Optimal Classification with Multivariate Losses

Nagarajan Natarajan T-NANATA@MICROSOFT.COM
Microsoft Research, INDIA

Oluwasanmi Koyejo SANMI@ILLINOIS.EDU
Stanford University, CA & University of Illinois at Urbana-Champaign, IL, USA

Pradeep Ravikumar PRADEEPR@CS.UTEXAS.EDU
Inderjit S. Dhillon INDERJIT@CS.UTEXAS.EDU
The University of Texas at Austin, TX, USA

Abstract

Multivariate loss functions are extensively employed in several prediction tasks arising in Information Retrieval. Often, the goal in the tasks is to minimize expected loss when retrieving relevant items from a presented set of items, where the expectation is with respect to the joint distribution over item sets. Our key result is that for most multivariate losses, the expected loss is provably optimized by sorting the items by the conditional probability of label being positive and then selecting top k items. Such a result was previously known only for the F-measure. Leveraging on the optimality characterization, we give an algorithm for estimating optimal predictions in practice with runtime quadratic in size of item sets for many losses. We provide empirical results on benchmark datasets, comparing the proposed algorithm to state-of-the-art methods for optimizing multivariate losses.

1. Introduction

A recent flurry of theoretical results and practical algorithms highlights a growing interest in understanding and optimizing general multivariate losses (Joachims, 2005; Petterson and Caetano, 2011; Dembczynski et al., 2011; Ye et al., 2012; Koyejo et al., 2014; Narasimhan et al., 2014). Conventional losses such as the 0-1 loss (or the error) in binary or multiclass classification, and Hamming loss in case of multilabel learning fall short in many applications, such as medical diagnosis, fraud detection, and information retrieval, with imbalanced and rare event classification tasks (Lewis and Gale, 1994; Drummond and Holte, 2005; Gu et al., 2009; He and Garcia, 2009). Practitioners employ multivariate loss functions such as the F-measure that capture non-linear trade-off between the entries of the "confusion matrix", namely, true positives, false positives, true negatives and false negatives. Multivariate loss functions are defined on vector-valued predictions, such as label vectors in multilabel learning, subsets of classes in multiclass classification, etc. The goal in the prediction tasks is to minimize expected loss when retrieving relevant items from a presented set of items, where the expectation is with respect to the joint distribution over item sets (that models uncertainty in the data).

Algorithmic approaches for minimizing expected multivariate losses such as structured support vector machines have been proposed (Joachims, 2005; Petterson and Caetano, 2011). Interestingly, the optimization can be performed efficiently for most multivariate losses that can be written as a function of the entries of the confusion matrix. However, there are no known consistency guarantees for these approaches; in fact, recently, Dembczynski et al. (2013) showed that structured loss minimization is not consistent for the F-measure. On the other hand, an important theoretical question is if and when one can explicitly characterize the optimal solution for the loss minimization problem. A key difficulty in the analysis of general multivariate losses is that they are often *non-decomposable*, i.e., the loss on prediction vectors does not decompose into the sum of losses over individual predictions. Two decades ago, Lewis (1995) showed that the optimal solution to minimizing expected F-measure, under the assumption that the labels are conditionally independent, admits a simple form — in particular, it requires only the knowledge of the marginals $\mathbb{P}(Y_i|\mathbf{x})$. Then, Dembczynski et al. (2011) showed that the optimal solution for F-measure can be de-

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

scribed using $O(n^2)$ parameters for a general distribution $\mathbb{P}$ over sets of n items. Since then, there has been extensive work focusing on binary and multilabel F-measure (Dembczyński et al., 2012; Ye et al., 2012; Dembczynski et al., 2013; Waegeman et al., 2014; Lipton et al., 2014). Yet, the question of characterizing optimal prediction for many other losses used in practice remains largely unknown.

In this paper, we provide a general theoretical analysis of expected loss minimization for general, non-decomposable, multivariate losses in binary, multiclass and multilabel prediction problems. Under conditional independence assumption on the underlying distribution, we show that the optimal solution for most losses exhibit a simple form, and depends only on the conditional probability of the label being positive. In multiclass classification, the joint distribution is a multinomial, and we show that a similar result holds. In particular, we identify a natural sufficient condition for any loss under which it allows such a simple characterization of optimal — we require the loss to be monotonically decreasing in true positives; this is satisfied by most, if not all, loss functions including the monotonic family studied by Narasimhan et al. (2014), and the linear fractional family studied by Koyejo et al. (2014; 2015). As a special case of our analysis, we naturally recover the F-measure result of Lewis (1995) for binary classification, and the result of Coz Velasco et al. (2009) for multiclass classification.

Minimizing (and even evaluating) expected multivariate losses can involve exponential-time computation, even when given access to the exact label distribution. As we show, in light of our main result characterizing optimal predictions, and with careful implementation, computations can be greatly simplified. We give an algorithm that runs in $O(n^3)$ time for a general loss, where n is the size of the item set (number of instances or labels or classes as the case maybe). For special cases such as F_β and Jaccard, the algorithm can be implemented to run in time $O(n^2)$. We prove that our overall procedure for computing the optimal in practice is consistent. We also support our theoretical results with experimental evaluation on synthetic and real-world datasets.

Related Work: We highlight some of the key results relating to prediction with general multivariate losses. Existing theoretical analysis has focused on two distinct approaches for characterizing the *population* version of multivariate losses: identified by Ye et al. (2012) as decision theoretic analysis (DTA) and empirical utility maximization (EUM). In DTA, the goal is to minimize the expected loss of a classifier on sets of predictions, which is the setting in our work here, while in EUM, the goal is to minimize the loss applied to population confusion matrix with *expected* values as entries. We can interpret DTA as minimizing the average loss over an infinite set of test sets, each of a fixed size, while EUM as minimizing the loss of a classifier over a single infinitely large test set. More recently, there have been several theoretical and algorithmic advances relating to general performance measures (Parambath et al., 2014; Koyejo et al., 2014; Narasimhan et al., 2014; Kar et al., 2014; Narasimhan et al., 2015; Koyejo et al., 2015) used in binary, multiclass and multilabel classification in the EUM setting. In stark contrast, we know much less about the setting in our paper; several authors have proposed algorithms for empirical optimization of the expected F_β measure including Chai (2005), Jansche (2007) and Dembczynski et al. (2011). Ye et al. (2012) compare the DTA and the EUM analyses for F_β, showing an asymptotic equivalence as the number of test samples goes to infinity. Quevedo et al. (2012) propose a cubic complexity dynamic programming algorithm for computing optimal labeling under conditional label independence assumption, albeit without providing an optimality characterization or consistency.

2. Multivariate Loss Minimization

We consider the problem of making multivariate predictions $\mathbf{y} = \{y_1, y_2, \ldots, y_n\} \in \{0,1\}^n$, for a given set of instances (described by their features) $\mathbf{x} \in \mathcal{X}$. Let $\mathbf{X}$ denote the random variable for instances, $\mathbf{Y}$ denote the random variable for label vectors of length n (with Y_i denoting the random variable for ith label). Given a multivariate loss L, the goal is to minimize the expected loss wrt. the underlying joint distribution $\mathbb{P}$ over $\mathbf{X}$ and $\mathbf{Y}$:

$$\mathbf{h}^* = \operatorname*{arg\,min}_{\mathbf{h}:\mathbf{x}\mapsto\mathbf{y}} \; \mathbf{E}_{(\mathbf{X},\mathbf{Y})\sim\mathbb{P}} L(\mathbf{h}(\mathbf{X}), \mathbf{Y}).$$

Note that $\mathbf{h}^*$ optimizes the expected loss wrt. to the conditional $\mathbb{P}(\mathbf{Y}|\mathbf{x})$ at each $\mathbf{x}$; therefore it is sufficient to analyze the optimal predictions at a given $\mathbf{x}$.

$$\mathbf{h}^*(\mathbf{x}) = \operatorname*{arg\,min}_{\mathbf{h}:\mathbf{x}\mapsto\mathbf{y}} \; \mathbf{E}_{\mathbf{Y}\sim\mathbb{P}(\mathbf{Y}|\mathbf{x})} L(\mathbf{h}(\mathbf{x}), \mathbf{Y}). \qquad (1)$$

The choices of $\mathbf{x}, \mathbf{y}$ and L depend on the prediction task:

- In **binary classification**, $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ is a set of instances, $y_i \in \{0,1\}$ corresponds to the label of instance x_i and $h_i(\mathbf{x}) \in \{0,1\}$ denotes the predicted label for x_i. One may be interested in retrieving the positive instances with high precision or high recall (or a function of the two, such as the F-measure). We let $L : \{0,1\}^n \times \{0,1\}^n \to \mathbb{R}_+$.
- In **multiclass classification** with n classes, $\mathbf{x} = \{x\}$ is an instance, $y \in [n]$ corresponds to the class of instance x, whereas the prediction $\mathbf{h}(x) \subseteq [n]$ corresponds to a *subset* of predicted classes (represented by a binary vector in $\{0,1\}^n$). Here we want to retrieve the true class in a predicted subset of small size

$k \ll n$; so, $L : [n] \times \{0,1\}^n \to \mathbb{R}_+$. This problem can be thought of as learning non-deterministic classifiers (Coz Velasco et al., 2009).

- In **multilabel learning** with n labels, $\mathbf{x} = \{x\}$ is an instance, $\mathbf{y}, \mathbf{h}(x) \in \{0,1\}^n$ correspond to the set of relevant and predicted labels respectively for instance x. Popular choices of loss functions include multilabel F-measure and the Hamming loss. Here, $L : \{0,1\}^n \times \{0,1\}^n \to \mathbb{R}_+$.

2.1. Hardness of the general setting

The optimization problem (1) essentially entails a search over binary vectors of length n. Efficient search for optimal solution or obtaining a consistent estimator can be notoriously hard depending on the loss function and the joint distribution. Of course, if the loss function is *decomposable* over instances (such as the Hamming loss in case of multilabel learning, or 0-1 loss in case of binary classification), a consistent estimator of the optimal solution can be obtained via empirical risk minimization. Dembczynski et al. (2011) showed that for the F-measure, and arbitrary distribution $\mathbb{P}$, optimal solution to (1) can be obtained provided we can estimate $O(n^2)$ parameters of $\mathbb{P}$ — in particular, $\mathbb{P}(Y_i = 1, \sum_{i=1}^n Y_i = s|\mathbf{x}), i, s \in [n] \times [n]$. In case of the subset 0/1 loss defined as $L(\mathbf{h}(\mathbf{x}), \mathbf{y}) = 1$ if $\mathbf{h}(\mathbf{x}) \neq \mathbf{y}$ and $L(\mathbf{h}(\mathbf{x}), \mathbf{y}) = 0$ otherwise, $\mathbf{h}^*(\mathbf{x})$ is the mode of the distribution $\mathbb{P}(\mathbf{Y}|\mathbf{x})$, which is infeasible to estimate for arbitrary $\mathbb{P}$. For general multivariate losses, one popular algorithmic approach is to employ structural support vector machines (Joachims, 2005; Petterson and Caetano, 2011) which optimize a convex upper bound of the expected loss on training data. However, there are no consistency results known for the approach. In fact, Dembczynski et al. (2013) show that structural SVMs are inconsistent for arbitrary $\mathbb{P}$, in case of the F-measure. More recently, Wang et al. (2015) study the multivariate loss minimization problem from an adversarial point of view, and provide a game-theoretic solution. Inevitably, they require solving sub-problems of the form (1), which (as they discuss) can be worked out for a few specific losses (such as the F-measure) but are hard in general.

2.2. Conditional independence

Consider the setting when $\mathbb{P}(\mathbf{Y}|\mathbf{x})$ satisfies conditional independence. In case of binary classification, instances are typically assumed to be i.i.d., and therefore conditional independence $\mathbb{P}(\mathbf{Y}|\mathbf{x}) = \Pi_{i=1}^n \mathbb{P}(Y|x_i)$ holds. In case of multilabel learning, conditional label independence $\mathbb{P}(\mathbf{Y}|x) = \Pi_{i=1}^n \mathbb{P}(Y_i|x)$ may be strong, as labels are likely to be correlated in practice. It has been known for a long time that for the F-measure, under conditional independence (Lewis, 1995), the optimal solution to (1) can be computed by simply sorting the instances according to $\mathbb{P}(Y|x_i)$ and setting the labels for the top k instances to 1 and the rest to 0 (for some $0 \leq k \leq n$). As a consequence, we only require estimates of the marginals to compute the optimal solution. For convenience, denote $\mathbf{h}(\mathbf{x})$ by $\mathbf{s} \in \{0,1\}^n$ (and the optimal solution to (1) by $\mathbf{s}^*$).

Theorem (Lewis (1995))**.** *Consider:*

$$L_{F_\beta}(\mathbf{s}, \mathbf{y}) = 1 - \frac{(1+\beta)^2 \sum_{i=1}^n s_i y_i}{\sum_{i=1}^n s_i + \beta^2 \sum_{i=1}^n y_i}. \tag{2}$$

Let x_i's be sorted in decreasing order of the marginals $\mathbb{P}(Y|x_i)$. Then, the optimal predictions $\mathbf{s}^*$ (1) for the F_β loss in (2) is given by $s_i^* = 1$, for $i \in [k^*]$, $s_i^* = 0$ otherwise, for some $0 \leq k^* \leq n$ that may depend on $\mathbf{x}$.

But such a characterization of optimality is not known for other multivariate losses used in practice.

2.3. Multinomial distributions

Note that in case of multiclass classification with n classes, $C_1, C_2, \ldots, C_n$, one can alternatively think of a joint distribution $\mathbb{P}(\mathbf{Y}|x)$ over label vectors $\mathbf{y} \in \{0,1\}^n$ such that the distribution is supported only on $\mathbf{y}$ satisfying $\sum_{i=1}^n y_i = 1$. Letting $e_i \in \{0,1\}^n$ be the indicator vector for class C_i, define:

$$\begin{aligned} \mathbb{P}(\mathbf{Y} = e_i|x) &= \mathbb{P}(Y = C_i|x), \\ \mathbb{P}(\mathbf{Y} = \mathbf{y}|x) &= 0, \text{ otherwise.} \end{aligned} \tag{3}$$

Characterizing optimal solution is simple in this setting as well. Coz Velasco et al. (2009) proved a result very similar to that of (Lewis, 1995) (although the proof turns out to be much simpler in this case), in the context of multiclass classification with F_β-measure.

Theorem (Coz Velasco et al. (2009))**.** *Consider:*

$$\mathbf{s}^* = \operatorname*{arg\,min}_{\mathbf{s} \in \{0,1\}^n} \mathbf{E}_{Y \sim \mathbb{P}(Y|x)} L_{F_\beta}(\mathbf{s}, e_Y),$$

where L_{F_β} is defined as in (2). Let the classes C_i be indexed in the decreasing order of the conditionals $\mathbb{P}(Y = C_i|x)$. Then, $\mathbf{s}^*$ satisfies $s_i^* = 1$, for $i \in [k^*]$, $s_i^* = 0$ otherwise, for some $0 \leq k^* \leq n$ that may depend on $\mathbf{x}$.

Next, we show that the above results indeed hold for general multivariate losses.

3. Main Results

We now show that for most multivariate losses L, the corresponding optimal predictions (1) can be explicitly characterized for multinomial distributions (as in the case of multiclass classification) and joint distributions satisfying conditional independence (as in the cases of binary classification and multilabel learning with conditional label independence assumption). To this end, we consider general losses that are functions of the entries of the so-called confusion matrix, namely true positives, true negatives, false positives and false negatives. The empirical confusion matrix is computed as $\widehat{\mathbf{C}}(\mathbf{s},\mathbf{y}) = \begin{bmatrix} \widehat{\text{TP}} & \widehat{\text{FN}} \\ \widehat{\text{FP}} & \widehat{\text{TN}} \end{bmatrix}$ with entries:

$$\widehat{\text{TP}} = \frac{1}{n}\sum_{i=1}^{n} s_i y_i,\ \widehat{\text{TN}} = \frac{1}{n}\sum_{i=1}^{n}(1-s_i)(1-y_i),$$

$$\widehat{\text{FP}} = \frac{1}{n}\sum_{i=1}^{n} s_i(1-y_i),\ \widehat{\text{FN}} = \frac{1}{n}\sum_{i=1}^{n}(1-s_i)y_i.$$

Most multivariate loss functions used in practice (see Table 1) admit the above form. For example, L_{F_β} in (2) can be written as:

$$L_{F_\beta}(\mathbf{s},\mathbf{y}) := L_{F_\beta}(\widehat{\mathbf{C}}(\mathbf{s},\mathbf{y})) = 1 - \frac{(1+\beta^2)\widehat{\text{TP}}}{(1+\beta^2)\widehat{\text{TP}} + \beta^2\widehat{\text{FN}} + \widehat{\text{FP}}}.$$

Without loss of generality, we let $L : [0,1]^4 \mapsto \mathbb{R}_+$ denote a multivariate loss function evaluated on the entries of the confusion matrix. We begin by observing the following equivalent representation for loss L. All the proofs in the manuscript are supplied in the Appendix due to limited space.

Proposition 1. *Let* $u(\mathbf{s},\mathbf{y}) = \widehat{TP}(\mathbf{s},\mathbf{y}), v = v(\mathbf{s}) := \frac{1}{n}\sum_i s_i$ *and* $p = p(\mathbf{y}) := \frac{1}{n}\sum_i y_i$*, then* $\exists\ \Phi : [0,1]^3 \to \mathbb{R}_+$ *such that* $L(\widehat{\mathbf{C}}(\mathbf{s},\mathbf{y})) = \Phi(u(\mathbf{s},\mathbf{y}), v(\mathbf{s}), p(\mathbf{y}))$.

Next, we define a certain monotonicity property which can be seen to be satisfied by popular loss functions.

Definition 1 (TP Monotonicity)**.** *A loss function L is said to be TP monotonic if for any v, p, and $u_1 > u_2$, it holds that $\Phi(u_{1,}v, p) < \Phi(u_2, v, p)$.*

In other words, L satisfies TP monotonicity if the corresponding representation Φ (Proposition 1) is monotonically decreasing in its first argument. It is easy to verify, for instance that $\Phi_{F_\beta}(u,v,p) = 1 - \frac{(1+\beta^2)u}{\beta^2 p + v}$ is monotonically decreasing in u. We are now ready to state our main result regarding minimizing expected losses wrt. distributions satisfying conditional independence.

Theorem 1 (Binary Losses)**.** *Assume the joint distribution $\mathbb{P}$ satisfies conditional independence. Let L be a multivariate loss that satisfies TP monotonicity. Then, the optimal predictions $\mathbf{s}^* := \mathbf{h}^*(\mathbf{x})$ in* (1) *satisfy:*

$$\min\{\mathbb{P}(Y=1|x_i)|s_i^*=1\} \geq \max\{\mathbb{P}(Y=1|x_i)|s_i^*=0\}.$$

In other words, $s_i^* = 1$, for all $i \in [k^*]$, for some $0 \leq k^* \leq n$ that may depend on $\mathbf{x}$, and $s_i^* = 0$ for $i \notin [k^*]$.

Note that the above result also applies to multilabel classification, albeit under conditional label independence assumption. Next, we give a result for multiclass classification, where the joint distribution is multinomial.

Theorem 2 (Multiclass Losses)**.** *Fix a multinomial distribution $\mathbb{P}$* (3)*. Let L be a multivariate loss that satisfies TP monotonicity. Let the classes C_i be indexed in the decreasing order of the conditionals $\mathbb{P}(Y = C_i|x)$. Then the optimal predictions $\mathbf{s}^* := \mathbf{h}^*(\mathbf{x})$ in* (1) *is given by: $s_i^* = 1$, for $i \in [k^*]$, for some $0 \leq k^* \leq n$ that may depend on $\mathbf{x}$, and $s_i^* = 0$ otherwise.*

The proof essentially is a direct consequence of TP monotonicity and thus gives a generalization of the result in (Coz Velasco et al., 2009).

3.1. Recovered and New Results

It is clear that our results generalize those by Lewis (1995) and by Coz Velasco et al. (2009) for expected loss minimization. Now, we draw attention to some of the recent optimality results in classification using general loss functions, where the objective is different from (1). The motivation of this section is to highlight the generality of our analysis and place our contributions in the context of recent results in a closely related setting. For a binary classifier $h : \mathcal{X} \mapsto \{0,1\}$ and a distribution $\mathbb{P}$ over $\mathcal{X} \times \{0,1\}$, let $\mathbf{C}(h;\mathbb{P}) = \begin{bmatrix} \text{TP} & \text{FN} \\ \text{FP} & \text{TN} \end{bmatrix}$ represent the *population* confusion matrix with entries:

$$\text{TP} = \mathbb{P}(h(x)=1, y=1),\quad \text{TN} = \mathbb{P}(h(x)=0, y=0),$$
$$\text{FP} = \mathbb{P}(h(x)=1, y=0),\quad \text{FN} = \mathbb{P}(h(x)=0, y=1).$$

Koyejo et al. (2014) and Narasimhan et al. (2014) are interested in optimizing the following notion of expected loss:

$$h^* = \operatorname*{arg\,min}_{h:\mathcal{X}\to\{0,1\}} L(\mathbf{C}(h;\mathbb{P}))$$

i.e., L is applied to the population confusion matrix. Under mild assumptions on the distribution $\mathbb{P}$, they show for a large family of loss functions, that the optimal solution satisfies a thresholded form, i.e. $h^*(x) = \text{sign}(\mathbb{P}(Y|x) - \delta^*)$, where δ^* is a constant that depends only on the distribution and the loss itself. In contrast, for the expected loss minimization in (1), there need not be a threshold in general that minimizes the expected loss on a given $\mathbf{x}$ (as also discussed in (Lewis, 1995)).

The Fractional Linear Family: Koyejo et al. (2014; 2015) studied a large family L_{FL} of losses that are represented by:

$$\Phi_{\text{FL}}(\widehat{\text{TP}}(\mathbf{s},\mathbf{y}), v(\mathbf{s}), p(\mathbf{y})) = \frac{c_0 + c_1\widehat{\text{TP}} + c_2 v + c_3 p}{d_0 + d_1\widehat{\text{TP}} + d_2 v + d_3 p} \tag{4}$$

for constants $c_i, d_i, i = \{0, 1, 2, 3\}$. We identify a subclass of L_{FL} where our results apply. The following result can be proven by inspection and is stated without proof.

Proposition 2. *If $c_1 < d_1$, then L_{FL} satisfies TP monotonicity.*

Note that this essentially constitutes the most useful losses in L_{FL} where increase in true positives (for a fixed total number of predictions) leads to decrease in loss.

Metrics from Narasimhan et al. (2014): The following alternative three-parameter representation of losses L was studied by Narasimhan et al. (2014). Let $p = p(\mathbf{y}) := \frac{1}{n}\sum_i y_i$, $r_p = \widehat{\text{TPR}}(\mathbf{s}, \mathbf{y}) = \frac{\widehat{\text{TP}}(\mathbf{s},\mathbf{y})}{p(\mathbf{y})}$ and $r_n = \widehat{\text{TNR}}(\mathbf{s}, \mathbf{y}) = \frac{\widehat{\text{TN}}(\mathbf{s},\mathbf{y})}{1-p(\mathbf{y})}$, then $\exists\, \Gamma : [0, 1]^3 \to \mathbb{R}_+$ such that:

$$L(\widehat{\mathbf{C}}(\mathbf{s}, \mathbf{y})) = \Gamma(\widehat{\text{TPR}}(\mathbf{s}, \mathbf{y}), \widehat{\text{TNR}}(\mathbf{s}, \mathbf{y}), p(\mathbf{y})). \qquad (5)$$

As shown in Table 1, many losses used in practice are easily represented in this form. Representation for additional losses is simplified by including the empirical precision, given by $\widehat{\text{Prec}}(\mathbf{s}, \mathbf{y}) = \frac{\widehat{\text{TP}}(\mathbf{s},\mathbf{y})}{v(\mathbf{s})}$, where $v(\mathbf{s}) := \frac{1}{n}\sum_i s_i = \widehat{\text{TP}} + \widehat{\text{FP}}$. Consider the following monotonicity property relevant to the representation (5).

Definition 2 (TPR/TNR Monotonicity)**.** *A loss L is said to be TPR/TNR monotonic if when $r_{p1} > r_{p2}$ and $r_{n1} > r_{n2}$ and p fixed, then $\Gamma(r_{p1}, r_{n1}, p) < \Gamma(r_{p2}, r_{n2}, p)$.*

In other words, L satisfies TPR/TNR monotonicity if the corresponding Γ in (5) is monotonically decreasing in its first two arguments. It can be shown that all the losses listed in Table 1 satisfy TPR/TNR monotonicity. The following proposition states that any loss function that satisfies TPR/TNR monotonicity also satisfies TP monotonicity.

Proposition 3. *If L satisfies TPR/TNR monotonicity, then L satisfies TP monotonicity.*

We can verify from the third column of Table 1 that each of the TPR/TNR monotonic losses $\Phi(u, v, p)$ is monotonically decreasing in u. So any loss that satisfies TPR/TNR monotonicity admits optimal classifier (1) as stated in Theorems 1 and 2.

The area under the ROC curve (AUC): AUC is an important special case, which reduces to Φ_{AUC} (see Table 1) in case of binary predictions (Joachims, 2005). It is clear that following proposition follows directly, and is stated without proof:

Proposition 4. *Φ_{AUC} satisfies TP monotonicity.*

While prior work on AUC has focused on optimizing prediction of continuous scores, our approach is able to optimize explicit label predictions. Note that optimality results for continuous scores need not trivially extend to optimality results for discrete label decisions.

Table 1. Examples of TP monotonic losses.

LOSS	DEFINITION	$\Phi(u, v, p)$
AM	$1 - \frac{\widehat{\text{TPR}}+\widehat{\text{TNR}}}{2}$	$1 - \frac{u+p(1-v-p)}{p(1-p)}$
F_β	$1 - \frac{1+\beta^2}{\frac{\beta^2}{\widehat{\text{Prec}}}+\frac{1}{\widehat{\text{TPR}}}}$	$1 - \frac{(1+\beta^2)u}{\beta^2 p+v}$
Jaccard	$\frac{\widehat{\text{FP}}+\widehat{\text{FN}}}{\widehat{\text{TP}}+\widehat{\text{FP}}+\widehat{\text{FN}}}$	$\frac{p+v-2u}{p+v-u}$
G-TP/PR	$1 - \sqrt{\widehat{\text{TPR}}.\widehat{\text{Prec}}}$	$1 - \frac{u}{\sqrt{p.v}}$
G-Mean	$1 - \sqrt{\widehat{\text{TPR}}.\widehat{\text{TNR}}}$	$1 - \frac{u(1-v-p+u)}{p(1-p)}$
H-Mean	$1 - 2/\left(\frac{1}{\widehat{\text{TPR}}} + \frac{1}{\widehat{\text{TNR}}}\right)$	$1 - \frac{2u(1-v-p+u)}{(1-v-p)p+u}$
AUC	$\frac{\widehat{\text{FP}}\ .\ \widehat{\text{FN}}}{(\widehat{\text{TP}}+\widehat{\text{FN}})(\widehat{\text{FP}}+\widehat{\text{TN}})}$	$\frac{(v-u)(p-u)}{p(1-p)}$

4. Algorithms

For multiclass classification, Theorem 2 immediately suggests $O(n^2)$ algorithm for computing the optimal solution when $\mathbb{P}(Y = C_i|x)$ is known — for each $k = 1, 2, \ldots, n$, evaluate the expected loss in selecting the top k classes (classes are sorted by $\mathbb{P}(Y = C_i|x)$), which can be done in $O(n)$ time. However, in binary and multilabel learning scenarios, even when $\mathbb{P}(Y_i = 1|\mathbf{x})$ is known exactly, characterization of optimality in Theorem 1 falls short in practice — it is not obvious how to compute the expectation in (1) without exhaustively enumerating 2^n possible $\mathbf{y}$ vectors. In this section, we present efficient algorithms for computing estimators for optimal predictions given $\mathbf{x}$ and a TP monotonic loss function L. We also prove the consistency of the proposed algorithms.

4.1. Computing Optimal in Practice

We observe that by evaluating the loss L through the function Φ (as defined in Proposition 1), we can generalize the approach suggested by Ye et al. (2012) for F_β-measure to obtain a template algorithm for optimizing any TP monotonic L. Consider the vector $\mathbf{s} \in \{0, 1\}^n$ with the top k values set to 1 and the rest to 0, and let $S_{i:j} := \sum_{l=i}^{j} y_l$. Note that for any $\mathbf{y} \in \{0, 1\}^n$ that satisfies $S_{1:k} = k_1$ and $S_{k+1:n} = k_2$, $L(\mathbf{s}, \mathbf{y})$ can simply be evaluated as $\Phi(\frac{1}{n}k_1, \frac{1}{n}k, \frac{1}{n}(k_1 + k_2))$. Thus $\sum_{\mathbf{y}\in\{0,1\}^n} \mathbb{P}(\mathbf{y}|\mathbf{x})L(\mathbf{s}, \mathbf{y})$ can be evaluated as a sum over possible values of k_1 and k_2, where the expectation is computed wrt. $P(S_{1:k} = k_1)P(S_{k+1:n} = k_2)$ with $0 \le k_1 \le k$ and $0 \le k_2 \le n - k$. Now, it remains to compute $P(S_{1:k} = k_1)$ and $P(S_{k+1:n} = k_2)$ efficiently.

Let $\eta_i = \mathbb{P}(Y_i = 1|\mathbf{x})$. A consistent estimate of this quantity may be obtained by minimizing a strongly proper loss function such as logistic loss (Reid and Williamson, 2009). Using the conditional independence assumption, we can show that $P(S_{1:k} = k_1)$ and $P(S_{k+1:n} = k_2)$ are the coefficients of z^{k_1} and z^{k_2} in $\Pi_{j=1}^{k}[\eta_j z + (1 - \eta_j)]$ and $\Pi_{j=k+1}^{n}[\eta_j z + (1 - \eta_j)]$ respectively, each of which can be computed in time $O(n^2)$ for fixed k. Note that the loss L itself can be evaluated in constant time. Let L_k denote the expected loss wrt. the estimated conditionals η_i, corresponding to setting the predictions of indices with top k highest conditional probabilities to 1. The resulting algorithm is presented in Algorithm 1. Though the computational complexity of Algorithm 1 is $O(n^3)$, we find that in practice it suffices to run the outer iterations until k^* where k^* is the first k s.t. $L_k \leq L_{k+1}$ (or $k^* = n$ if no such k exists), because for all $k > k^*$, $L_k \geq L_{k^*}$ holds. It is not obvious if this is a property enjoyed by all TP monotonic loss functions under conditional independence, but it would be interesting to theoretically establish this. The improvement in runtime is significant as multivariate losses are typically employed in scenarios where there is heavy imbalance in the distribution (in case of binary classification) or typically a small set of labels are relevant for a given instance (in case of multilabel learning), so $k^* \ll n$, for large n. Finally, we note that for a sub-family of fractional-linear losses studied by Koyejo et al. (2014; 2015), we can get a faster algorithm that runs in time $O(n^2)$ using the trick by Ye et al. (2012). Due to limited space, the resulting algorithm is presented in Appendix B.1.

Remark on multinomial distributions. For multiclass classification, note that computing the expectation wrt. the multinomial distribution is straight-forward. We simply replace steps 4-6 of Algorithm 1 with the following step:

$$L_k \leftarrow \sum_{j \leq k} \Phi\left(\frac{1}{n}, \frac{k}{n}, \frac{1}{n}\right) . \eta_j + \sum_{k < j \leq n} \Phi\left(0, \frac{k}{n}, \frac{1}{n}\right) . \eta_j$$

4.2. Consistency Analysis

Consider a procedure that minimizes (1) computed with respect to a consistent estimate $\hat{\mathbb{P}}(Y_i|\mathbf{x})$ of the conditional $\mathbb{P}(Y_i|\mathbf{x})$. Here, we show that any such procedure is consistent.

Theorem 3. *Let* $\eta_i = \mathbb{P}(Y_i = 1|\mathbf{x})$, *and assume the estimate* $\hat{\eta}_i$ *satisfies* $\hat{\eta}_i \xrightarrow{p} \eta_i$. *Given a bounded loss function* L *and a set of instances* $\mathbf{x}$, *let* $\mathbf{s}^* = \arg\min_{\mathbf{s} \in \{0,1\}^n} \mathbf{E}_{\mathbf{Y} \sim \mathbb{P}(.|\mathbf{x})} L(\mathbf{s}, \mathbf{Y})$ *be the optimal prediction with respect to* $\mathbb{P}$ *and* $\hat{\mathbf{s}} = \arg\min_{\mathbf{s} \in \{0,1\}^n} \mathbf{E}_{\mathbf{Y} \sim \hat{\mathbb{P}}(.|\mathbf{x})} L(\mathbf{s}, \mathbf{Y})$ *be the optimal prediction with respect to the consistent estimate* $\hat{\mathbb{P}}$, *then*

$$\mathbf{E}_{\mathbf{Y} \sim \mathbb{P}(.|\mathbf{x})}\left[L(\hat{\mathbf{s}}, \mathbf{Y}) - L(\mathbf{s}^*, \mathbf{Y})\right] \xrightarrow{p} 0.$$

Our consistency result also applies to previous algorithms proposed for F_β in different settings e.g. by Lewis (1995); Chai (2005); Jansche (2007); Coz Velasco et al. (2009); Ye et al. (2012), where analysis of consistency with empirical probability estimates was lacking. For TP monotonic multivariate losses, it is immediate from Theorem 3 that Algorithm 1 is consistent.

Algorithm 1 Computing $\mathbf{s}^*$ for TP Monotonic L

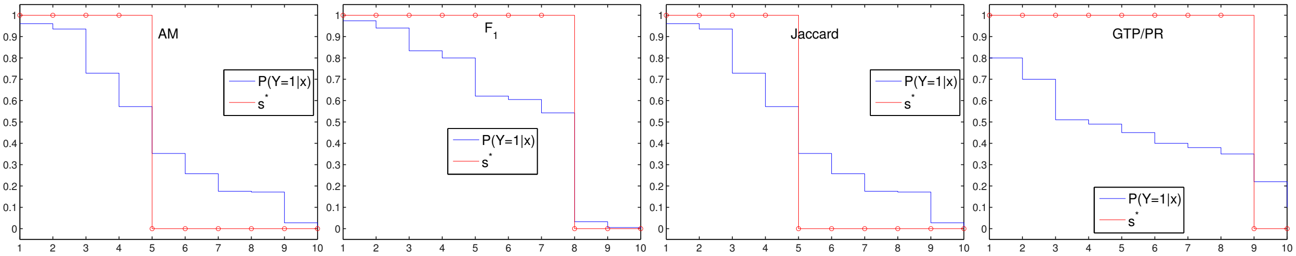
```
1: **Input:** $L$ and estimates of $\eta_i = \mathbb{P}(Y_i = 1|\mathbf{x})$, $i = 1, 2, \ldots, n$ sorted wrt. $\eta_i$
2: Init $s_i^* = 0, \forall i \in [n]$.
3: **for** k = $1, 2, \ldots, n$ **do**
4:     For $0 \leq i \leq k$, set $C_k[i]$ as the coefficient of $z^i$ in $\Pi_{i=1}^{k}\big(\eta_i z + (1 - \eta_i)\big)$.
5:     For $0 \leq i \leq n - k$, set $D_k[i]$ as the coefficient of $z^i$ in $\Pi_{i=k+1}^{n}\big(\eta_i z + (1 - \eta_i)\big)$.
6:     $L_k \leftarrow \sum_{\substack{0 \leq k_1 \leq k \\ 0 \leq k_2 \leq n-k}} C_k[k_1] D_k[k_2] \Phi(\frac{k_1}{n}, \frac{k}{n}, \frac{1}{n}(k_1 + k_2))$.
7: **end for**
8: $k^* \leftarrow \arg\min_k L_k$.
9: return $\mathbf{s}^*$ s.t. $s_i^* \leftarrow 1$ for $i \in [k^*]$.
```

5. Experiments

We present two sets of experiments. The first is an experimental validation on synthetic data with known ground truth probabilities. The results serve to verify our main result (Theorem 1) for some of the losses in Table 1. The second set is an experimental evaluation of the proposed algorithm for computing optimal prediction on benchmark datasets, with comparisons to baseline and state-of-the-art algorithms for classification with general losses.

5.1. Synthetic data: Verification of the theory

We consider four losses from Table 1 based on the performance metrics AM, Jaccard, F_1 (harmonic mean of Precision and Recall) and G-TP/PR (geometric mean of Precision and Recall). To simulate, we sample a set of ten 2-dimensional vectors $\mathbf{x} = \{x_1, x_2, \ldots, x_{10}\}$ from the standard Gaussian. The conditional probability is modeled using a sigmoid function: $\eta_i = \mathbb{P}(Y = 1|x_i) = \frac{1}{1+\exp(-w^T x_i)}$, for a random vector w also sampled from the standard Gaussian. The optimal predictions $\mathbf{s}^*$ (1) are then obtained by exhaustive search over the 2^{10} possible label vectors. For each loss, we plot the conditional probabilities (in decreasing order) and $\mathbf{s}^*$ in Figure 1. We observe that it is optimal to assign positive labels to top k^* instances with highest $\mathbb{P}(Y|x)$, as given in Theorem 1, where k^* depends on the instances, the distribution and the loss itself.



(a) AM (b) F_1 (c) Jaccard (d) G-TP/PR

Figure 1. Optimal predictions for losses from Table 1 demonstrated on synthetic data. In each case, we verify that $\mathbf{s}^*$ conforms to the ordering of $\mathbb{P}(Y_i|\mathbf{x})$ as stated in Theorem 1.

DATASET	T	Proposed F_1	Baseline F_1	Tuning δ F_1	S-SVM F_1	Proposed Jaccard	Baseline Jaccard	Tuning δ Jaccard	S-SVM Jaccard
REUTERS (65)	1	0.4125	0.4849	0.5020	**0.3254**	0.5239	0.5692	0.5691	**0.4977**
	10	**0.1753**	0.2376	0.2401	0.1604	0.3199	0.3591	**0.3090**	**0.2983**
	50	**0.1003**	0.1572	0.1490	0.1533	**0.2485**	0.2552	**0.2422**	0.2528
	100	**0.0144**	0.0325	0.0331	0.0360	**0.0602**	**0.0625**	0.0643	0.0689
LETTERS (26)	1	**0.2890**	0.5173	0.4255	0.4395	0.5728	0.6368	0.5682	**0.5029**
SCENE (6)	1	**0.3084**	0.8023	0.5603	0.3937	0.6460	0.9794	0.7920	**0.5541**
WEB PAGE	1	**0.1606**	0.3145	0.3191	0.4700	0.5363	**0.4785**	0.4806	0.5758
SPAMBASE	1	0.1552	0.1202	**0.1161**	0.2263	0.2686	0.2133	**0.1997**	0.3461
IMAGE	1	0.1458	0.1429	0.1419	**0.1227**	0.2545	02500	0.2377	**0.2235**
BREAST CANCER	1	**0.0207**	0.0411	**0.0234**	**0.0270**	0.0658	0.0789	**0.0519**	**0.0526**

Table 2. Comparison of methods: Fractional-linear losses, F_1 and Jaccard defined in Table 1. Reported values correspond to losses on test data (lower values are better). Baseline refers to thresholding $\hat{\eta}(x)$ at 0.5; 'Tuning δ' refers to the plugin-estimator combined with threshold selection method in Koyejo et al. (2014); S-SVM refers to the structured SVM method proposed by Joachims (2005). First three are multiclass datasets (number of classes indicated in parenthesis).

5.2. Benchmark data: Proposed algorithms

We perform classification using the proposed approach: (i) obtain a model for the conditional distribution $\eta(x) = \mathbb{P}(Y = 1|x)$ using training data and (ii) compute $\mathbf{s}^*$ for the test data using estimated conditionals in the proposed Algorithm 1 (and in the faster Algorithm 2 in the Appendix, when applicable). We use logistic loss on the training samples (with L_2 regularization) to obtain an estimate $\hat{\eta}(x)$ of $\mathbb{P}(Y = 1|x)$. In our experiments, we consider losses based on four performance metrics AM, F_1, Jaccard and G-TP/PR. For AM and G-TP/PR we use Algorithm 1, while for the fractional-linear losses Jaccard and F_1 we use the more efficient Algorithm 2. We report the loss on the test data.

We compare our approach with that of structured hinge loss minimization for multivariate losses (Joachims, 2005). We use the MATLAB wrapper provided by Vedaldi (2011), which internally uses the fast `svm-struct` implementation of (Joachims, 2005). The solver is based on cutting-plane method and for each of the aforementioned losses, we can implement the constraint generation step efficiently. We also compare with that of the plugin-estimator combined with threshold selection (which we refer to as 'Tuning δ') proposed by Koyejo et al. (2014) and Narasimhan et al. (2014), that minimizes the loss on the *population* confusion matrix (as discussed in Section 3.1). In this case, the optimal classifier is given by sign$(\hat{\eta}(x) - \delta^*)$. The training data is split into two sets, one set is used for estimating $\hat{\eta}(x)$ and the other for selecting the optimal δ for the loss function. The predictions are then made by thresholding $\hat{\eta}(x)$ of the test data points at δ. We also compare to the standard baseline method for binary classification — thresholding $\hat{\eta}(x)$ at 1/2.

We report results on seven benchmark datasets (used in (Koyejo et al., 2014; Ye et al., 2012)). (1) REUTERS, consisting of 8293 news articles categorized into 65 topics. We present results for averaging over topics with at least T positives in the training (5946 articles) as well as the test (2347 articles) data; (2) LETTERS dataset consisting of 20000 handwritten characters (16000 training and 4000 test instances) categorized into 26 letters; (3) SCENE (a UCI benchmark dataset) consisting of 2230 images (1137 training and 1093 test instances) categorized into 6 scene types; (4) WEBPAGE binary dataset, consisting of 34780 web pages (6956 train and 27824 test); highly imbalanced, with only about 182 positive instances in the

Dataset	T	Proposed AM	Baseline AM	Tuning δ AM	S-SVM AM	Proposed G-TP/PR	Baseline G-TP/PR	Tuning δ G-TP/PR	S-SVM G-TP/PR
Reuters (65)	1	0.1166	0.2777	0.2267	**0.0598**	**0.2711**	0.4553	0.4782	**0.2700**
	10	**0.0480**	0.1640	0.0889	**0.0368**	0.1934	0.2200	0.1924	**0.1579**
	50	**0.0341**	0.0983	0.0418	**0.0374**	0.1505	0.1559	**0.1309**	0.1527
	100	**0.0217**	0.0239	**0.0219**	0.0235	**0.0313**	**0.0325**	0.0328	0.0369
Letters (26)	1	**0.1285**	0.2980	**0.1280**	0.1800	0.4213	0.4936	0.4098	**0.3465**
Scene (6)	1	0.4160	0.4935	0.4190	**0.2250**	0.4931	0.9395	0.6152	**0.3935**
Web page	1	**0.1311**	0.1795	**0.1250**	0.1528	0.3383	**0.3133**	**0.3114**	0.4384
Spambase	1	0.1220	**0.0990**	**0.0910**	0.1650	0.1506	0.1169	**0.1087**	0.2204
Image	1	0.1959	0.1808	0.1931	**0.1599**	0.1324	0.1423	0.1298	**0.1198**
Breast Cancer	1	0.0204	0.0399	**0.0170**	0.0205	0.0340	0.0410	**0.0266**	**0.0270**

Table 3. Comparison of methods: AM and G-TP/PR losses defined in Table 1. Reported values correspond to losses on test data (lower values are better). Baseline refers to thresholding $\hat{\eta}(x)$ at 0.5; 'Tuning δ' refers to the plugin-estimator combined with threshold selection method in Narasimhan et al. (2014). First three are multiclass datasets (number of classes indicated in parenthesis).

train; (5) Image, with 1300 train and 1010 test images; (6) Breast Cancer, with 463 train and 220 test instances, and (7) Spambase with 3071 train and 1530 test instances. See (Koyejo et al., 2014; Ye et al., 2012) for more details on the datasets. Note that in case of the multiclass datasets, we report results (using *one-versus-all* classifiers) averaged over classes.

The results for F_1 and Jaccard losses are presented in Table 2. Note that the reported values correspond to losses — smaller values are better. We find that our proposed algorithm almost always achieves the least F_1 loss compared to other methods. However, in case of the Jaccard loss, tuning a threshold on training data seems to perform better though it is theoretically not optimal. The structured SVM method, though not known to have strong consistency properties, is competitive here. The results for AM and G-TP/PR losses are presented in Table 3. We find that the proposed approach is competitive across many datasets for the AM loss. Again, we find that the structured SVM method is competitive in case of the G-TP/PR loss. Perhaps, structured loss minimization may hold promise for certain data distributions and it is an interesting future research direction.

6. Conclusions

We study optimizing expected multivariate losses used in several prediction tasks. Our analysis shows that for losses that satisfy an intuitive monotonicity property, optimal predictions can be computed given the knowledge of the conditional probability of the positive class. We propose efficient and consistent estimators for computing optimal predictions in practice. Our results are complementary to some of the recent advances in the understanding of optimal classification (Koyejo et al., 2014; Narasimhan et al., 2014) and a leap forward in the direction initiated two decades ago by Lewis (1995).

Acknowledgments: We acknowledge the support of NSF via CCF-1320746 and IIS-1546459, and NIH via R01 GM117594-01 as part of the Joint DMS/NIGMS Initiative to Support Research at the Interface of the Biological and Mathematical Sciences.

References

Kian Ming Adam Chai. 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*, pages 593–594. ACM, 2005.

Juan José del Coz Velasco, Jorge Díez Peláez, and Antonio Bahamonde Rionda. Learning nondeterministic classifiers. *Journal of Machine Learning Research, 10*, 2009.

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

Krzysztof Dembczynski, Arkadiusz Jachnik, Wojciech Kotlowski, Willem Waegeman, and Eyke Hüllermeier. Optimizing the F-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. In *Proceedings of The 30th Intl. Conf. on Machine Learning*, pages 1130–1138, 2013.

Krzysztof J Dembczynski, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. An exact algorithm for F-measure maximization. In *Advances in Neural Information Processing Systems*, pages 1404–1412, 2011.

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.

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.

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

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

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

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

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

Oluwasanmi O Koyejo, Nagarajan Natarajan, Pradeep K Ravikumar, and Inderjit S Dhillon. Consistent multilabel classification. In *Advances in Neural Information Processing Systems*, pages 3303–3311, 2015.

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

David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual Intl. ACM SIGIR Conf.*, pages 3–12. Springer-Verlag New York, Inc., 1994.

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

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

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

Shameem Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet. Optimizing F-measures by cost-sensitive classification. In *Advances in Neural Information Processing Systems*, pages 2123–2131, 2014.

James Petterson and Tibério S Caetano. Submodular multilabel learning. In *Advances in Neural Information Processing Systems*, pages 1512–1520, 2011.

José Ramón Quevedo, Oscar Luaces, and Antonio Bahamonde. Multilabel classifiers with a probabilistic thresholding strategy. *Pattern Recognition*, 45(2):876–883, 2012.

Mark D Reid and Robert C Williamson. Surrogate regret bounds for proper losses. In *Proceedings of the 26th Intl. Conf. on Machine Learning*, pages 897–904. ACM, 2009.

A. Vedaldi. A MATLAB wrapper of SVM$^{\text{struct}}$. `http://www.vlfeat.org/~vedaldi/code/svm-struct-matlab`, 2011.

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

Hong Wang, Wei Xing, Kaiser Asif, and Brian Ziebart. Adversarial prediction games for multivariate losses. In *Advances in Neural Information Processing Systems*, pages 2710–2718, 2015.

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 Intl. Conf. on Machine Learning*, 2012.

A. Appendix A

A.1. Proof of Theorem 1

The proof is by contradiction. Fix a distribution $\mathbb{P}$ satisfying conditional independence, and let $\mathbf{x}$ denote a fixed set of instances. Denote $\mathbb{P}(Y = 1|x_i) = \eta_i$ and the optimal classifier by $\mathbf{s}^* \in \{0,1\}^n$. Suppose there exist indices j, k such that $s_j^* = 1, s_k^* = 0$ and $\eta_j < \eta_k$. Let $\mathbf{s}' \in \{0,1\}^n$ be such that $s_j' = 0$ and $s_k' = 1$, but identical to $\mathbf{s}^*$ otherwise i.e. $s_i^* = s_i' \ \forall i \in [n]\backslash\{j,k\}$. Note that $\sum_{i=1}^n s_i^* = \sum_{i=1}^n s_i^{'}$. For convenience, define:

$$\mathcal{U}^L(\mathbf{s}; \mathbb{P}) := \mathbf{E}_{\mathbf{Y}\sim\mathbb{P}(.|\mathbf{x})} L(\mathbf{s}, \mathbf{Y}) \ .$$

By optimality of $\mathbf{s}^*$,

$$\mathcal{U}^L(\mathbf{s}^*; \mathbb{P}) - \mathcal{U}^L(\mathbf{s}'; \mathbb{P}) \leq 0. \tag{6}$$

Consider the LHS, $\mathcal{U}^L(\mathbf{s}^*; \mathbb{P}) - \mathcal{U}^L(\mathbf{s}'; \mathbb{P})$ is equal to:

$$\sum_{\mathbf{y}\in\{0,1\}^n} P(\mathbf{y}|\mathbf{x})[L(\mathbf{s}^*, \mathbf{y}) - L(\mathbf{s}', \mathbf{y})] = \\ \sum_{\mathbf{y}\in\{0,1\}^n : y_j \neq y_k} P(\mathbf{y}|\mathbf{x})[L(\mathbf{s}^*, \mathbf{y}) - L(\mathbf{s}', \mathbf{y})] \\ + \sum_{\mathbf{y}\in\{0,1\}^n : y_j = y_k} P(\mathbf{y}|\mathbf{x}) \underbrace{[L(\mathbf{s}^*, \mathbf{y}) - L(\mathbf{s}', \mathbf{y})]}_{(*)}$$

Note that when $y_j = y_k$, $\sum_{i=1}^n s_i^* y_i = \sum_{i=1}^n s_i' y_i$, so $L(\mathbf{s}^*, \mathbf{y}) - L(\mathbf{s}', \mathbf{y}) = 0$. It follows that the term $(*)$ equals 0.

Next we apply the representation of Proposition 1 with $v(\mathbf{s}) = \frac{1}{n}\sum_{i=1}^n s_i$ and $p(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^n y_i$. Let $z \in \{0,1\}^{n-2}$ denote the vector corresponding to $n-2$ indices $\{y_i, \ i \in [n] \setminus \{j,k\}\}$, then $\mathcal{U}^L(\mathbf{s}^*; \mathbb{P}) - \mathcal{U}^L(\mathbf{s}'; \mathbb{P})$ is given by:

$$\sum_{\mathbf{y}\in\{0,1\}^n : y_j \neq y_k} \mathbb{P}(\mathbf{y}|\mathbf{x})[L(\mathbf{s}^*, \mathbf{y}) - L(\mathbf{s}', \mathbf{y})] = \\ \sum_{\mathbf{z}\in\{0,1\}^{n-2}} \mathbb{P}(\mathbf{z}, y_j = 1, y_k = 0|\mathbf{x}) \big[\Phi(\widehat{\text{TP}}(\mathbf{s}^*, \mathbf{y}), v(\mathbf{s}^*), p(\mathbf{y})) \\ - \Phi(\widehat{\text{TP}}(\mathbf{s}', \mathbf{y}), v(\mathbf{s}'), p(\mathbf{y}))\big] \\ + \mathbb{P}(\mathbf{z}, y_j = 0, y_k = 1|\mathbf{x}) \big[\Phi(\widehat{\text{TP}}(\mathbf{s}^*, \mathbf{y}), v(\mathbf{s}^*), p(\mathbf{y})) \\ - \Phi(\widehat{\text{TP}}(\mathbf{s}', \mathbf{y}), v(\mathbf{s}'), p(\mathbf{y}))\big]$$

Let $\tilde{\mathbf{s}} = \{s_i^* \ \forall i \in [n] \setminus \{j,k\}\}$ and define $\#TP(\mathbf{z}) := \sum_i \tilde{\mathbf{s}}_i z_i$ and $\#p(\mathbf{z}) = z_i$ (where the # prefix indicates counts rather than normalized values), and note that $v(\mathbf{s}^*) = v(\mathbf{s}')$. With these substitutions, $\mathcal{U}^L(\mathbf{s}^*; \mathbb{P}) - \mathcal{U}^L(\mathbf{s}'; \mathbb{P})$ is given by:

$$\sum_{\mathbf{z}\in\{0,1\}^{n-2}} \mathbb{P}(\mathbf{z}, y_j = 1, y_k = 0|\mathbf{x}) \\ \left[\Phi\left(\frac{1}{n}(\#TP(\mathbf{z}) + 1), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z}) + 1)\right)\right. \\ \left. - \Phi\left(\frac{1}{n}\#TP(\mathbf{z}), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z}) + 1)\right)\right] \\ + \mathbb{P}(\mathbf{z}, y_j = 0, y_k = 1|\mathbf{x}) \left[\Phi\left(\frac{1}{n}\#TP(\mathbf{z}), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z}) + 1)\right)\right. \\ \left. -\Phi(\frac{1}{n}(\#TP(\mathbf{z}) + 1), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z}) + 1)\Big)\right]$$

By conditional independence, we have that $P(\mathbf{z}, y_j, y_k|\mathbf{x}) = P(\mathbf{z}|\mathbf{x})P(y_j|\mathbf{x})P(y_k|\mathbf{x})$, so that the equation further simplifies to:

$$\sum_{\mathbf{z}\in\{0,1\}^{n-2}} \mathbb{P}(\mathbf{z}|\mathbf{x}) \left[\Phi\left(\frac{1}{n}(\#TP(\mathbf{z}) + 1), v(\mathbf{s}'),\right.\right. \\ \left.\left.\frac{1}{n}(\#p(\mathbf{z}) + 1)\right) - \Phi\left(\frac{1}{n}\#TP(\mathbf{z}), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z}) + 1)\right)\right] \\ \Big[\eta_j(1 - \eta_k) - \eta_k(1 - \eta_j)\Big] = \\ (\eta_j - \eta_k) \sum_{\mathbf{z}\in\{0,1\}^{n-2}} \mathbb{P}(\mathbf{z}|\mathbf{x}) \left[\Phi\left(\frac{1}{n}(\#TP(\mathbf{z}) + 1), v(\mathbf{s}'),\right.\right. \\ \left.\left.\frac{1}{n}(\#p(\mathbf{z}) + 1)\right) - \Phi\left(\frac{1}{n}\#TP(\mathbf{z}), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z}) + 1)\right)\right]$$

Note that for each $\mathbf{z} \in \{0,1\}^{n-2}$:

- $\Phi\left(\frac{1}{n}(\#TP(\mathbf{z}) + 1), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z}) + 1)\right)$ can be interpreted as L computed on the vectors $\mathbf{y} \in \mathbb{R}^n$ defined as $\{y_i = z_i \ \forall \ i \in [n] \setminus \{j,k\}\} \cup \{y_j = 1\} \cup \{y_k = 0\}$, and $\mathbf{s}^* \in \mathbb{R}^n$ (which is the assumed optimal).

- $\Phi\left(\frac{1}{n}\#TP(\mathbf{z}), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z}) + 1)\right)$ can be interpreted as L computed on the vectors $\mathbf{y} \in \mathbb{R}^n$ defined as above and $\mathbf{s}' \in \mathbb{R}^n$.

By TP monotonicity of L, for each $\mathbf{z}$, the difference term

$$\Phi\left(\frac{1}{n}(\#TP(\mathbf{z}) + 1), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z}) + 1)\right) \\ -\Phi\left(\frac{1}{n}\#TP(\mathbf{z}), v(\mathbf{s}'), \frac{1}{n}(\#p(\mathbf{z}) + 1)\right) < 0.$$

This combined with (6) implies that $\eta_j - \eta_k \geq 0$ which is a contradiction.

A.2. Proof of Theorem 2

Fix a multinomial distribution $\mathbb{P}$, and instance $\mathbf{x}$. Let the classes $C_1, C_2, \ldots, C_n$ be indexed according to the descending order of $\eta_i := \mathbb{P}(Y = C_i|\mathbf{x})$. First, observe that it suffices to show that for any fixed $0 \leq k \leq n$, the optimal solution denoted by $\mathbf{s}^*(k)$ that minimizes the expected loss restricted to subset of vectors $\mathcal{S}_k = \{\mathbf{s} \in \{0,1\}^n \mid \sum_{i=1}^n s_i = k\}$ satisfies $s_1^*(k) = s_2^*(k) = \cdots = s_k^*(k) = 1$, and $s_{k+1}^*(k) = \cdots = s_n^*(k) = 0$. Define $[[P]] = 1$ if the predicate P is true or 0 otherwise. Now, for any $\mathbf{s} \in \mathcal{S}_k$, we have,

$$\mathbf{E}_{Y\sim\mathbb{P}(.|\mathbf{x})}[L(\mathbf{s},Y)] = \sum_{i\in[n]} \Phi\left(\frac{1}{n}[[s_i = 1]], \frac{k}{n}, \frac{1}{n}\right)\eta_i$$

$$= \sum_{i:s_i=1} \Phi\left(\frac{1}{n}, \frac{k}{n}, \frac{1}{n}\right)\eta_i + \sum_{i:s_i=0} \Phi\left(0, \frac{k}{n}, \frac{1}{n}\right)\eta_i$$

$$= \Phi\left(\frac{1}{n}, \frac{k}{n}, \frac{1}{n}\right) \sum_{i:s_i=1} \eta_i + \Phi\left(0, \frac{k}{n}, \frac{1}{n}\right)\left(1 - \sum_{i:s_i=1} \eta_i\right)$$

$$= \left(\Phi\left(\frac{1}{n}, \frac{k}{n}, \frac{1}{n}\right) - \Phi\left(0, \frac{k}{n}, \frac{1}{n}\right)\right) \sum_{i:s_i=1} \eta_i + \Phi\left(0, \frac{k}{n}, \frac{1}{n}\right)$$

By TP monotonicity of L, we have,

$$\Phi\left(\frac{1}{n}, \frac{k}{n}, \frac{1}{n}\right) < \Phi\left(0, \frac{k}{n}, \frac{1}{n}\right).$$

So, to minimize the RHS of the above set of equations, we need to maximize $\sum_{i:s_i=1} \eta_i$. Restricting to $\mathcal{S}_k$, the sum is maximized when we choose classes with k largest η_i values. We conclude that $\mathbf{s}^*(k)$ is the minimizer. This completes the proof.

A.3. Sufficiency of TP Monotonicity

TP monotonicity of L is sufficient but not necessary for the optimality characterization we show in the paper. For instance, consider the subclass of losses where $\Phi(\cdot, v, p)$ is independent of the first argument i.e. independent of $\widehat{\text{TP}}$. SEC is an example of a loss in this family with $\Phi_{\text{SEC}}(\widehat{\text{TP}}, v, p) = 2 - v - p$. But then, it is straight-forward to characterize optimal solution for such losses:

Proposition 5. *Let $L = \Phi(\widehat{TP}, v, p)$ be a loss independent of $\widehat{TP}$, then the optimal* (1) *under L satisfies the ordering of marginal probabilities as in Theorem 1.*

Proof. Suppose $\Phi(\cdot, v, p)$ is independent of its first argument. Let $\mathbf{s}^*$ be an optimal classifier, with $v^* = v(\mathbf{s}^*)$. If $\mathbf{s}^*$ does not already satisfy the property, then simply sort $\mathbf{s}^*$ with respect to $\mathbb{P}(Y_i|\mathbf{x})$ to obtain a new classifier $\tilde{\mathbf{s}}$. Clearly, $v(\tilde{\mathbf{s}}) = v^*$, and $\Phi(\cdot, v(\tilde{\mathbf{s}}^*), p) = \Phi(\cdot, v^*, p)$. □

A.4. Proof of Proposition 3

Suppose L satisfies TPR/TNR monotonicity. Let $u_1 = \text{TP}(\mathbf{s}_1, \mathbf{y}_1)$ and $u_2 = \text{TP}(\mathbf{s}_2, \mathbf{y}_2)$, $v = v(\mathbf{s}_1) = v(\mathbf{s}_2)$ and $p = p(\mathbf{y}_1) = p(\mathbf{y}_2)$. Note that $\Phi(u_1, v, p) = \Gamma(\frac{u_1}{p}, \frac{1-v-p+u_1}{1-p}, p)$ (and similarly equality holds for $\Phi(u_2, v, p)$). Now, whenever $u_1 = \widehat{\text{TP}}(\mathbf{s}_1, \mathbf{y}_1) > \widehat{\text{TP}}(\mathbf{s}_2, \mathbf{y}_2) = u_2$, $v(\mathbf{s}_1) = v(\mathbf{s}_2) = v$, and $p(\mathbf{y}_1) = p(\mathbf{y}_2) = p$, we have $\widehat{\text{TPR}}(\mathbf{s}_1, \mathbf{y}_1) > \widehat{\text{TPR}}(\mathbf{s}_2, \mathbf{y}_2), \widehat{\text{TNR}}(\mathbf{s}_1, \mathbf{y}_1) > \widehat{\text{TNR}}(\mathbf{s}_2, \mathbf{y}_2)$, and

$$\begin{aligned}
\Phi(u_1, v, p) &= \Gamma(\frac{u_1}{p}, \frac{1-v-p+u_1}{1-p}, p) \\
&= \Gamma(\widehat{\text{TPR}}(\mathbf{s}_1, \mathbf{y}_1), \widehat{\text{TNR}}(\mathbf{s}_1, \mathbf{y}_1), p) \\
&\overset{(*)}{<} \Gamma(\widehat{\text{TPR}}(\mathbf{s}_2, \mathbf{y}_2), \widehat{\text{TNR}}(\mathbf{s}_2, \mathbf{y}_2), p) \\
&= \Gamma(u_2.p, \frac{1-v-p+u_2}{1-p}, p) \\
&= \Phi(u_2, v, p)
\end{aligned}$$

where $(*)$ follows from TPR/TNR monotonicity of L. Thus L satisfies TP monotonicity.

B. Appendix B

B.1. Faster Algorithm for Fractional-Linear Losses

We focus our attention on the fractional-linear family of losses studied by Koyejo et al. (2014; 2015). A fractional-linear loss can be represented by Φ_{FL} as given in (4). As shown in Proposition 2, L_{FL} satisfies TP monotonicity when $c_1 < d_1$. When $c_3 = 0$ and the constants $\{d_0, d_1, d_2, d_3\}$ are rational in (4), we can get a quadratic-time procedure for computing $\mathbf{s}^*$ appealing to the method proposed by Ye et al. (2012). Formally, we consider the sub-family of TP monotonic fractional-linear losses:

$$\{L_{SFL} : \Phi_{FL}(u, v, p) = \frac{c_0 + c_1 u + c_2 v}{d_0 + d_1 u + d_2 v + d_3 p}, \; c_1 < d_1, \text{ and } d_0, d_1, d_2, d_3 \text{ are rational}\}, \quad (7)$$

which includes the loss based on Jaccard measure and others not studied by Ye et al. (2012). Consider Step 6 of Algorithm 1 for a loss in family (7):

$$L_k \leftarrow \sum_{0\leq k_1\leq k} C[k_1](c_0 n + c_1 k_1 + c_2 k) \; .$$

$$\sum_{0\leq k_2\leq n-k} D[k_2]/(d_0 n + (d_1 + d_3)k_1 + d_2 k + d_3 k_2).$$

Define $b(k, \alpha) = \sum_{0\leq k_2\leq n-k} D_k[k_2]/(\alpha + d_3 k_2)$. Verify that $b(n, \alpha) = 1/\alpha$. From the fact that $D_{k-1}[i] = \eta_k D_k[i-1] + (1-\eta_k) D_k[i]$, it follows that:

$$b(k-1, \alpha) = \eta_k b(k, \alpha + d_3) + (1 - p_k) b(k, \alpha).$$

Now, when d_i's are rational, i.e. $d_i = q_i/r_i$, the above induction can be implemented using an array to store the values of b, for possible values of α.

Algorithm 2 Computing $\mathbf{s}^*$ for L_{SFL} in the family (7)

```
1: Input: Estimates of $\eta_i = \mathbb{P}(Y_i = 1|\mathbf{x}), i = 1, 2, \ldots, n$
   sorted wrt. $\eta_i$, and $c_0, c_1, c_2, d_i = q_i/r_i, i = 0, 1, 2, 3$
   corresponding to $L_{\text{SFL}}$
2: Init $s_i^* = 0, \forall i \in [n]$.
3: Set $j_0 \leftarrow r_1 r_2 r_3 q_0,\ j_{u,1} \leftarrow r_0 r_2 r_3 q_1,\ j_{u,2} \leftarrow r_0 r_1 r_2 q_3,\ j_v \leftarrow r_0 r_1 r_3 q_2$
4: for $1 \le i \le (|j_{u,1}| + |j_{u,2}| + |j_v|)n$ do
5:     set $S[i] \leftarrow r_0 r_1 r_2 r_3/(i + j_0 n)$.
6: end for
7: for $k = n$ to 1 do
8:     For $0 \le i \le k$, set $C_k[i]$ as the coefficient of $z^i$ in
       $\Pi_{i=1}^{k}\big(\eta_i z + (1 - \eta_i)\big)$.
9:     $L_{\text{SFL};k} \leftarrow \sum_{0 \le k_1 \le k} (c_0 n + c_1 k_1 + c_2 k) C_k[k_1] S[(j_{u,1} + j_{u,2})k_1 + j_v k]$.
10:    for $i = 1$ to $(|j_{u,1}| + |j_{u,2}| + |j_v|)(k - 1)$ do
11:        $S[i] \leftarrow (1 - \eta_k) S[i] + \eta_k S[i + j_{u,2}]$.
12:    end for
13: end for
14: Set $k^* \leftarrow \arg\min_k L_{\text{SFL};k}$ and $s_i^* \leftarrow 1$ for $i \in [k^*]$.
15: return $\mathbf{s}^*$
```

Correctness of Algorithm 2: When $d_3 \neq 0$, at line 7 of Algorithm 2, we can verify that $S[i] = b(k, (i + j_0 n)d_3/j_{u,2})$, and therefore at line 9, $S[(j_{u,1} + j_{u,2})k_1 + j_v k] = b(k, (j_{u,1} + j_{u,2})k_1 + j_v k + j_0 n)d_3/j_{u,2}) = b(k, (d_1 + d_3)k_1 + d_2 k + d_0 n)$ as desired. When $d_3 = 0$, $b(k, \alpha) = b(k - 1, \alpha)$ for all $1 \le k \le n$. Let $q_3 = 0$ and $r_3 = 1$. Then, line 5 sets $S[i] = r_0 r_1 r_2/(i + j_0 n)$, line 11 maintains this invariant as $j_{u,2} = 0$ in this case, and therefore at line 9, $S[(j_{u,1} + j_{u,2})k_1 + j_v k] = 1/(d_1 k_1 + d_2 k + d_0 n)$ as desired.

B.2. Proof of Theorem 3

For convenience, define:

$$\mathcal{U}^L(\mathbf{s}; \mathbb{P}) := \mathbf{E}_{\mathbf{Y} \sim \mathbb{P}(.|\mathbf{x})} L(\mathbf{s}, \mathbf{Y}) \ .$$

Let $\mathcal{U}_*^L := \mathcal{U}^L(\mathbf{s}^*; \mathbb{P})$ and let $\widehat{\mathcal{U}}^L = \mathcal{U}^L(\hat{\mathbf{s}}; \mathbb{P})$. Also define the empirical distribution:

$$\hat{\mathbb{P}}(\mathbf{y}|\mathbf{x}) = \Pi_{i=1}^{n} \hat{\eta}_i^{y_i} (1 - \hat{\eta}_i)^{1 - y_i}.$$

Now consider:

$$\begin{aligned} \widehat{\mathcal{U}}^L - \mathcal{U}_*^L = &\quad \widehat{\mathcal{U}}^L + \mathcal{U}^L(\hat{\mathbf{s}}; \hat{\mathbb{P}}) - \mathcal{U}^L(\hat{\mathbf{s}}; \hat{\mathbb{P}}) - \mathcal{U}_*^L \\ \le &\quad \widehat{\mathcal{U}}^L + \mathcal{U}^L(\mathbf{s}^*; \hat{\mathbb{P}}) - \mathcal{U}^L(\hat{\mathbf{s}}; \hat{\mathbb{P}}) - \mathcal{U}_*^L \\ \le &\quad 2 \max_{\mathbf{s}} \big|\mathcal{U}^L(\mathbf{s}; \mathbb{P}) - \mathcal{U}^L(\mathbf{s}; \hat{\mathbb{P}})\big| \qquad (8) \end{aligned}$$

For any fixed $\mathbf{s} \in \{0, 1\}^n$, we have:

$$\begin{aligned} &\big|\mathcal{U}^L(\mathbf{s}; \mathbb{P}) - \mathcal{U}^L(\mathbf{s}; \hat{\mathbb{P}})\big| = \\ &\big| \sum_{y \in \{0,1\}^n} \hat{\mathbb{P}}(\mathbf{y}|\mathbf{x}) L(\mathbf{s}, \mathbf{y}) - \sum_{y \in \{0,1\}^n} \mathbb{P}(\mathbf{y}|\mathbf{x}) L(\mathbf{s}, \mathbf{y}) \big| \\ &\le \sum_{y \in \{0,1\}^n} \big|\hat{\mathbb{P}}(\mathbf{y}|\mathbf{x}) - \mathbb{P}(\mathbf{y}|\mathbf{x})\big| L(\mathbf{s}, \mathbf{y}) \qquad (9) \end{aligned}$$

Let $\eta(x)$ denote the empirical estimate obtained using m training samples. Now because $\hat{\eta}(x) \xrightarrow{p} \eta(x)$, we have that for sufficiently large set of training examples, $\hat{\mathbb{P}}(\mathbf{y}|\mathbf{x}) \xrightarrow{p} \mathbb{P}(\mathbf{y}|\mathbf{x})$; i.e. for any given $\epsilon > 0$, there exists m_ϵ such that for all $m > m_\epsilon$, $|\hat{\mathbb{P}}(\mathbf{y}|\mathbf{x}) - \mathbb{P}(\mathbf{y}|\mathbf{x})| < \epsilon$, with high probability. It follows that, with high probability, (9) $\le \epsilon \sum_{y \in \{0,1\}^n} L(\mathbf{s}, \mathbf{y})$. Assuming L is bounded, we have that for any fixed s, $\big|\mathcal{U}^L(\mathbf{s}; \mathbb{P}) - \mathcal{U}^L(\mathbf{s}; \hat{\mathbb{P}})\big| \le C\epsilon$, for some constant C that depends only on the metric L and (fixed) test set size n. The uniform convergence also follows because the $\max$ in (8) is over finitely many vectors $\mathbf{s}$. Putting together, we have that for any given $\delta, \epsilon' > 0$, there exists training sample size $m_{\epsilon', \delta}$ such that the output $\hat{\mathbf{s}}$ of our procedure satisfies, with probability at least $1 - \delta, \widehat{\mathcal{U}}^L - \mathcal{U}_*^L < \epsilon'$; when L is unbounded, we have that $\mathbf{s}^* = \underset{\mathbf{s} \in \{0,1\}^n}{\arg\min}\ L(\mathbf{s}, \cdot)$ over all unbounded $L(\mathbf{s}, \mathbf{y})$. Thus all that is required is support consistency i.e. $\{\mathbf{y}|\hat{\mathbb{P}}(\mathbf{y}|\mathbf{x}) > 0\} \xrightarrow{p} \{\mathbf{y}|\mathbb{P}(\mathbf{y}|\mathbf{x}) > 0\}$ which is a much weaker condition than distribution consistency. The proof is complete.

C. Appendix C

EUM and DTA Classification

A recent flurry of theoretical results and practical algorithms highlights a growing interest in understanding and optimizing non-decomposable metrics (Dembczynski et al., 2011; Ye et al., 2012; Koyejo et al., 2014; Narasimhan et al., 2014). Existing theoretical analysis has focused on two distinct approaches for characterizing the *population* version of the non-decomposable metrics: identified by Ye et al. (2012) as decision theoretic analysis (DTA) and empirical utility maximization (EUM). DTA population utilities measure the expected gain of a classifier on a fixed-size test set, while EUM population utilities are a function of the population confusion matrix. In other words, DTA population utilities measure the the average utility over an infinite set of test sets, each of a fixed size, while EUM population utilities evaluate the performance of a classifier over a single infinitely large test set.

It has recently been shown that for EUM based population utilities, the optimal classifier for large classes of non-decomposable binary classification metrics is just the sign of the thresholded conditional probability of the positive class with a metric-dependent threshold (Koyejo et al., 2014; Narasimhan et al., 2014). In addition, practical algorithms have been proposed for such EUM consistent classification based on direct optimization for the threshold on a held-out validation set. In stark contrast to this burgeoning understanding of EUM optimal classification, we are aware of only two metrics for which DTA consistent classifiers have been derived and shown to exhibit a simple form; namely, the F_β metric (Lewis, 1995; Dembczynski et al., 2011; Ye et al., 2012) and squared error in counting (SEC) studied by Lewis (1995).

While the optimal classifiers of both EUM and DTA population utilities associated with the performance metrics we study comprise signed thresholding of the conditional probability of the positive class, the evaluation and optimization for EUM and DTA utilities require quite different techniques. Given a classifier and a distribution, evaluating a population DTA utility can involve exponential-time computation, even leaving aside maximizing the utility on a fixed test set. As we show, in light of the probability ranking principle, and with careful implementation, this can actually be reduced to cubic complexity. These computations can be further reduced to quadratic complexity in a few special cases (Ye et al., 2012). To this end, we propose two algorithms for optimal DTA classification. The first algorithm runs in $O(n^3)$ time for a general metric, where n is the size of the test set and the second algorithm runs in time $O(n^2)$ for special cases such as F_β and Jaccard. We show that our overall procedure for decision-theoretic classification is consistent. More recently, Parambath et al. (2014) gave a theoretical analysis of the binary and multi-label F_β measure in the EUM setting. Dembczynski et al. (2011) analyzed the F_β measure in the DTA setting including the case where the data is non i.i.d., and also proposed efficient algorithms for optimal classification.