Consistency Analysis for Binary Classification Revisited

Krzysztof Dembczyński[1] **Wojciech Kotłowski**[1] **Oluwasanmi Koyejo**[2] **Nagarajan Natarajan**[3]

Abstract

Statistical learning theory is at an inflection point enabled by recent advances in understanding and optimizing a wide range of metrics. Of particular interest are non-decomposable metrics such as the F-measure and the Jaccard measure which cannot be represented as a simple average over examples. Non-decomposability is the primary source of difficulty in theoretical analysis, and interestingly has led to two distinct settings and notions of consistency. In this manuscript we analyze both settings, from statistical and algorithmic points of view, to explore the connections and to highlight differences between them for a wide range of metrics. The analysis complements previous results on this topic, clarifies common confusions around both settings, and provides guidance to the theory and practice of binary classification with complex metrics.

1. Introduction

Real-world applications of binary classification to complex decision problems have led to the design of a wide range of evaluation metrics (Choi & Cha, 2010). Prominent examples include area under the ROC curve (AUC) for imbalanced labels (Menon et al., 2013), F-measure for information retrieval (Lewis, 1995), and precision at the top (Kar et al., 2014; 2015; Jasinska et al., 2016). To this end, several algorithms have been proposed for optimizing many of these metrics, primarily focusing on large-scale learning, without a conscious emphasis on statistical consequences of choosing models and their asymptotic behavior (Kar et al., 2015; Joachims, 2005). Wide use of such complex metrics has also re-invigorated research into their theoretical properties, which can then serve as a guide to practice (Koyejo et al., 2014a; Narasimhan et al., 2014a; Dembczyński et al., 2012; Waegeman et al., 2014; Natarajan et al., 2016).

Complex evaluation metrics for binary classification are best described as *set* metrics, or *non-decomposable* metrics – as, in general, the evaluation for a set of predictions cannot be decomposed into the average of individual instance evaluations. This is in contrast to *decomposable* metrics such as accuracy which are defined as the empirical average of the instance evaluations. This property is the primary source of difficulty in theoretical analysis, and interestingly has led to two distinct settings and notions of consistency. On one hand, Population Utility (PU) focuses on *estimation* – so a consistent PU classifier is one which correctly estimates the population optimal utility as the size of the training set (equiv. test set) increases. The PU approach has strongest roots in classical statistical analysis which often deals with asymptotically optimal estimation. On the other hand, Expected Test Utility (ETU) focuses on *generalization*. Thus, the consistent ETU classifier is one which optimizes the expected prediction error over test sets of a pre-defined size. The ETU approach has strongest roots in statistical machine learning which prizes generalization as the primary goal. Importantly, these distinctions are irrelevant when the metric is a *linear* function of the confusion matrix e.g. (weighted) accuracy and other linear metrics. To the best of our knowledge, this dichotomy was first explicitly noted by Ye et al. (2012) in the context of F-measure.[1] Like in Ye et al. (2012), our goal is not to adjudicate the correctness of either approach, but instead to explore deep connections, and highlight significant differences between both approaches for a wide range of metrics.

Contributions: We present a variety of results comparing and contrasting the PU and ETU approaches for consistent classification:

- We show that for a wide range of metrics, PU and ETU are asymptotically equivalent with respect to the size of the test set, subject to a certain p-Lipschitzness

Authors listed in the alphabetical order [1]Institute of Computing Science, Poznan University of Technology, Poland [2]Department of Computer Science, University of Illinois at Urbana-Champaign, USA [3]Microsoft Research, India. Correspondence to: Wojciech Kotłowski <wkotlowski@cs.put.poznan.pl>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

[1] Note that Ye et al. (2012) termed the two approaches Empirical Utility Maximization (EUM) and Decision Theoretic Approach (DTA), respectively. We have instead chosen the more descriptive names Population Utility (PU) and Expected Test Utility (ETU).

condition which is satisfied by many metrics of interest. This further implies asymptotic equivalence of the Bayes optimal classifiers (Section 3.1). Similar results were previously only known for F-measure.

- We provide lower bounds for the difference between PU and ETU metrics for finite test sets, and for certain metrics – thereby highlighting the difference between PU and ETU consistent classifiers with small test sets (Section 3.2).
- We analyze approximate ETU classification using low order Taylor approximations, showing that the approximation can be computed with effectively linear complexity, yet achieves low error under standard assumptions (Section 4.1).
- We consider the effects of model mis-specification and find that ETU may be more sensitive than PU, but this may be alleviated by properly calibrating the estimated probabilities (Section 4.2).

In addition, we present experimental results using simulated and real data to evaluate our theoretical claims (Section 5).

2. Preliminaries and Problem Setup

We consider the binary classification problem, where the input is a feature vector $x \in X$, and the output is a label $y \in \{0,1\}$. We assume the examples (x, y) are generated i.i.d. according to $\mathbb{P}(x, y)$. A *classifier* is a mapping $h \colon X \to \{0,1\}$. We let $\mathbb{1}_C$ denote the indicator function i.e. equal to one if C is satisfied, and zero otherwise.

Given a distribution $\mathbb{P}$ and a binary classifier h, define:

$$\begin{aligned}\mathrm{TP}(h) &= \mathbb{P}(h=1, y=1), \quad & \mathrm{TN}(h) &= \mathbb{P}(h=0, y=0),\\ \mathrm{FP}(h) &= \mathbb{P}(h=1, y=0), \quad & \mathrm{FN}(h) &= \mathbb{P}(h=0, y=1),\end{aligned}$$

which are entries of the so-called confusion matrix, namely true positives, true negatives, false positives and false negatives. In this paper, we are interested in optimizing performance metrics $\Phi(h, \mathbb{P})$ (we use explicit dependence on $\mathbb{P}$ because we will also consider the empirical version of Φ) that are functions of the above four quantities. However, since the entries of the confusion matrix are interdependent, it suffices to only use their three independent combinations. Following Natarajan et al. (2016), we parametrize $\Phi(h, \mathbb{P}) = \Phi(u(h), v(h), p)$ by means of:

$$u(h) = \mathrm{TP}(h), \quad v(h) = \mathbb{P}(h = 1), \text{ and } p = \mathbb{P}(y = 1).$$

As argued by Natarajan et al. (2016), any metric being a function of the confusion matrix can be parameterized in this way. Table 1 lists popular examples of such metrics with explicit parameterization $\Phi(u, v, p)$. Throughout the paper we assume $\Phi(u, v, p)$ is bounded from above and from below.[2]

Metric	Definition	$\Phi(u, v, p)$
Accuracy	$\mathrm{TP} + \mathrm{TN}$	$1 + 2u - v - p$
AM	$\frac{\mathrm{TP}/2}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}/2}{\mathrm{TN}+\mathrm{FP}}$	$\frac{u+p(1-v-p)}{p(1-p)}$
F_β	$\frac{(1+\beta^2)\mathrm{TP}}{(1+\beta^2)\mathrm{TP}+\beta^2\mathrm{FN}+\mathrm{FP}}$	$\frac{(1+\beta^2)u}{\beta^2 p+v}$
Jaccard	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$	$\frac{p+v-2u}{p+v-u}$
G-Mean	$\sqrt{\frac{\mathrm{TP}\cdot\mathrm{TN}}{(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})}}$	$\frac{u(1-v-p+u)}{p(1-p)}$
AUC	$\frac{\mathrm{FP}\cdot\mathrm{FN}}{(\mathrm{TP}+\mathrm{FN})(\mathrm{FP}+\mathrm{TN})}$	$\frac{(v-u)(p-u)}{p(1-p)}$

Table 1. Examples of performance metrics.

2.1. Formal Definitions of PU and ETU

Definition 1 (Population Utility (PU)). *Given a distribution $\mathbb{P}$ and classifier h, the* PU *of h for a performance metric Φ is defined as $\Phi(u(h), v(h), p)$. We let h^*_{PU} denote any maximizer of the* PU,

$$h^*_{\mathrm{PU}} \in \operatorname*{argmax}_h \Phi(u(h), v(h), p)\,.$$

In words, the PU is obtained by taking the value of metric Φ evaluated at the *expected confusion matrix* of h over $\mathbb{P}$. Thus, one can think of the PU as evaluating the classifier h on a "single test set of infinite size" drawn i.i.d. from $\mathbb{P}$.

In contrast, ETU evaluates the expected utility for a *fixed*-size test set. Formally, given a sample $S = \{(x_i, y_i)\}_{i=1}^n$ of size n, generated i.i.d. from $\mathbb{P}$, we let $\widehat{u}(h), \widehat{v}(h), \widehat{p}$ denote the corresponding empirical quantities:

$$\widehat{u}(h) = \frac{1}{n}\sum_{i=1}^n h(x_i) y_i,\ \widehat{v}(h) = \frac{1}{n}\sum_{i=1}^n h(x_i),\ \widehat{p} = \frac{1}{n}\sum_{i=1}^n y_i,$$

and the empirical value of metric Φ is then $\Phi(\widehat{u}(h), \widehat{v}(h), \widehat{p})$.

Definition 2 (Expected Test Utility (ETU)). *Let $\boldsymbol{x} = (x_1, \ldots, x_n) \in X^n$ be an arbitrary sequence of inputs. Given a distribution $\mathbb{P}$ and a classifier h, the* ETU *of h for a performance metric Φ conditioned on $\boldsymbol{x}$ is defined as:*[3]

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\Phi\big(\widehat{u}(h), \widehat{v}(h), \widehat{p}\big)\right],$$

[2] In fact, for essentially all metrics used in practice it holds $0 \le \Phi(u, v, p) \le 1$.

[3] The conditional expectation $\boldsymbol{y}|\boldsymbol{x}$ is defined up to a zero-measure set (over $\boldsymbol{x}$), but this does not create any problems as we always consider $\boldsymbol{x}$ being sampled from the data distribution.

where the expectation over $\boldsymbol{y} = (y_1, \ldots, y_n)$ is with respect to the conditional distribution $\mathbb{P}(y|\boldsymbol{x})$ i.i.d. over the examples. We let $h^*_{\text{ETU}}(\boldsymbol{x})$ denote any maximizer of the ETU*,*

$$h^*_{\text{ETU}}(\boldsymbol{x}) \in \operatorname*{argmax}_h \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \left[\Phi\big(\widehat{u}(h), \widehat{v}(h), \widehat{p}\big) \right].$$

One can think of ETU as evaluating the classifier h on "infinitely many test sets of size n" drawn i.i.d. from $\mathbb{P}$. We will see (in Section 4) that the optimal predictions (in both PU and ETU approaches) can be accurately estimated using the conditional probabilities $\mathbb{P}(y_i|x_i)$. In practice, we first obtain an estimator of the conditional probability and then compute the optimal predictions on test data based on their conditional probability estimates.

Remark 1. *More generally,* ETU *optimizes the expected utility $\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}} \left[\Phi\big(\widehat{u}(h), \widehat{v}(h), \widehat{p}\big) \right]$. However, clearly, it is sufficient to analyze the predictions at any given $\boldsymbol{x}$ (Natarajan et al., 2016) as in Definition 2.*

2.2. Well-behaved Performance Metrics

The two frameworks treat the metrics as utility measures (i.e., they are to be maximized). Further, it is reasonable to expect that $\Phi(h, \mathbb{P})$ is non-decreasing in true positive and true negative rates (and indeed, virtually all performance measures used in practice behave this way). As shown by Natarajan et al. (2016), such monotonicity in true positive and true negative rates implies another property, called *TP monotonicity*, which is better suited to the parameterization employed here.

Definition 3 (TP monotonicity)**.** *$\Phi(u, v, p)$ is said to be TP monotonic if for any v, p and $u_1 > u_2$, it holds that $\Phi(u_1, v, p) > \Phi(u_2, v, p)$.*

It is easy to verify that all measures in Table 1 are TP monotonic.

A contribution in this work is to develop a notion of regularity for metrics, that helps establish statistical connections between the two frameworks and their optimal classifiers. We call it p-Lipschitzness, defined next.

Definition 4 (p-Lipschitzness)**.** *$\Phi(u, v, p)$ is said to be p-Lipschitz if:*

$$|\Phi(u, v, p) - \Phi(u', v', p')| \le U_p|u - u'| + V_p|v - v'| + P_p|p - p'|,$$

for any feasible u, v, p, u', v', p'. The Lipschitz constants U_p, V_p, P_p are allowed to depend on p, in contrast to the standard Lipschitz functions.

The rationale behind p-Lipschitzness is that we want to control the change in value of the measure under small changes in their arguments. This property turns out to be essential to show equivalence between ETU and PU approaches. On the other hand, if we simply used a standard definition of Lipschitz function (with global constants), it would not be satisfied by many interesting measures. Hence, we weaken the Lipschitz property by allowing the constant to vary as a function of p. One can also show that general linear-fractional performance metrics studied in (Koyejo et al., 2014a; Narasimhan et al., 2015; Kotłowski & Dembczyński, 2016) satisfy p-Lipschitzness under mild conditions (Appendix A).

Proposition 1. *All measures in Table 1 are p-Lipschitz.*

Proof. We only give a proof for F_β-measure here (See Appendix A for the rest). For ease, let us denote $F_\beta(u, v, p)$ by F_β and $F_\beta(u', v', p')$ by F'_β. Let $\Delta u = u - u'$, $\Delta v = v - v'$, $\Delta p = p - p'$. We have:

$$\begin{aligned}
|F_\beta - F'_\beta| &= (1 + \beta^2) \frac{|u(\beta^2 p' + v') - u'(\beta^2 p + v)|}{(\beta^2 p + v)(\beta^2 p' + v')} \\
&= (1 + \beta^2) \frac{|\Delta u(\beta^2 p' + v') - u'\beta^2 \Delta p - u'\Delta v|}{(\beta^2 p + v)(\beta^2 p' + v')} \\
&\le \frac{1 + \beta^2}{\beta^2 p + v} \left(|\Delta u| + \frac{\beta^2 u'}{\beta^2 p' + v'} |\Delta p| + \frac{u'}{\beta^2 p' + v'} |\Delta v| \right).
\end{aligned}$$

Since $u' \le \min\{p', v'\}$, we have $\frac{\beta^2 u'}{\beta^2 p' + v'} \le 1$, $\frac{u'}{\beta^2 p' + v'} \le 1$, and thus we can choose $U_p = V_p = P_p = \frac{1+\beta^2}{\beta^2 p}$. □

As for an example of a metric which is not p-Lipschitz, consider the *precision* defined as $\Phi(u, v, p) = \frac{u}{v}$. Indeed, if v is close to zero, choosing $v' = 2v$, $u' = u$ and $p' = p$ gives:

$$\Phi(u, v, p) - \Phi(u', v', p') = \frac{u}{2v},$$

which can be arbitrarily large for sufficiently small v, while the difference $|v - v'| = v$ is small. As it turns out in Section 3.2, this pathological behavior of the precision metric is responsible for a large deviation between PU and ETU, which suggests that p-Lipschitzness is in some sense necessary to establish connections.

3. Equivalence of PU and ETU

Most of the existing literature on optimizing non-decomposable classification metrics focus on one of the two approaches in isolation. In this section, we show that the two approaches are in fact asymptotically equivalent, for a range of well-behaved metrics. Informally, given a distribution $\mathbb{P}$ and a performance metric Φ, our first result is that for sufficiently large n, the PU of the associated h^*_{ETU} is arbitrarily close to that of h^*_{PU}, and likewise, the ETU of h^*_{PU} is arbitrarily close to that of h^*_{ETU}. In contrast, we also show that the PU and ETU optimal classifiers may suffer differences for small samples.

3.1. Asymptotic Equivalence

The intuition behind the equivalence lies in the observation that the optimal classifiers under the two approaches exhibit a very simple, similar form, under mild assumptions on the distribution (Koyejo et al., 2014b; Narasimhan et al., 2014b; Natarajan et al., 2016). Let $\eta(x) := \mathbb{P}(y = 1|x)$ denote the conditional probability of positive class as a function of x. The following lemma shows that for any fixed classifier h that thresholds $\eta(x)$, and sufficiently large sample size n, its performance measured with respect to PU and ETU are *close*, in particular, differ by a factor that decays as fast as $\tilde{O}(1/\sqrt{n})$. In fact, the result holds uniformly over all such binary classifiers.

Lemma 1. *Let $\mathcal{H} = \{h \mid h = \mathbb{1}_{\eta(x)\geq\tau}, \tau \in [0, 1]\}$, be the class of thresholded binary decision functions. Let Φ be a performance metric which is p-Lipschitz. Then, with probability at least $1 - \delta$ over a random sample $S = \{(x_i, y_i)\}_{i=1}^n$ of size n generated i.i.d. from $\mathbb{P}$, it holds uniformly over all $h \in \mathcal{H}$,*

$$\Big|\Phi(u(h),v(h), p(h)) - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\Phi(\widehat{u}(h), \widehat{v}(h), \widehat{p}(h))\right]\Big| \leq 4L_p\sqrt{\frac{2\log(n+1)}{n}} + 3L_p\sqrt{\frac{\log\frac{4}{\delta}}{2n}} + \frac{L_p}{\sqrt{n}},$$

where $\widehat{u}(h), \widehat{v}(h), \widehat{p}(h)$ are empirical quantities evaluated on S, and $L_p = \max\{U_p, V_p, P_p\}$.

Remark 2. *Lemma 1 generalizes the result obtained by Ye et al. (2012) for F_β-measure to arbitrary p-Lipschitz metrics. Furthermore, using more careful bounding technique, we are able to get a better dependence on the sample size n, essentially $\tilde{O}(1/\sqrt{n})$ (neglecting logarithmic terms). In fact, this dependence cannot be improved any further in general (See Appendix B).*

The uniform convergence result in Lemma 1 enables the first main result of this work. In particular, the convergence holds when the optimal classifiers with respect to ETU and PU are of the thresholded form, i.e. $h^*_{\text{PU}} \in \mathcal{H}$, and $h^*_{\text{ETU}}(\boldsymbol{x}) \in \mathcal{H}$ almost surely (with respect to random sample of inputs $\boldsymbol{x}$), where $\mathcal{H} = \{h \mid h = \mathbb{1}_{\eta(x)\geq\tau}, \tau \in [0, 1]\}$ is the class of threshold functions on function $\eta(x)$. Several recent results have shown that the optimal classifier for many popular metrics (including all metrics in Table 1) indeed has the thresholded form (Narasimhan et al., 2014a; Lewis, 1995), under a mild condition related to continuity of the distribution of $\eta(x)$ (See the proof of Theorem 1 in Appendix B.2 for details):

Assumption 1. *The random variable $\eta(x)$ has a density (with respect to the Lebesgue measure) on $[0, 1]$.*

We are now ready to state the result. Proofs omitted in the main text are supplied in the Appendix.

Theorem 1. *Let Φ be a performance metric that is TP monotonic and p-Lipschitz, and $\mathbb{P}$ be a distribution satisfying Assumption 1. Consider the* ETU *optimal classifier h^*_{ETU} (Definition 2) and the* PU *optimal classifier h^*_{PU} (Definition 1). Then, for any given ϵ and δ, we can choose n large enough (in Definition 2 of* ETU*), such that, with probability at least $1 - \delta$ over the random choice of the sample of inputs $\boldsymbol{x}$, we have:*

$$\Big|\Phi\big(u(h^*_{\text{ETU}}(\boldsymbol{x})), v(h^*_{\text{ETU}}(\boldsymbol{x})), p\big) - \Phi\big(u(h^*_{\text{PU}}), v(h^*_{\text{PU}}), p\big)\Big| \leq \epsilon.$$

Similarly, for large enough n, with probability $1 - \delta$,

$$\Big|\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\Big[\Phi\big(\widehat{u}(h^*_{\text{ETU}}(\boldsymbol{x})), \widehat{v}(h^*_{\text{ETU}}(\boldsymbol{x})), \widehat{p}\big)\Big] - \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\Big[\Phi\big(\widehat{u}(h^*_{\text{PU}}), \widehat{v}(h^*_{\text{PU}}), \widehat{p}\big)\Big]\Big| \leq \epsilon.$$

Remark 3. *In essence, Theorem 1 suggests that, for large sample sizes, the optimal in the sense of one approach gives an accurate estimate (or a proxy) of the optimal in the sense of the other approach. Our characterization of p-Lipschitzness is key to showing the equivalence.*

3.2. Finite Sample Regime

The aforementioned result is asymptotic; to elucidate the point, we now give an example where optimal classifiers corresponding to PU and ETU differ. It is important to be aware of such extremities, especially when one applies a learned model to test data of modest sample sizes. The way we argue a lower bound is by specifying a metric and a distribution, such that on a randomly obtained test set of modest size, say m, the gap in the empirical metric computed on the test data for the two optimal classifiers can be large. As one is typically primarily interested in the empirical metric on a given test set, focusing on the empirical metric ensures fairness and forbids favoring either definition.

Example. For some constant $\alpha > 0$, consider the (adjusted) empirical precision metric defined as:

$$\Phi_{\text{Prec}}(\widehat{u}(h(\boldsymbol{x})), \widehat{v}(h(\boldsymbol{x})), p) = \frac{\widehat{u}(h(\boldsymbol{x}))}{\widehat{v}(h(\boldsymbol{x})) + \alpha}.$$

Note that $\Phi_{\text{Prec}} \in [0, \frac{1}{1+\alpha}]$; Furthermore, it is p-Lipschitz, with Lipschitz constant $V_p \propto \frac{1}{\alpha}$ (see Definition 4). Thus, choosing very small values of α implies very high Lipschitz constant, and in turn the metric becomes less "stable". To establish the desired lower bound, we choose a small $0 <$

$\alpha \ll 1$. Let $\{\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3\}$ denote a partition of the instance space $\mathcal{X}$, i.e. $\cup_{i=1}^{3} \mathcal{X}_i = \mathcal{X}$ and $\mathcal{X}_i \cap \mathcal{X}_j = 0$, for any pair (i, j). Consider the joint distribution $\mathbb{P}$ defined as:

$$\begin{aligned} \mathbb{P}(y=1|x\in\mathcal{X}_1) = 1 \quad &, \quad \mathbb{P}(y=1|x\in\mathcal{X}_3) = 0, \\ \mathbb{P}(y=1|x\in\mathcal{X}_2) \quad &= \quad 1-\epsilon = 1-\sqrt{\alpha}, \\ \mathbb{P}(\mathcal{X}_1)+\mathbb{P}(\mathcal{X}_3) = \epsilon_2 \quad &, \quad \mathbb{P}(\mathcal{X}_2) = 1-\epsilon_2. \end{aligned} \tag{1}$$

for some $1 \gg \epsilon_2 > 0$ and note that the distribution is defined to be dependent on our choice of α. The last line in the above set of equations suggests that the distribution has a small region where labels are deterministically positive or negative, but overwhelmingly positive elsewhere.

Theorem 2. *Let $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ denote a set of instances drawn i.i.d. from the distribution $\mathbb{P}$. Let $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$ denote their labels drawn from the same distribution. With probability at least $(1 - \epsilon_2 - \epsilon^n)$,*

$$\begin{aligned} &\Phi_{\text{Prec}}(\widehat{u}(h^*_{\text{ETU}}(\boldsymbol{x})), \widehat{v}(h^*_{\text{ETU}}(\boldsymbol{x})), \hat{p}) - \\ &\quad \Phi_{\text{Prec}}(\widehat{u}(h^*_{\text{PU}}(\boldsymbol{x})), \widehat{v}(h^*_{\text{PU}}(\boldsymbol{x})), \hat{p}) \geq \quad \frac{1}{n(1+\alpha)} . \end{aligned}$$

4. Algorithms: Optimization and Conditional Probability Estimation

Characterization of the optimal classifier as a thresholding of the conditional probability yields simple and efficient PU consistent estimators. The idea is to first obtain an estimator for the conditional probability using training data, and then search for an optimal threshold on a separate validation set (Narasimhan et al., 2014b; Koyejo et al., 2014b). Threshold search can be efficiently implemented in linear time (assuming probabilities are pre-sorted). In contrast, although a similar thresholding characterization exists for ETU (Natarajan et al., 2016), evaluation and prediction require the computation of an expensive expectation (Definition 2). For general metrics, there is an $O(n^3)$ procedure to determine the optimal test set labeling (Jansche, 2007; Chai, 2005; Natarajan et al., 2016), and the procedure can be sped up to $O(n^2)$ in some special cases (Ye et al., 2012; Natarajan et al., 2016). Here, we consider an approximation to ETU that requires only $O(n)$ computation, yet achieves error $O(n^{-3/2})$ compared to exact optimization.

4.1. Approximation Algorithms

Recall that ETU seeks to find the classifier of the form:

$$h^*_{\text{ETU}}(\boldsymbol{x}) = \underset{h}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \left[\Phi(\widehat{u}(h), \widehat{v}(h), \widehat{p}) \right] .$$

Following (Lewis, 1995; Natarajan et al., 2016) we know that when Φ is TP monotonic, it suffices to sort observations in decreasing order according to $\eta(x)$ and assign positive labels to top k of them, for $k = 0, \ldots, n$. Unfortunately, for each k, we need to calculate the expected utility measure, which is time consuming – requiring $O(n^2)$ in general. Our goal here is to approximate this term, so that it can be computed in $O(n)$ time, then the whole procedure can be implemented in amortized time $O(n)$.

Algorithm 1 Approximate ETU Consistent Classifier

```
Input: $\Phi$ and sorted estimates of $\eta_i$, $i = 1, 2, \ldots, n$
Init $s^*_i = 0, \forall i \in [n]$, $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\widehat{u}_0 = 0$
Set $\Phi_0 = \Phi(0, 0, \widehat{p})$
for k = $1, 2, \ldots, n$ do
    Set $\widehat{u}_k = \frac{(k-1)\widehat{u}_{k-1} + \eta_k}{k}$, $\widehat{v}_k = \frac{k}{n}$
    Set $\Phi_k = \Phi(\widehat{u}_k, \widehat{v}_k, \widehat{p})$ (via Lemmas 2 or 3)
end for
$k^* \leftarrow \arg\max_{k=0,\ldots,n} \Phi_k$.
return $\mathbf{s}^*$ s.t. $s^*_i \leftarrow 1$ for $i \in [k^*]$.
```

Fix a binary classifier $h \colon X \to \{0, 1\}$ and the input sample $\boldsymbol{x} = (x_1, \ldots, x_n)$. Let $\widehat{u}(h), \widehat{v}(h), \widehat{p}$ denote the empirical quantities, as defined in Section 3.1. Furthermore, we define *semi-empirical* quantities:

$$\widetilde{u}(h) = \frac{1}{n}\sum_{i=1}^{n} h(x_i)\eta(x_i), \qquad \text{and} \quad \widetilde{p} = \frac{1}{n}\sum_{i=1}^{n} \eta(x_i)$$

(there is no need to define $\widetilde{v}(h)$). Note that $\widetilde{u}(h) = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \left[\widehat{u}(h) \right]$, and $\widetilde{p} = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \left[\widehat{p} \right]$.

Zeroth-order approximation. Our first approximation is based on Taylor-expanding the measure up to the second order:

Lemma 2. *If Φ is twice-differentiable in (u, p) and all its second-order derivatives are bounded by constant A, then:*

$$\left| \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \left[\Phi(\widehat{u}(h), \widehat{v}(h), \widehat{p}) \right] - \Phi(\widetilde{u}(h), \widehat{v}(h), \widetilde{p}) \right| \leq \frac{A}{2n}.$$

We note that the first order terms vanish in the Taylor approximation (proof in Appendix). This constitutes a simple, yet powerful method for approximating ETU utility. Algorithm 1 outlines the resulting algorithm. As shown, the classifier can be computed in $O(n)$ time overall, assuming the data is already sorted according to $\eta(x_i)$ (otherwise, the procedure is dominated by sorting time $O(n \log n)$). We note that (Lewis, 1995) proposed a similar first order approximation, albeit without any rigorous guarantee.

Second order approximation. Naturally, we can get a better approximation by Taylor-expanding the measure up to the third order.

Lemma 3. *Assume Φ is three times differentiable in (u, p) and assume all its third-order derivatives are bounded by constant B. Let $\nabla^2_{uu}, \nabla^2_{up}, \nabla^2_{pp}$ denote the second-order* *derivative terms evaluated at* $(\widetilde{u}, \widehat{p})$*, and likewise define* $\nabla^2_{up}, \nabla^2_{pp}$*. We then have:*

$$\left|\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\Phi(\widehat{u}(h), \widehat{v}(h), \widehat{p})\right] - \Phi_{\text{appr}}(h)\right| \le \frac{B}{3n^{3/2}},$$

where:

$$\begin{aligned}\Phi_{\text{appr}}(h) &= \Phi(\widetilde{u}(h), \widehat{v}(h), \widehat{p}) \\ &\quad + \frac{1}{2}(\nabla^2_{uu} + 2\nabla^2_{up})s_u + \nabla^2_{pp}s_p,\end{aligned}$$

and

$$\begin{aligned}s_p &:= \frac{1}{n^2}\sum_{i=1}^n \eta(x_i)(1-\eta(x_i)), \\ s_u &:= \frac{1}{n^2}\sum_{i=1}^n h(x_i)\eta(x_i)(1-\eta(x_i)).\end{aligned}$$

Theorem 3 (Consistency)**.** *Given* n *instances* $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$*, sort them in decreasing order of* $\eta(x_i)$*. For* $0 \le k \le n$*, let* $\mathbf{s}^{(k)}$ *denote the vector with positions corresponding to top* k *of the sorted instances set to 1, and 0 otherwise. (a) Suppose first order derivatives are bounded by* A*, let:*

$$h^*_a = \arg\max_{\mathbf{s}^{(k)}} \Phi(\widetilde{u}(\mathbf{s}^{(k)}), \widehat{v}(\mathbf{s}^{(k)}), \widehat{p}),$$

We have:

$$\Phi(h^*_{\text{ETU}}) - \Phi(\widetilde{u}(h^*_a), \widehat{v}(h^*_a), \widehat{p}) \le \frac{A}{2n}\ .$$

(b) Suppose second order derivatives are bounded by B*, let:*

$$h^*_b = \arg\max_{\mathbf{s}^{(k)}} \Phi_{\text{appr}}(\mathbf{s}^{(k)}),$$

where $\Phi_{\text{appr}}(h)$ *is defined in Lemma 3. We have:*

$$\Phi(h^*_{\text{ETU}}) - \Phi_{\text{appr}}(h^*_b) \le \frac{2B}{3n^{3/2}}\ .$$

As before, the approximation can be computed in $O(n)$ total time. We could also expand the function up to orders higher than the third order, and get better approximations (still with $O(n)$ computation if the order of the expansion is independent of n) at the cost of an even more complicated approximation formula. In experiments, we find that on real datasets with test data sets of size 100 or more, even the zeroth order approximation is highly accurate.

4.2. Conditional Probability Estimation and Model Misspecification

So far, we assumed that we have access to the true class conditional density $\eta(x) = \mathbb{P}(y = 1|x)$ and the resulting classifier is a threshold function on $\eta(x)$. In practice, one employs some probability estimation procedure and gets $\widehat{\eta}(x)$, which we call a *model*.[4] Then, one uses $\widehat{\eta}(x)$ as if it were a true conditional probability $\eta(x)$ to obtain PU or ETU classifiers. Note that since $\eta(x)$ is unknown, and we only have access to $\widehat{\eta}(x)$, *the best we can hope for* is to choose the optimal threshold on $\widehat{\eta}(x)$ (for PU) or choose the optimal number of test set observations k to be classified as positive after sorting them according to $\widehat{\eta}(x)$ (for ETU). Next, we investigate these finite sample effects in practical PU and ETU procedures. For this analysis, we treat $\widehat{\eta}(x)$ as given and fixed, make no other assumptions on how it was obtained. Let $\widehat{\mathcal{H}} = \{h \mid h = \mathbb{1}_{\widehat{\eta}(x) \ge \tau}, \tau \in [0, 1]\}$ denote the class of binary threshold functions on $\widehat{\eta}(x)$.

Consider PU first, and let h^* be the PU-optimal classifier from $\widehat{\mathcal{H}}$, i.e.:

$$h^* = \operatorname*{argmax}_{h \in \widehat{\mathcal{H}}} \Phi(u(h), v(h), p).$$

In practice, however, one does not have access to $\mathbb{P}$, and thus $u(h), v(h), p$ cannot be computed. Instead, given $\widehat{\eta}(x)$, one uses a *validation sample* $S = \{(x_i, y_i)\}_{i=1}^n$ to choose a threshold on $\widehat{\eta}(x)$ (and thus, a classifier from $\widehat{\mathcal{H}}$), by directly optimizing the empirical version of the metric on S:

$$\widehat{h} = \operatorname*{argmax}_{h \in \widehat{\mathcal{H}}} \Phi(\widehat{u}(h), \widehat{v}(h), \widehat{p}).$$

We would like to assess how close is $\widehat{h}$ to h^*. By following the proof of Lemma 1 (which never assumes the class $\mathcal{H}$ is based on thresholding $\eta(x)$), it is easy to show that with high probability,

$$\left|\Phi(u(\widehat{h}), v(\widehat{h}), p) - \Phi(u(h^*), v(h^*), p)\right| \le O\Big(\frac{1}{\sqrt{n}}\Big).$$

Thus, if we have a sufficiently large validation sample at our disposal, we can set the threshold which maximizes the empirical version of the metric, and our performance is guaranteed to be $\widetilde{O}(1/\sqrt{n})$ close to the performance of the Φ-optimal classifier from $\widehat{\mathcal{H}}$. In other words, PU does not require to know the true distribution in order to select the best classifier in $\widehat{\mathcal{H}}$, only a sufficiently large validation sample is required.

In contrast, ETU procedure is inherently based on using $\widehat{\eta}(x)$ as a replacement for $\eta(x)$ (which we do not know) to decide upon label assignments. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be the input sample of size n. Assume for simplicity the distribution of $\eta(x)$ and $\widehat{\eta}(x)$ are continuous on $[0, 1]$, so that for any $i \neq j$, $\eta(x_i) \neq \eta(x_j)$ with probability one, and

[4] For instance, $\widehat{\eta}(x)$ could be obtained from logistic regression or neural network with soft-max function on the final layer.

similarly for $\widehat{\eta}$. Then, given $\boldsymbol{x}$ and $\widehat{\eta}$, the ETU procedure chooses the classifier of the form:

$$\widehat{h} = \operatorname*{argmax}_{h \in \widehat{\mathcal{H}}} \mathbb{E}_{\boldsymbol{y} \sim \widehat{\eta}(\boldsymbol{x})} \left[\Phi(\widehat{u}(h), \widehat{v}(h), \widehat{p}) \right].$$

Likewise, the optimal ETU classifier in $\widehat{\mathcal{H}}$ is given by:

$$h^* = \operatorname*{argmax}_{h \in \widehat{\mathcal{H}}} \mathbb{E}_{\boldsymbol{y} \sim \eta(\boldsymbol{x})} \left[\Phi(\widehat{u}(h), \widehat{v}(h), \widehat{p}) \right],$$

i.e. by definition, the optimal classifier in the restricted class $\mathcal{H}$ involves the expectation with respect to the true η. Let us denote $\Phi_{\text{ETU}} = \mathbb{E}_{\boldsymbol{y} \sim \eta(\boldsymbol{x})} \left[\Phi(\widehat{u}(h), \widehat{v}(h), \widehat{p}) \right]$, so that h^* maximizes Φ_{ETU}. In the supplementary material, we show that under some mild assumptions on Φ:

$$\mathbb{E}_{\boldsymbol{x}} \left[\left| \Phi_{\text{ETU}}(\widehat{h}) - \Phi_{\text{ETU}}(h^*) \right| \right] \leq \tilde{O}\Big(\frac{1}{\sqrt{n}} \Big) + P_p |p - p_{\widehat{\eta}}| + \sup_{h \in \widehat{\mathcal{H}}} U_p |u(h) - u_{\widehat{\eta}}(h)|,$$

where $p_{\widehat{\eta}} = \mathbb{E}\left[\widehat{\eta}(x) \right]$ and $u_{\widehat{\eta}}(h) = \mathbb{E}\left[h(x) \widehat{\eta}(x) \right]$, are the quantities corresponding to p and $u(h)$, which were calculated by replacing the conditional probability η with its estimate $\widehat{\eta}$. Thus, while for the PU procedure, the difference between $\widehat{h}$ and h^* diminishes as n grows, it is not the case of ETU, as there are two bias terms $|p - p_{\widehat{\eta}}|$ and $|u(h) - u_{\widehat{\eta}}(h)|$ which do not depend on n. These terms correspond to using incorrect conditional probability $\widehat{\eta}$ while selecting the classifier, and are present even if the sample size tends to infinity. Thus, it seems crucial for the success of ETU procedure to have $\widehat{\eta}$ *calibrated* with respect to the true distribution.

A popular choice for class probability estimation is to use logistic regression. However, if the model is misspecified, which happens often in practice, the aforementioned discussion suggests that the desired ETU solution may not be achieved. Therefore, we need to learn the class probability function more carefully. Here, we consider two variants.

The first is to use the **Isotron** algorithm. In case of the generalized linear model, i.e. $\mathbb{P}(y|x) = \gamma(\mathbf{w}^*, x)$ for some unknown *link* function γ and model $\mathbf{w}^*$, Kalai & Sastry (2009) proposed a simple and elegant algorithm (see Appendix E) that alternatively learns $\mathbf{w}^*$ and the link function γ (approximated by a piecewise linear function). It provably learns the model under certain assumptions on $\mathbb{P}$. The model $\mathbf{w}^*$ and link function γ are learned using training data, and at prediction time, the link function and the scores of training data (i.e., $x_i^T \mathbf{w}$) are used to calibrate the class probabilities $\eta(x)$ of test instances.

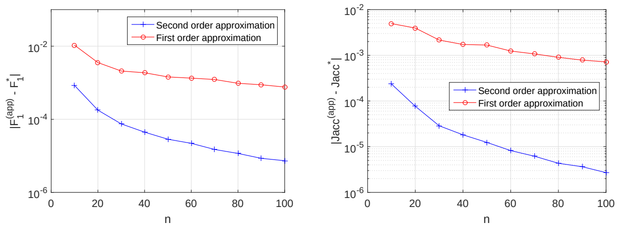
We also consider using a **recalibrated logistic model**, i.e., we first estimate the class probabilities via standard logistic regression, and recalibrate the probabilities by running one update of the γ function in Isotron algorithm (which essentially solves a quadratic problem known as the Pool of Adjacent Violators). At test time, we use the learnt γ and the logistic model to estimate $\eta(x)$ for test instances.

5. Experiments

We empirically evaluate the effectiveness and accuracy of ETU approximations introduced in Section 4.1, on synthetic as well as real datasets. We also show on several benchmark datasets that, by carefully calibrating the conditional probabilities in ETU, we can improve the classification performance.

5.1. Convergence of Approximations

We consider F_1 and Jaccard metrics from Table 1. We sample conditional probabilities η_i for n instances from the uniform distribution. The optimal predictions (see Definition 2) are obtained using Algorithm 1 of (Natarajan et al., 2016) (which is equivalent to searching over 2^n possible label vectors). Then we compute the approximate optimal predictions using the first and the second order approximations discussed in Section 4.1. For each metric, we measure the deviation between the true and the approximate optimal values with increasing sample size in Figure 1. We observe linear convergence for the first order approximation and quadratic convergence for the second order approximation. This suggests that the bounds in Theorem 3 indeed can be improved for some metrics, if not in general.



(a) F_1 (b) Jaccard

Figure 1. Convergence of approximations demonstrated on synthetic data.

5.2. Approximations on Real Data

We report results on seven multiclass and multilabel benchmark datasets: (1) LETTERS: 16000 train, 4000 test instances, (2) SCENE: 1137 train, 1093 test (3) YEAST: 1500 train, 917 test (4) WEBPAGE: 6956 train, 27824 test (5) IMAGE: 1300 train, 1010 test (6) BREAST CANCER: 463 train, 220 test instances, (7) SPAMBASE: 3071 train, 1530 test instances.[5] In case of multiclass datasets, we report results (using one-*vs*-all classifiers) averaged over classes (as

[5] See (Koyejo et al., 2014b; Ye et al., 2012) for details.

Dataset	Exact F_1	Approx (first) F_1	Approx (second) F_1	Exact Jaccard	Approx (first) Jaccard	Approx (second) Jaccard
Letters (26)	0.5666	0.5666	0.5666	0.4273	0.4272	0.4273
Scene (6)	0.6916	0.6917	0.6916	0.5374	0.5376	0.5374
Yeast (14)	0.4493	0.4494	0.4493	0.3242	0.3242	0.3242
Web page	0.6336	0.6336	0.6336	0.4637	0.4637	0.4637
Spambase	0.8448	0.8457	0.8448	0.7314	0.7326	0.7314
Image	0.8542	0.8542	0.8542	0.7455	0.7455	0.7455
Breast Cancer	0.9660	0.9660	0.9660	0.9342	0.9342	0.9342

Table 2. Comparison of ETU approximation methods: F_1 and Jaccard metrics defined in Table 1. Reported values correspond to performance evaluated on heldout data (higher values are better). For multiclass datasets (number of classes indicated in parenthesis), average performance over classes is reported. Evidently, the ETU approximations are accurate to at least 4 decimal digits in several cases.

Dataset	Logistic F_1	Isotron F_1	Recalibrated Logistic: F_1	PU F_1	Logistic Jaccard	Isotron Jaccard	Recalibrated Logistic: Jaccard	PU Jaccard
Letters (26)	0.5666	**0.5905**	0.5722	0.5745	0.4273	**0.4447**	0.4304	0.4318
Scene (6)	**0.6916**	0.6222	0.6809	0.4397	**0.5374**	0.4673	0.5362	0.4318
Yeast (14)	0.4493	0.4666	0.4629	**0.4730**	0.3242	0.3399	**0.3414**	**0.3422**
Web page	0.6336	**0.7362**	0.6756	0.6809	0.4637	**0.5825**	0.5037	0.5194
Spambase	0.8448	0.7825	**0.8905**	0.8839	0.7314	0.5729	**0.7837**	0.8003
Image	0.8542	**0.8683**	0.8569	0.8581	0.7455	0.7673	**0.7704**	0.7623
Breast Cancer	0.9660	0.9669	**0.9799**	0.9766	0.9342	0.9359	**0.9605**	0.9481

Table 3. Modeling conditional probabilities in ETU: Logistic model vs calibrated model using Isotron: F_1 and Jaccard metrics defined in Table 1. Reported values correspond to performance evaluated on heldout data (higher values are better). For multiclass datasets (number of classes indicated in parenthesis), average performance over classes is reported.

in Natarajan et al. (2016)).

We compare the exact ETU optimal, computed using the algorithm of Natarajan et al. (2016), with the approximations. The results for F_1 and Jaccard metrics are presented in Table 2. The results convincingly show that the approximations are highly accurate, and almost always indistinguishable from optimizing true metrics, on real datasets. Note that even the first-order approximation (in fact, this is zeroth-order, as the first order term is zero; see Section 4) achieves high accuracy, as the test set sizes are relatively large.

5.3. Model Misspecification

We now study how class probability estimation (CPE) and model misspecification affects the performances of PU and ETU approaches, on the seven benchmark datasets. We compare four methods: (a) ETU with logistic regression based CPE, (b) ETU with Isotron based CPE (discussed in Section 4.1), (c) ETU with recalibrated logistic regression based CPE (discussed in Section 4.1), and (d) PU using logistic regression based CPE followed by threshold tuning on validation set (Koyejo et al., 2014b). Additional comparisons to structured SVM (Joachims, 2005) and other classifiers are available in previously published work by others (Koyejo et al., 2014b; Natarajan et al., 2016), and are omitted here.

The results are presented in Table 3. We observe that the logistic model (column 1) is insufficient for many of the datasets. The results improve in several cases using the estimated generalized linear model with Isotron (column 2). However, there is a confounding factor that the two algorithms are very different, and noticed improvement may not necessarily be due to better CPE. To isolate this, recalibrated logistic model results are presented in column 3. The results are in general much better than the standard logistic model, which suggests that it is indeed the case of model misspecification in these datasets. Finally, we present the results with PU algorithm in column 4. We find that the results closely match that of the recalibrated logistic model (except in the case of Scene dataset); thus, correcting for model misspecification helps demonstrate the theorized asymptotic equivalence of PU and ETU approaches in practice.

6. Conclusions and Future Work

We have presented new results which elucidate the relationship between the two notions of consistency for complex binary classification metrics. Next, we plan to explore surrogates to further improve training efficiency non-decomposable metrics. We will also extend to more complex prediction problems such as multilabel classification, where a similar dichotomy exists.

Acknowledgments

W. Kotłowski has been supported by the Polish National Science Centre under Grant No. 2013/11/D/ST6/03050.

References

Chai, Kian Ming Adam. Expectation of F-measures: tractable exact computation and some empirical observations of its properties. In *Proceedings of the 28th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 593–594. ACM, 2005.

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

Dembczyński, Krzysztof, Waegeman, Willem, Cheng, Weiwei, and Hüllermeier, Eyke. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1-2):5–45, 2012.

Jansche, Martin. A maximum expected utility framework for binary sequence labeling. In *Annual Meeting of the Association of Computational Linguistics*, volume 45, pp. 736, 2007.

Jasinska, Kalina, Dembczynski, Krzysztof, Busa-Fekete, Róbert, Pfannschmidt, Karlson, Klerx, Timo, and Hüllermeier, Eyke. Extreme f-measure maximization using sparse probability estimates. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pp. 1435–1444, 2016.

Joachims, Thorsten. A support vector method for multivariate performance measures. In *Proceedings of the 22nd Intl. Conf. on Machine Learning*, pp. 377–384. ACM, 2005.

Kalai, Adam and Sastry, Ravi. The Isotron algorithm: High-dimensional isotonic regression. In *Conference on Learning Theory (COLT)*, 2009.

Kar, Purushottam, Narasimhan, Harikrishna, and Jain, Prateek. Online and stochastic gradient methods for non-decomposable loss functions. In *Advances in Neural Information Processing Systems*, pp. 694–702, 2014.

Kar, Purushottam, Narasimhan, Harikrishna, and Jain, Prateek. Surrogate functions for maximizing precision at the top. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 189–198, 2015.

Kotłowski, Wojciech and Dembczyński, Krzysztof. Surrogate regret bounds for generalized classification performance metrics. *Machine Learning Journal*, DOI 10.1007/s10994-016-5591-7, 2016.

Koyejo, Oluwasanmi, Natarajan, Nagarajan, Ravikumar, Pradeep K., and Dhillon, Inderjit S. Consistent binary classification with generalized performance metrics. In *Neural Information Processing Systems (NIPS)*, 2014a.

Koyejo, Oluwasanmi O, Natarajan, Nagarajan, Ravikumar, Pradeep K, and Dhillon, Inderjit S. Consistent binary classification with generalized performance metrics. In *Advances in Neural Information Processing Systems*, pp. 2744–2752, 2014b.

Lewis, David D. Evaluating and optimizing autonomous text classification systems. In *Proceedings of the 18th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 246–254. ACM, 1995.

Menon, Aditya, Narasimhan, Harikrishna, Agarwal, Shivani, and Chawla, Sanjay. On the statistical consistency of algorithms for binary classification under class imbalance. In *Proceedings of the 30th Intl. Conf. on Machine Learning*, pp. 603–611, 2013.

Mohri, Mehryar, Rostamizadeh, Afshin, and Talwalkar, Ameet. *Foundations of Machine Learning*. The MIT Press, 2012.

Narasimhan, Harikrishna, Vaish, Rohit, and Agarwal, Shivani. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In *Neural Information Processing Systems (NIPS)*, 2014a.

Narasimhan, Harikrishna, Vaish, Rohit, and Agarwal, Shivani. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In *Advances in Neural Information Processing Systems*, pp. 1493–1501, 2014b.

Narasimhan, Harikrishna, Ramaswamy, Harish, Saha, Aadirupa, and Agarwal, Shivani. Consistent multiclass algorithms for complex performance measures. In *Proceedings of the 32nd Intl. Conf. on Machine Learning*, pp. 2398–2407, 2015.

Natarajan, Nagarajan, Koyejo, Oluwasanmi, Ravikumar, Pradeep, and Dhillon, Inderjit. Optimal classification with multivariate losses. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1530–1538, 2016.

Waegeman, Willem, Dembczynski, Krzysztof, Jachnik, Arkadiusz, Cheng, Weiwei, and Hüllermeier, Eyke. On the bayes-optimality of F-measure maximizers. *Journal of Machine Learning Research*, 15:3333–3388, 2014.

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