN}(\theta, \mathbb{P}) &= \mathbb{P}(Y = 0, \theta = 0). \end{aligned} \tag{1}$$

These quantities may be further decomposed as:

$$\text{FP}(\theta, \mathbb{P}) = \gamma(\theta) - \text{TP}(\theta), \quad \text{FN}(\theta, \mathbb{P}) = \pi - \text{TP}(\theta), \quad \text{TN}(\theta, \mathbb{P}) = 1 - \gamma(\theta) - \pi + \text{TP}(\theta). \tag{2}$$

Let $\mathcal{L} : \Theta \times \mathbb{P} \mapsto \mathbb{R}$ be a performance metric of interest. Without loss of generality, we assume that $\mathcal{L}$ is a utility metric, so that larger values are better. The *Bayes utility* $\mathcal{L}^*$ is the optimal value of the performance metric, i.e., $\mathcal{L}^* = \sup_{\theta \in \Theta} \mathcal{L}(\theta, \mathbb{P})$. The *Bayes classifier* θ^* is the classifier that optimizes the performance metric, so $\mathcal{L}^* = \mathcal{L}(\theta^*)$, where:

$$\theta^* = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta, \mathbb{P}).$$

We consider a family of classification metrics computed as the ratio of linear combinations of these fundamental population quantities (1). In particular, given constants (representing costs or weights) $\{a_{11}, a_{10}, a_{01}, a_{00}, a_0\}$ and $\{b_{11}, b_{10}, b_{01}, b_{00}, b_0\}$, we consider the measure:

$$\mathcal{L}(\theta, \mathbb{P}) = \frac{a_0 + a_{11}\text{TP} + a_{10}\text{FP} + a_{01}\text{FN} + a_{00}\text{TN}}{b_0 + b_{11}\text{TP} + b_{10}\text{FP} + b_{01}\text{FN} + b_{00}\text{TN}} \tag{3}$$

where, for clarity, we have suppressed dependence of the population quantities on θ and $\mathbb{P}$. Examples of performance metrics in this family include the AM measure [7], the F_β measure [6], the Jaccard similarity coefficient (JAC) [16] and Weighted Accuracy (WA):

$$\text{AM} = \frac{1}{2}\left(\frac{\text{TP}}{\pi} + \frac{\text{TN}}{1-\pi}\right) = \frac{(1-\pi)\text{TP} + \pi\text{TN}}{2\pi(1-\pi)}, \; F_\beta = \frac{(1+\beta^2)\text{TP}}{(1+\beta^2)\text{TP} + \beta^2\text{FN} + \text{FP}} = \frac{(1+\beta^2)\text{TP}}{\beta^2\pi + \gamma},$$
$$\text{JAC} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} = \frac{\text{TP}}{\pi + \text{FP}} = \frac{\text{TP}}{\gamma + \text{FN}}, \quad \text{WA} = \frac{w_1\text{TP} + w_2\text{TN}}{w_1\text{TP} + w_2\text{TN} + w_3\text{FP} + w_4\text{FN}}.$$

Note that we allow the constants to depend on $\mathbb{P}$. Other examples in this class include commonly used ratios such as the true positive rate (also known as recall) (TPR), true negative rate (TNR), precision (Prec), false negative rate (FNR) and negative predictive value (NPV):

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \; \text{TNR} = \frac{\text{TN}}{\text{FP} + \text{TN}}, \; \text{Prec} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \; \text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}}, \; \text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}.$$

Interested readers are referred to [17] for a list of additional metrics in this class.

By decomposing the population measures (1) using (2) we see that any performance metric in the family (3) has the equivalent representation:

$$\mathcal{L}(\theta) = \frac{c_0 + c_1\text{TP}(\theta) + c_2\gamma(\theta)}{d_0 + d_1\text{TP}(\theta) + d_2\gamma(\theta)} \tag{4}$$

with the constants:

$$\begin{aligned} c_0 &= a_{01}\pi + a_{00} - a_{00}\pi + a_0, & c_1 &= a_{11} - a_{10} - a_{01} + a_{00}, & c_2 &= a_{10} - a_{00} \quad \text{and} \\ d_0 &= b_{01}\pi + b_{00} - b_{00}\pi + b_0, & d_1 &= b_{11} - b_{10} - b_{01} + b_{00}, & d_2 &= b_{10} - b_{00}. \end{aligned}$$

Thus, it is clear from (4) that the family of performance metrics depends on the classifier θ only through the quantities $\text{TP}(\theta)$ and $\gamma(\theta)$.

Optimal Classifier

We now characterize the optimal classifier for the family of performance metrics defined in (4). Let ν represent the dominating measure on $\mathcal{X}$. For the rest of this manuscript, we make the following assumption:

Assumption 1. *The marginal distribution $\mathbb{P}(X)$ is absolutely continuous with respect to the dominating measure ν on $\mathcal{X}$ so there exists a density μ that satisfies $d\mathbb{P} = \mu d\nu$.*

To simplify notation, we use the standard $d\nu(x) = dx$. We also define the conditional probability $\eta_x = \mathbb{P}(Y = 1|X = x)$. Applying Assumption 1, we can expand the terms $\text{TP}(\theta) = \int_{x\in\mathcal{X}} \eta_x\theta(x)\mu(x)dx$ and $\gamma(\theta) = \int_{x\in\mathcal{X}} \theta(x)\mu(x)dx$, so the performance metric (4) may be represented as:

$$\mathcal{L}(\theta, \mathbb{P}) = \frac{c_0 + \int_{x\in\mathcal{X}}(c_1\eta_x + c_2)\theta(x)\mu(x)dx}{d_0 + \int_{x\in\mathcal{X}}(d_1\eta_x + d_2)\theta(x)\mu(x)}.$$

Our first main result identifies the Bayes classifier for all utility functions in the family (3), showing that they take the form $\theta^*(x) = \text{sign}(\eta_x - \delta^*)$, where δ^* is a metric-dependent threshold, and the sign function is given by $\text{sign} : \mathbb{R} \mapsto \{0, 1\}$ as $\text{sign}(t) = 1$ if $t \geq 0$ and $\text{sign}(t) = 0$ otherwise.

Theorem 2. *Let $\mathbb{P}$ be a distribution on $\mathcal{X} \times [0, 1]$ that satisfies Assumption 1, and let $\mathcal{L}$ be a performance metric in the family* (3)*. Given the constants $\{c_0, c_1, c_2\}$ and $\{d_0, d_1, d_2\}$, define:*

$$\delta^* = \frac{d_2\mathcal{L}^* - c_2}{c_1 - d_1\mathcal{L}^*}. \tag{5}$$

1. *When $c_1 > d_1\mathcal{L}^*$, the Bayes classifier θ^* takes the form $\theta^*(x) = \text{sign}(\eta_x - \delta^*)$*
2. *When $c_1 < d_1\mathcal{L}^*$, the Bayes classifier takes the form $\theta^*(x) = \text{sign}(\delta^* - \eta_x)$*

The proof of the theorem involves examining the first-order optimality condition (see Appendix B).

Remark 3. *The specific form of the optimal classifier depends on the sign of $c_1 - d_1\mathcal{L}^*$, and $\mathcal{L}^*$ is often unknown. In practice, one can often estimate loose upper and lower bounds of $\mathcal{L}^*$ to determine the classifier.*

A number of useful results can be evaluated directly as instances of Theorem 2. For the F_β measure, we have that $c_1 = 1 + \beta^2$ and $d_2 = 1$ with all other constants as zero. Thus, $\delta^*_{F_\beta} = \frac{\mathcal{L}^*}{1+\beta^2}$. This matches the optimal threshold for F_1 metric specified by Zhao et al. [14]. For precision, we have that $c_1 = 1, d_2 = 1$ and all other constants are zero, so $\delta^*_{\text{Prec}} = \mathcal{L}^*$. This clarifies the observation that in practice, precision can be maximized by predicting only high confidence positives. For true positive rate (recall), we have that $c_1 = 1, d_0 = \pi$ and other constants are zero, so $\delta^*_{\text{TPR}} = 0$ recovering the known result that in practice, recall is maximized by predicting all examples as positives. For the Jaccard similarity coefficient $c_1 = 1, d_1 = -1, d_2 = 1, d_0 = \pi$ and other constants are zero, so $\delta^*_{\text{JAC}} = \frac{\mathcal{L}^*}{1+\mathcal{L}^*}$.

When $d_1 = d_2 = 0$, the generalized metric is simply a linear combination of the four fundamental quantities. With this form, we can then recover the optimal classifier outlined by Elkan [5] for cost sensitive classification.

Corollary 4. *Let $\mathbb{P}$ be a distribution on $\mathcal{X} \times [0, 1]$ that satisfies Assumption 1, and let $\mathcal{L}$ be a performance metric in the family* (3)*. Given the constants $\{c_0, c_1, c_2\}$ and $\{d_0, d_1 = 0, d_2 = 0\}$, the optimal threshold* (5) *is $\delta^* = -\frac{c_2}{c_1}$.*

Classification accuracy is in this family, with $c_1 = 2, c_2 = -1$, and it is well-known that $\delta^*_{\text{ACC}} = \frac{1}{2}$. Another case of interest is the AM metric, where $c_1 = 1, c_2 = -\pi$, so $\delta^*_{\text{AM}} = \pi$, as shown in Menon et al. [7].

3 Algorithms

The characterization of the Bayes classifier for the family of performance metrics (4) given in Theorem 2 enables the design of practical classification algorithms with strong theoretical properties. In particular, the algorithms that we propose are intuitive and easy to implement. Despite their simplicity, we show that the proposed algorithms are *consistent* with respect to the measure of interest; a desirable property for a classification algorithm. We begin with a description of the algorithms, followed by a detailed analysis of consistency. Let $\{X_i, Y_i\}_{i=1}^n$ denote iid. training instances drawn from a fixed unknown distribution $\mathbb{P}$. For a given $\theta : \mathcal{X} \to \{0, 1\}$, we define the following empirical quantities based on their population analogues: $\text{TP}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \theta(X_i)Y_i$, and $\gamma_n(\theta) = \frac{1}{n}\sum_{i=1}^n \theta(X_i)$. It is clear that $\text{TP}_n(\theta) \xrightarrow{n\to\infty} \text{TP}(\theta; \mathbb{P})$ and $\gamma_n(\theta) \xrightarrow{n\to\infty} \gamma(\theta; \mathbb{P})$.

Consider the empirical measure:

$$\mathcal{L}_n(\theta) = \frac{c_1 \text{TP}_n(\theta) + c_2\gamma_n(\theta) + c_0}{d_1 \text{TP}_n(\theta) + d_2\gamma_n(\theta) + d_0}\,, \tag{6}$$

corresponding to the population measure $\mathcal{L}(\theta; \mathbb{P})$ in (4). It is expected that $\mathcal{L}_n(\theta)$ will be close to the $\mathcal{L}(\theta; \mathbb{P})$ when the sample is sufficiently large (see Proposition 8). For the rest of this manuscript, we assume that $\mathcal{L}^* \leq \frac{c_1}{d_1}$ so $\theta^*(x) = \text{sign}(\eta_x - \delta^*)$. The case where $\mathcal{L}^* > \frac{c_1}{d_1}$ is solved identically.

Our first approach (Two-Step Expected Utility Maximization) is quite intuitive (Algorithm 1): Obtain an estimator $\hat{\eta}_x$ for $\eta_x = \mathbb{P}(Y = 1|x)$ by performing ERM on the sample using a proper loss function [11]. Then, maximize $\mathcal{L}_n$ defined in (6) with respect to the threshold $\delta \in (0, 1)$. The optimization required in the third step is one dimensional, thus a global minimizer can be computed efficiently in many cases [18]. In experiments, we use (regularized) logistic regression on a training sample to obtain $\hat{\eta}$.

Algorithm 1: Two-Step EUM

Input: Training examples $\mathcal{S} = \{X_i, Y_i\}_{i=1}^n$ and the utility measure $\mathcal{L}$.
1. Split the training data $\mathcal{S}$ into two sets $\mathcal{S}_1$ and $\mathcal{S}_2$.
2. Estimate $\hat{\eta}_x$ using $\mathcal{S}_1$, define $\hat{\theta}_\delta = \text{sign}(\hat{\eta}_x - \delta)$
3. Compute $\hat{\delta} = \arg\max_{\delta \in (0,1)} \mathcal{L}_n(\hat{\theta}_\delta)$ on $\mathcal{S}_2$.

Return: $\hat{\theta}_{\hat{\delta}}$

Our second approach (Weighted Empirical Risk Minimization) is based on the observation that empirical risk minimization (ERM) with suitably weighted loss functions yields a classifier that thresholds η_x appropriately (Algorithm 2). Given a convex surrogate $\ell(t, y)$ of the 0-1 loss, where t is a real-valued prediction and $y \in \{0, 1\}$, the δ-weighted loss is given by [9]:

$$\ell_\delta(t, y) = (1 - \delta)1_{\{y=1\}}\ell(t, 1) + \delta 1_{\{y=0\}}\ell(t, 0).$$

Denote the set of real valued functions as Φ; we then define $\hat{\theta}_\delta$ as:

$$\hat{\phi}_\delta = \arg\min_{\phi \in \Phi} \frac{1}{n} \sum_{i=1}^{n} \ell_\delta(\phi(X_i), Y_i) \tag{7}$$

then set $\hat{\theta}_\delta(x) = \text{sign}(\hat{\phi}_\delta(x))$. Scott [9] showed that such an estimated $\hat{\theta}_\delta$ is consistent with $\theta_\delta = \text{sign}(\eta_x - \delta)$. With the classifier defined, maximize $\mathcal{L}_n$ defined in (6) with respect to the threshold $\delta \in (0, 1)$.

Algorithm 2: Weighted ERM

Input: Training examples $\mathcal{S} = \{X_i, Y_i\}_{i=1}^n$, and the utility measure $\mathcal{L}$.
1. Split the training data $\mathcal{S}$ into two sets $\mathcal{S}_1$ and $\mathcal{S}_2$.
2. Compute $\hat{\delta} = \arg\max_{\delta \in (0,1)} \mathcal{L}_n(\hat{\theta}_\delta)$ on $\mathcal{S}_2$.

Sub-algorithm: Define $\hat{\theta}_\delta(x) = \text{sign}(\hat{\phi}_\delta(x))$ where $\hat{\phi}_\delta(x)$ is computed using (7) on $\mathcal{S}_1$.

Return: $\hat{\theta}_{\hat{\delta}}$

Remark 5. *When $d_1 = d_2 = 0$, the optimal threshold does not depend on $\mathcal{L}^*$ (Corollary 4). We may then employ simple sample-based plugin estimates $\hat{\delta}_S$.*

A benefit of using such plugin estimates is that the classification algorithms can be simplified while maintaining consistency. Given such a sample-based plugin estimate $\hat{\delta}_S$, Algorithm 1 then reduces to estimating $\hat{\eta}_x$, and then setting $\hat{\theta}_{\hat{\delta}_S} = \text{sign}(\hat{\eta}_x - \hat{\delta}_S)$, Algorithm 2 reduces to a single ERM (7) to estimate $\hat{\phi}_{\hat{\delta}_S}(x)$, and then setting $\hat{\theta}_{\hat{\delta}_S}(x) = \text{sign}(\hat{\phi}_{\hat{\delta}_S}(x))$. In the case of AM measure, the threshold is given by $\delta^* = \pi$. A consistent estimator for π is all that is required (see [7]).

3.1 Consistency of the proposed algorithms

An algorithm is said to be $\mathcal{L}$-consistent if the learned classifier $\hat{\theta}$ satisfies $\mathcal{L}^* - \mathcal{L}(\hat{\theta}) \xrightarrow{\text{p}} 0$ i.e., for every $\epsilon > 0$, $\mathbb{P}(|\mathcal{L}^* - \mathcal{L}(\hat{\theta})| < \epsilon) \to 1$, as $n \to \infty$.

We begin the analysis from the simplest case when δ^* is independent of $\mathcal{L}^*$ (Corollary 4). The following proposition, which generalizes Lemma 1 of [7], shows that maximizing $\mathcal{L}$ is equivalent to minimizing δ^*-weighted risk. As a consequence, it suffices to minimize a suitable surrogate loss ℓ_{δ^*} on the training data to guarantee $\mathcal{L}$-consistency.

Proposition 6. *Assume $\delta^* \in (0,1)$ and δ^* is independent of $\mathcal{L}^*$, but may depend on the distribution $\mathbb{P}$. Define δ^*-weighted risk of a classifier θ as*

$$R_{\delta^*}(\theta) = E_{(x,y)\sim\mathbb{P}}\big[(1-\delta^*)1_{\{y=1\}}1_{\{\theta(x)=0\}} + \delta^* 1_{\{y=0\}}1_{\{\theta(x)=1\}}\big],$$
$$\textit{then,} \quad R_{\delta^*}(\theta) - \min_{\theta} R_{\delta^*}(\theta) = \frac{1}{c_1}(\mathcal{L}^* - \mathcal{L}(\theta)).$$

The proof is simple, and we defer it to Appendix B. Note that the key consequence of Proposition 6 is that if we know δ^*, then simply optimizing a weighted surrogate loss as detailed in the proposition suffices to obtain a consistent classifier. In the more practical setting where δ^* is not known exactly, we can then compute a sample based estimate $\hat{\delta}_S$. We briefly mentioned in the previous section how the proposed Algorithms 1 and 2 simplify in this case. Using the plug-in estimate $\hat{\delta}_S$ such that $\hat{\delta}_S \xrightarrow{p} \delta^*$ in the algorithms directly guarantees consistency, under mild assumptions on $\mathbb{P}$ (see Appendix A for details). The proof for this setting essentially follows the arguments in [7], given Proposition 6.

Now, we turn to the general case, i.e. when $\mathcal{L}$ is an arbitrary measure in the class (4) such that δ^* is difficult to estimate directly. In this case, both the proposed algorithms estimate δ to optimize the empirical measure $\mathcal{L}_n$. We employ the following proposition which establishes bounds on $\mathcal{L}$.

Proposition 7. *Let the constants a_{ij}, b_{ij} for $i,j \in \{0,1\}$, a_0, and b_0 be non-negative and, without loss of generality, take values from $[0,1]$. Then, we have:*

1. *$-2 \le c_1, d_1 \le 2, -1 \le c_2, d_2 \le 1$, and $0 \le c_0, d_0 \le 2(1+\pi)$.*
2. *$\mathcal{L}$ is bounded, i.e. for any θ, $0 \le \mathcal{L}(\theta) \le L := \frac{a_0 + \max_{i,j\in\{0,1\}} a_{ij}}{b_0 + \min_{ij\in\{0,1\}} b_{ij}}$.*

The proofs of the main results in Theorem 10 and 11 rely on the following Lemmas 8 and 9 on how the empirical measure converges to the population measure at a steady rate. We defer the proofs to Appendix B.

Lemma 8. *For any $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|\mathcal{L}_n(\theta) - \mathcal{L}(\theta)| < \epsilon) = 1$. Furthermore, with probability at least $1-\rho$, $|\mathcal{L}_n(\theta) - \mathcal{L}(\theta)| < \frac{(C+LD)r(n,\rho)}{B - Dr(n,\rho)}$, where $r(n,\rho) = \sqrt{\frac{1}{2n}\ln\frac{4}{\rho}}$, L is an upper bound on $\mathcal{L}(\theta)$, $B \ge 0, C \ge 0, D \ge 0$ are constants that depend on $\mathcal{L}$ (i.e. c_0, c_1, c_2, d_0, d_1 and d_2).*

Now, we show a uniform convergence result for $\mathcal{L}_n$ with respect to maximization over the threshold $\delta \in (0,1)$.

Lemma 9. *Consider the function class of all thresholded decisions $\Theta = \{1_{\{\phi(x)>\delta\}} \ \forall \delta \in (0,1)\}$ for a $[0,1]$-valued function $\phi : \mathcal{X} \to [0,1]$. Define $\tilde{r}(n,\rho) = \sqrt{\frac{32}{n}\big[\ln(en) + \ln\frac{16}{\rho}\big]}$. If $\tilde{r}(n,\rho) < \frac{B}{D}$ (where B and D are defined as in Lemma 8) and $\epsilon = \frac{(C+LD)\tilde{r}(n,\rho)}{B - D\tilde{r}(n,\rho)}$, then with prob. at least $1-\rho$,*

$$\sup_{\theta\in\Theta} |\mathcal{L}_n(\theta) - \mathcal{L}(\theta)| < \epsilon.$$

We are now ready to state our main results concerning the consistency of the two proposed algorithms.

Theorem 10. *(Main Result 2) If the estimate $\hat{\eta}_x$ satisfies $\hat{\eta}_x \xrightarrow{p} \eta_x$, Algorithm 1 is $\mathcal{L}$-consistent.*

Note that we can obtain an estimate $\hat{\eta}_x$ with the guarantee that $\hat{\eta}_x \xrightarrow{p} \eta_x$ by using a strongly proper loss function [19] (e.g. logistic loss) (see Appendix B).

Theorem 11. *(Main Result 3) Let $\ell : \mathbb{R} : [0, \infty)$ be a* classification-calibrated *convex (margin) loss (i.e. $\ell'(0) < 0$) and let ℓ_δ be the corresponding weighted loss for a given δ used in the weighted ERM* (7)*. Then, Algorithm 2 is $\mathcal{L}$-consistent.*

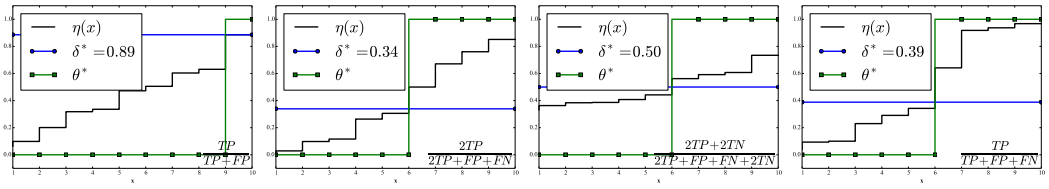
Note that loss functions used in practice such as hinge and logistic are *classification-calibrated* [8].

4 Experiments

We present experiments on synthetic data where we observe that measures from our family indeed are maximized by thresholding η_x. We also compare the two proposed algorithms on benchmark datasets on two specific measures from the family.

4.1 Synthetic data: Optimal decisions

We evaluate the Bayes optimal classifiers for common performance metrics to empirically verify the results of Theorem 2. We fix a domain $\mathcal{X} = \{1, 2, \dots 10\}$, then we set $\mu(x)$ by drawing random values uniformly in $(0, 1)$, and then normalizing these. We set the conditional probability using a sigmoid function as $\eta_x = \frac{1}{1+\exp(-wx)}$, where w is a random value drawn from a standard Gaussian. As the optimal threshold depends on the Bayes risk $\mathcal{L}^*$, the Bayes classifier cannot be evaluated using plug-in estimates. Instead, the Bayes classifier θ^* was obtained using an exhaustive search over all 2^{10} possible classifiers. The results are presented in Fig. 1. For different metrics, we plot η_x, the predicted optimal threshold δ^* (which depends on $\mathbb{P}$) and the Bayes classifier θ^*. The results can be seen to be consistent with Theorem 2 i.e. the (exhaustively computed) Bayes optimal classifier matches the thresholded classifier detailed in the theorem.



(a) Precision (b) F_1 (c) Weighted Accuracy (d) Jaccard

Figure 1: Simulated results showing η_x, optimal threshold δ^* and Bayes classifier θ^*.

4.2 Benchmark data: Performance of the proposed algorithms

We evaluate the two algorithms on several benchmark datasets for classification. We consider two measures, F_1 defined as in Section 2 and Weighted Accuracy defined as $\frac{2(TP+TN)}{2(TP+TN)+FP+FN}$. We split the training data $\mathcal{S}$ into two sets $\mathcal{S}_1$ and $\mathcal{S}_2$: $\mathcal{S}_1$ is used for estimating $\hat{\eta}_x$ and $\mathcal{S}_2$ for selecting δ. For Algorithm 1, we use logistic loss on the samples (with L_2 regularization) to obtain estimate $\hat{\eta}_x$. Once we have the estimate, we use the model to obtain $\hat{\eta}_x$ for $x \in \mathcal{S}_2$, and then use the values $\hat{\eta}_x$ as candidate δ choices to select the optimal threshold (note that the empirical best lies in the choices). Similarly, for Algorithm 2, we use a weighted logistic regression, where the weights depend on the threshold as detailed in our algorithm description. Here, we grid the space $[0, 1]$ to find the best threshold on $\mathcal{S}_2$. Notice that this step is embarrassingly parallelizable. The granularity of the grid depends primarily on class imbalance in the data, and varies with datasets. We also compare the two algorithms with the standard empirical risk minimization (ERM) - regularized logistic regression with threshold $1/2$.

First, we optimize for the F_1 measure on four benchmark datasets: (1) REUTERS, consisting of news 8293 articles categorized into 65 topics (obtained the processed dataset from [20]). For each topic, we obtain a highly imbalanced binary classification dataset with the topic as the positive class and the rest as negative. We report the average F_1 measure over all the topics (also known as macro-F_1 score). Following the analysis in [6], we present results for averaging over topics that had at least C positives in the training (5946 articles) as well as the test (2347 articles) data. (2) LETTERS dataset consisting of 20000 handwritten letters (16000 training and 4000 test instances)

from the English alphabet (26 classes, with each class consisting of at least 100 positive training instances). (3) SCENE dataset (UCI benchmark) consisting of 2230 images (1137 training and 1093 test instances) categorized into 6 scene types (with each class consisting of at least 100 positive instances). (4) WEBPAGE binary text categorization dataset obtained from [21], consisting of 34780 web pages (6956 train and 27824 test), with only about 182 positive instances in the train. All the datasets, except SCENE, have a high class imbalance. We use our algorithms to optimize for the F_1 measure on these datasets. The results are presented in Table 1. We see that both algorithms perform similarly in many cases. A noticeable exception is the SCENE dataset, where Algorithm 1 is better by a large margin. In REUTERS dataset, we observe that as the number of positive instances C in the training data increases, the methods perform significantly better, and our results align with those in [6] on this dataset. We also find, albeit surprisingly, that using a threshold $1/2$ performs competitively on this dataset.

DATASET	C	ERM	Algorithm 1	Algorithm 2
REUTERS (65 classes)	1	0.5151	0.4980	0.4855
	10	0.7624	0.7600	0.7449
	50	0.8428	0.8510	0.8560
	100	0.9675	0.9670	0.9670
LETTERS (26 classes)	1	0.4827	0.5742	0.5686
SCENE (6 classes)	1	0.3953	0.6891	0.5916
WEB PAGE (binary)	1	0.6254	0.6269	0.6267

Table 1: Comparison of methods: F1 measure. First three are multi-class datasets: F1 is computed individually for each class that has at least C positive instances (in both the train and the test sets) and then averaged over classes (macro-F1).

Next we optimize for the Weighted Accuracy measure on datasets with less class imbalance. In this case, we can see that $\delta^* = 1/2$ from Theorem 2. We use four benchmark datasets: SCENE (same as earlier), IMAGE (2068 images: 1300 train, 1010 test) [22], BREAST CANCER (683 instances: 463 train, 220 test) and SPAMBASE (4601 instances: 3071 train, 1530 test) [23]. Note that the last three are binary datasets. The results are presented in Table 2. Here, we observe that all the methods perform similarly, which conforms to our theoretical guarantees of consistency.

DATASET	ERM	Algorithm 1	Algorithm 2
SCENE	0.9000	0.9000	0.9105
IMAGE	0.9060	0.9063	0.9025
BREAST CANCER	0.9860	0.9910	0.9910
SPAMBASE	0.9463	0.9550	0.9430

Table 2: Comparison of methods: Weighted Accuracy defined as $\frac{2(TP+TN)}{2(TP+TN)+FP+FN}$. Here, $\delta^* = 1/2$. We observe that the two algorithms are consistent (ERM thresholds at $1/2$).

5 Conclusions and Future Work

Despite the importance of binary classification, theoretical results identifying optimal classifiers and consistent algorithms for many performance metrics used in practice remain as open questions. Our goal in this paper is to begin to answer these questions. We have considered a large family of generalized performance measures that includes many measures used in practice. Our analysis shows that the optimal classifiers for such measures can be characterized as the sign of the thresholded conditional probability of the positive class, with a threshold that depends on the specific metric. This result unifies and generalizes known special cases. We have proposed two algorithms for consistent estimation of the optimal classifiers. While the results presented are an important first step, many open questions remain. It would be interesting to characterize the convergence rates of $\mathcal{L}(\hat{\theta}) \xrightarrow{p} \mathcal{L}(\theta^*)$ as $\hat{\theta} \xrightarrow{p} \theta^*$, using surrogate losses similar in spirit to how excess 0-1 risk is controlled through excess surrogate risk in [8]. Another important direction is to characterize the entire family of measures for which the optimal is given by thresholded $P(Y = 1|x)$. We would like to extend our analysis to the multi-class and multi-label domains as well.

Acknowledgments: This research was supported by NSF grant CCF-1117055 and NSF grant CCF-1320746. P.R. acknowledges the support of ARO via W911NF-12-1-0390 and NSF via IIS-1149803, IIS-1320894.

References

[1] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference*, pages 3–12. Springer-Verlag New York, Inc., 1994.

[2] Chris Drummond and Robert C Holte. Severe class imbalance: Why better algorithms aren't the answer? In *Machine Learning: ECML 2005*, pages 539–546. Springer, 2005.

[3] Qiong Gu, Li Zhu, and Zhihua Cai. Evaluation measures of the classification performance of imbalanced data sets. In *Computational Intelligence and Intelligent Systems*, pages 461–471. Springer, 2009.

[4] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284, 2009.

[5] Charles Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, volume 17, pages 973–978. Citeseer, 2001.

[6] Nan Ye, Kian Ming A Chai, Wee Sun Lee, and Hai Leong Chieu. Optimizing F-measures: a tale of two approaches. In *Proceedings of the International Conference on Machine Learning*, 2012.

[7] Aditya Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *Proceedings of The 30th International Conference on Machine Learning*, pages 603–611, 2013.

[8] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[9] Clayton Scott. Calibrated asymmetric surrogate losses. *Electronic J. of Stat.*, 6:958–992, 2012.

[10] Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26 (2):225–287, 2007.

[11] Mark D Reid and Robert C Williamson. Composite binary losses. *The Journal of Machine Learning Research*, 9999:2387–2422, 2010.

[12] Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *The Journal of Machine Learning Research*, 5:1225–1251, 2004.

[13] Ambuj Tewari and Peter L Bartlett. On the consistency of multiclass classification methods. *The Journal of Machine Learning Research*, 8:1007–1025, 2007.

[14] Ming-Jie Zhao, Narayanan Edakunni, Adam Pocock, and Gavin Brown. Beyond Fano's inequality: bounds on the optimal F-score, BER, and cost-sensitive risk and their implications. *The Journal of Machine Learning Research*, 14(1):1033–1090, 2013.

[15] Zachary Chase Lipton, Charles Elkan, and Balakrishnan Narayanaswamy. Thresholding classiers to maximize F1 score. *arXiv*, abs/1402.1892, 2014.

[16] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

[17] Seung-Seok Choi and Sung-Hyuk Cha. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, pages 43–48, 2010.

[18] Yaroslav D Sergeyev. Global one-dimensional optimization using smooth auxiliary functions. *Mathematical Programming*, 81(1):127–146, 1998.

[19] Mark D Reid and Robert C Williamson. Surrogate regret bounds for proper losses. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 897–904. ACM, 2009.

[20] Deng Cai, Xuanhui Wang, and Xiaofei He. Probabilistic dyadic data analysis with local and global consistency. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 105–112. ACM, 2009.

[21] John C Platt. Fast training of support vector machines using sequential minimal optimization. 1999.

[22] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing IX*, pages 41–48. IEEE, 1999.

[23] Steve Webb, James Caverlee, and Calton Pu. Introducing the webb spam corpus: Using email spam to identify web spam automatically. In *CEAS*, 2006.

[24] Stephen Poythress Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[25] Luc Devroye. *A probabilistic theory of pattern recognition*, volume 31. springer, 1996.

[26] Aditya Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance: Supplementary material. In *Proceedings of The 30th International Conference on Machine Learning*, pages 603–611, 2013.

Appendix A

Lemma 12. *Let $\mathcal{F} = \{f : \mathcal{X} \mapsto \mathfrak{R}\}$, the constraint set $\mathcal{C} \subset \mathcal{F}$, and the functional $\mathcal{G} : \mathcal{F} \mapsto \mathfrak{R}$, consider the optimization problem:*

$$f^* = \arg\max_{f \in \mathcal{F}} \mathcal{G}(f) \quad \textit{s.t.} \quad f \in \mathcal{C}$$

If the Fréchet derivative $\nabla \mathcal{G}(f)$ exists, then f^* is locally optimal iff. $f^* \in \mathcal{C}$ and:

$$\langle \nabla \mathcal{G}(f^*), f^* - f \rangle \geq 0 \quad \forall\, f \in \mathcal{C},$$

$$\implies \int_{x \in \mathcal{X}} [\nabla \mathcal{G}(f^*)]_x f^*(x) dx \geq \int_{x \in \mathcal{X}} [\nabla \mathcal{G}(f^*)]_x f(x) dx \quad \forall\, f \in \mathcal{C}.$$

Lemma 12 is a generalization of the well known first order condition for optimality of finite dimensional optimization problems [24, Section 4.2.3] to optimization of smooth functionals.

Proposition 13. *Let $\mathcal{L}$ be a measure of the form* (4)*, and $\hat{\delta}_S$ be some estimator of its optimal threshold δ^*. Assume $\hat{\delta}_S \in (0, 1)$ and $\hat{\delta}_S \xrightarrow{p} \delta^*$. Also assume the cumulative distribution of η_x conditioned on $Y = 1$ and on $Y = 0$, $F_{\eta_x|Y=1}(z) = \mathbb{P}(\eta_x \leq z|Y = 1)$ and $F_{\eta_x|Y=0}(z) = \mathbb{P}(\eta_x \leq z|Y = 0)$ are continuous at $z = \delta^*$. Let the classifier be given by one of the following:*

- *(a) the classifier $\hat{\theta}_{\hat{\delta}_S}(x) = \text{sign}(\hat{\eta}_x - \hat{\delta}_S)$, where $\hat{\eta}$ is a class probability estimate that satisfies $E_x[|\hat{\eta}_x - \eta_x|^r] \xrightarrow{p} 0$ for some $r \geq 1$,*
- *(b) the classifier $\hat{\theta}_{\hat{\delta}_S} = \text{sign}(\hat{\phi}_{\hat{\delta}_S})$, the empirical minimizer of the ERM* (7) *using a suitably calibrated convex loss $\ell_{\hat{\delta}_S}$* [9]*,*

then $\hat{\theta}_{\hat{\delta}_S}$ is $\mathcal{L}$-consistent.

Proof. Given Proposition 6, the proofs for parts (a) and (b) essentially follow from the arguments in [7] for consistency with respect to the AM measure. Under the stated assumptions, the decomposition Lemma (Lemma 2) of [7] holds: For a classifier $\hat{\theta}$, if

$$R_{\hat{\delta}_S}(\hat{\theta}) - \min_{\theta} R_{\hat{\delta}_S}(\theta) \xrightarrow{p} 0 \quad \text{then,} \quad \mathcal{L}^* - \mathcal{L}(\hat{\theta}) \xrightarrow{p} 0$$

This allows us to directly invoke Theorems 5 and Theorems 6 of [7] giving us the desired $\mathcal{L}$-consistency in parts (a) and (b) respectively. □

Appendix B: Proofs

Proof of Theorem 2

Proof. Let $\mathcal{F} = \{f : \mathcal{X} \mapsto \mathfrak{R}\}$, and note that $\Theta \subset \mathcal{F}$. We consider a continuous extension of (4) by extending the domain of $\mathcal{L}$ from Θ to $\mathcal{F}$. This results in the following optimization:

$$f^* = \arg\max_{f \in \mathcal{F}} \mathcal{L}(f) \quad \text{s.t.} \quad f \in \Theta \tag{8}$$

It is clear that (4) is equivalent to (8), and the minima coincide i.e. $f^* = \theta^*$. The Fréchet derivative of $\mathcal{L}$ evaluated at x is given by:

$$[\nabla \mathcal{L}(f)]_x = \frac{1}{(c_1 - d_1 L(f)) D_r(f)} \left[\eta_x - \frac{d_2 \mathcal{L}(f) - c_2}{c_1 - d_1 \mathcal{L}(f)}\right] \mu(x)$$

where $D_r(f)$ is denominator of $\mathcal{L}(f)$. A function $f^* \in \Theta$ optimizes $\mathcal{L}$ if $f^* \in \Theta$ and (Lemma 12):

$$\int_{x \in \mathcal{X}} [\nabla \mathcal{L}(f^*)]_x f(x) dx \geq \int_{x \in \mathcal{X}} [\nabla \mathcal{L}(f^*)]_x f^*(x) dx \quad \forall\, f \in \Theta.$$

Thus, when $c_1 \geq d_1 \mathcal{L}^*$, a necessary condition for local optimality is that the sign of f^* and the sign of $[\nabla \mathcal{L}(f^*)]$ agree pointwise wrt. x. This is equivalent to the condition that $\text{sign}(f^*) = \text{sign}(\eta_x - \delta^*)$. Combining this result with the constraint set $f \in \Theta$, we have that $f^* = \text{sign}(f^*)$, thus $f^* = \text{sign}(\eta_x - \delta^*)$ is locally optimal. Finally, we note that $f^* = \text{sign}(\eta_x - \delta^*)$ is unique for $f \in \Theta$, thus f^* is globally optimal. The proof for $c_1 < d_1 \mathcal{L}^*$ follows using similar arguments. □

Proof of Proposition 6

Proof. From Corollary 4 we know $\delta^* = -\frac{c_2}{c_1}$. Since $0 < \delta^* < 1$, and $c_1 < 1$ from Proposition 7, we have $1 > c_1 > 0$. We can rewrite $\mathcal{L}(\theta)$ as $\mathcal{L}(\theta) = c_1[(1-\delta^*)\text{TP} + \delta^*\text{TN}] + \tilde{A}$, where $\tilde{A}$ is a constant. We have:

$$\begin{aligned}
R_{\delta^*}(\theta) &= E_{(x,y)\sim\mathbb{P}}\left[\left((1-\delta^*)1_{\{y=1\}} + \delta^* 1_{\{y=0\}}\right).1_{\{\theta(x)\neq y\}}\right] \\
&= (1-\delta^*)P(y=1,\theta(x)=0) + \delta^* P(y=0,\theta(x)=1) \\
&= (1-\delta^*)\text{FN} + \delta^*\text{FP} \\
&= (1-\delta^*)(\pi - \text{TP}) + \delta^*(1-\pi-\text{TN}) \\
&= (1-\delta^*)\pi + \delta^*(1-\pi) - \left((1-\delta^*)\text{TP} + \delta^*\text{TN}\right) \\
&= (1-\delta^*)\pi + \delta^*(1-\pi) + \frac{\tilde{A}}{c_1} - \frac{1}{c_1}\mathcal{L}(\theta).
\end{aligned}$$

Observing that $(1-\delta^*)\pi + \delta^*(1-\pi) + \frac{\tilde{A}}{c_1}$ is a constant independent of θ, the proof is complete. □

Proof of Lemma 8

Proof. For a given $\theta, \epsilon_1 > 0, \rho > 0$, there exists an N such that for any $n > N$, $\mathbb{P}(|\text{TP}_n(\theta) - \text{TP}(\theta)| < \epsilon_1) > 1 - \rho/2$ and $\mathbb{P}(|\gamma_n(\theta) - \gamma(\theta)| < \epsilon_1) > 1 - \rho/2$. By union bound, the two events simultaneously hold with probability at least $1-\rho$. Let $\tilde{c}_1 = 1/|c_1|$ if $c_1 \neq 0$ else $\tilde{c}_1 = 0$. Define $\tilde{c}_2, \tilde{d}_1, \tilde{d}_2$ similarly. Now define $C = \max(\tilde{c}_1, \tilde{c}_2)$ and $D = \max(\tilde{d}_1, \tilde{d}_2)$. Observe that either $C > 0$ or $D > 0$ otherwise $\mathcal{L}$ is a constant. Now for a given $\epsilon > 0$, after some simple algebra, we need

$$\epsilon_1 \leq \frac{(d_1\text{TP}(\theta) + d_2\gamma(\theta) + d_0)\epsilon}{D(\mathcal{L}(\theta)+\epsilon) + C}.$$

Choosing some ϵ_1 satisfying the upper bound above guarantees $\mathcal{L}(\theta) - \epsilon < \mathcal{L}_n(\theta) < \mathcal{L}(\theta) + \epsilon$. Thus for all $n > N$ implied by this ϵ_1 and ρ, $P(|\mathcal{L}_n(\theta) - \mathcal{L}(\theta)| < \epsilon) > 1 - \rho$ holds.

Now, for the rate of convergence, Hoeffding's inequality with $\rho = 4e^{-2n\epsilon_1^2}$ (or $\epsilon_1 = \sqrt{\frac{1}{2n}\ln\frac{4}{\rho}}$) gives us $\mathbb{P}(|\text{TP}_n(\theta) - \text{TP}(\theta)| < \epsilon_1) > 1 - \rho/2$ and $\mathbb{P}(|\gamma_n(\theta) - \gamma(\theta)| < \epsilon_1) > 1 - \rho/2$. Choose $\epsilon_1 > 0$ as a function of ϵ such that it is sufficiently small, i.e. $\epsilon_1 \leq \frac{(d_1\text{TP}(\theta)+d_2\gamma(\theta)+d_0)\epsilon}{D(\mathcal{L}(\theta)+\epsilon)+C}$. We know $\mathcal{L}(\theta) \leq L$ for any θ (from Proposition 7), therefore $D(\mathcal{L}(\theta)+\epsilon) + C < D(L+\epsilon) + C$. Furthermore, $d_1\text{TP}(\theta) + d_2\gamma(\theta) + d_0 > b_0 + \min(b_{00}, b_{11}, b_{01}, b_{10}) := B$. We can choose $\epsilon_1 = \frac{B\epsilon}{D(L+\epsilon)+C} \leq \frac{(d_1\text{TP}(\theta)+d_2\gamma(\theta)+d_0)\epsilon}{D(\mathcal{L}(\theta)+\epsilon)+C}$ or $\epsilon = \frac{(C+LD)\epsilon_1}{B - D\epsilon_1}$. From the first part of the lemma, we know $P(|\mathcal{L}_n(\theta) - \mathcal{L}(\theta)| < \epsilon) > 1 - \rho$ holds with probability at least ρ. This completes the proof. □

Proof of Lemma 9

Proof. Let $\rho = 16e^{\ln(en) - n\epsilon_1^2/32}$, then $\epsilon_1 = \tilde{r}(n,\rho)$. Using Lemma 29.1 in [25], we obtain:

$$\mathbb{P}\left[\sup_{\theta\in\Theta}|\text{TP}_n(\theta) - \text{TP}(\theta)| < \epsilon_1\right] > 1 - \rho/2\,.$$

By union bound, the inequalities $\mathbb{P}\left[\sup_{\theta\in\Theta}|\text{TP}_n(\theta) - \text{TP}(\theta)| < \epsilon_1\right]$ and $\mathbb{P}\left[\sup_{\theta\in\Theta}|\gamma_n(\theta) - \gamma(\theta)| < \epsilon_1\right]$ simultaneously hold with probability at least $1-\rho$. If n is large enough that $\tilde{r}(n,\rho) < \frac{B}{D}$, then from Proposition 8 we know that, for any given θ, $|\mathcal{L}_n(\theta) - \mathcal{L}(\theta)| < \frac{(C+LD)\tilde{r}(n,\rho)}{B - D\tilde{r}(n,\rho)}$ with probability at least $1-\rho$. The lemma follows. □

Proof of Theorem 10

Proof. Using a strongly proper loss function [19] and its corresponding link function ψ, and an appropriate function class to minimize the empirical loss, we can obtain a class probability estimator $\hat{\eta}$ such that $E_x\left[|\hat{\eta}_x - \eta_x|^2\right] \to 0$ (from Theorem 5 in [26]). Convergence in mean implies convergence

in probability and so we have $\hat{\eta} \xrightarrow{p} \eta$. Now let $\theta^*_\delta = \text{sign}(\eta_x - \delta)$. Recall that $\hat{\delta}$ denotes the empirical maximizer obtained in Step 3. Now, since $\mathcal{L}_n(\theta^*_{\hat{\delta}}) \geq \mathcal{L}_n(\theta^*_{\delta^*})$, it follows that:

$$\begin{aligned}
\mathcal{L}^* - \mathcal{L}(\theta^*_{\hat{\delta}}) &= \mathcal{L}^* - \mathcal{L}_n(\theta^*_{\hat{\delta}}) + \mathcal{L}_n(\theta^*_{\hat{\delta}}) - \mathcal{L}(\theta^*_{\hat{\delta}}) \\
&\leq \mathcal{L}^* - \mathcal{L}_n(\theta^*_{\delta^*}) + \mathcal{L}_n(\theta^*_{\hat{\delta}}) - \mathcal{L}(\theta^*_{\hat{\delta}}) \\
&\leq 2 \sup_\delta |\mathcal{L}(\theta^*_\delta) - \mathcal{L}_n(\theta^*_\delta)| \\
&\leq 2\epsilon \xrightarrow{p} 0
\end{aligned}$$

where ϵ is defined as in Lemma 9. The last step is true by instantiating Lemma 9 with the thresholded classifiers corresponding to $\phi(x) = \eta_x$. □

Proof of Theorem 11

Proof. For a fixed δ, $E_{(X,Y)\sim\mathbb{P}}[\ell_\delta(\hat{\theta}_\delta(X), Y)] \to \min_\theta E_{(X,Y)\sim\mathbb{P}}[\ell_\delta(\theta(X), Y)]$. With the understanding that the surrogate loss ℓ_δ (i.e. the ℓ_δ-risk) satisfies regularity assumptions and the minimizer is unique, the weighted empirical risk minimizer also converges to the corresponding Bayes classifier [9]; i.e., we have $\hat{\theta}_\delta \xrightarrow{p} \theta^*_\delta$. In particular, $\hat{\theta}_{\hat{\delta}} \xrightarrow{p} \theta^*_{\hat{\delta}} = \text{sign}(\eta_x - \hat{\delta})$. Let $\hat{\delta}$ denote the empirical maximizer obtained in Step 2. Now, by using an argument identical to the one in Theorem 10 we can show that $\mathcal{L}^* - \mathcal{L}(\theta^*_{\hat{\delta}}) \leq 2\epsilon \xrightarrow{p} 0$. □