Consistent Multilabel Classification

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

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

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

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

Abstract

Multilabel classification is rapidly developing as an important aspect of modern predictive modeling, motivating study of its theoretical aspects. To this end, we propose a framework for constructing and analyzing multilabel classification metrics which reveals novel results on a parametric form for population optimal classifiers, and additional insight into the role of label correlations. In particular, we show that for multilabel metrics constructed as instance-, micro- and macro-averages, the population optimal classifier can be decomposed into binary classifiers based on the marginal instance-conditional distribution of each label, with a weak association between labels via the threshold. Thus, our analysis extends the state of the art from a few known multilabel classification metrics such as Hamming loss, to a general framework applicable to many of the classification metrics in common use. Based on the population-optimal classifier, we propose a computationally efficient and general-purpose plug-in classification algorithm, and prove its consistency with respect to the metric of interest. Empirical results on synthetic and benchmark datasets are supportive of our theoretical findings.

1 Introduction

Modern classification problems often involve the prediction of multiple labels simultaneously associated with a single instance e.g. image tagging by predicting multiple objects in an image. The growing importance of multilabel classification has motivated the development of several scalable algorithms [8, 12, 18] and has led to the recent surge in theoretical analysis [1, 3, 7, 16] which helps guide and understand practical advances. While recent results have advanced our knowledge of optimal population classifiers and consistent learning algorithms for particular metrics such as the Hamming loss and multilabel F-measure [3, 4, 5], a general understanding of learning with respect to multilabel classification metrics has remained an open problem. This is in contrast to the more traditional settings of binary and multiclass classification where several recently established results have led to a rich understanding of optimal and consistent classification [9, 10, 11]. This manuscript constitutes a step towards establishing results for multilabel classification at the level of generality currently enjoyed only in these traditional settings.

Towards a generalized analysis, we propose a framework for multilabel sample performance metrics and their corresponding population extensions. A classification *metric* is constructed to measure the utility[1] of a classifier, as defined by the practitioner or end-user. The utility may be measured using

*Equal contribution.
[1]Equivalently, we may define the loss as the negative utility.

the *sample* metric given a finite dataset, and further generalized to the *population* metric with respect to a given data distribution (i.e. with respect to infinite samples). Two distinct approaches have been proposed for studying the population performance of classifier in the classical settings of binary and multiclass classification, described by Ye et al. [17] as decision theoretic analysis (DTA) and empirical utility maximization (EUM). DTA population utilities measure the expected performance of a classifier on a fixed-size test set, while EUM population utilities are directly defined as a function of the population confusion matrix. However, state-of-the-art analysis of multilabel classification has so-far lacked such a distinction. The proposed framework defines both EUM and DTA multilabel population utility as generalizations of the aforementioned classic definitions. Using this framework, we observe that existing work on multilabel classification [1, 3, 7, 16] have exclusively focused on optimizing the DTA utility of (specific) multilabel metrics.

Averaging of binary classification metrics remains one of the most widely used approaches for defining multilabel metrics. Given a binary label representation, such metrics are constructed via averaging with respect to labels (instance-averaging), with respect to examples separately for each label (macro-averaging), or with respect to both labels and examples (micro-averaging). We consider a large sub-family of such metrics where the underlying binary metric can be constructed as a fraction of linear combinations of true positives, false positives, false negatives and true negatives [9]. Examples in this family include the ubiquitous Hamming loss, the averaged precision, the multilabel averaged F-measure, and the averaged Jaccard measure, among others. Our key result is that a Bayes optimal multilabel classifier for such metrics can be explicitly characterized in a simple form — the optimal classifier thresholds the label-wise conditional probability marginals, and the label dependence in the underlying distribution is relevant to the optimal classifier only through the threshold parameter. Further, the threshold is *shared* by all the labels when the metric is instance-averaged or micro-averaged. This result is surprising and, to our knowledge, a first result to be shown at this level of generality for multilabel classification. The result also sheds additional insight into the role of label correlations in multilabel classification – answering prior conjectures by Dembczyński et al. [3] and others.

We provide a plug-in estimation based algorithm that is efficient as well as theoretically consistent, i.e. the true utility of the empirical estimator approaches the optimal (EUM) utility of the Bayes classifier (Section 4). We also present experimental evaluation on synthetic and real-world benchmark multilabel datasets comparing different estimation algorithms (Section 5) for representative multilabel performance metrics selected from the studied family. The results observed in practice are supportive of what the theory predicts.

1.1 Related Work

We briefly highlight closely related theoretical results in the multilabel learning literature. Gao and Zhou [7] consider the consistency of multilabel learning with respect to DTA utility, with a focus on two specific losses – Hamming and rank loss (the corresponding measures are defined in Section 2). Surrogate losses are devised which result in consistent learning with respect to these metrics. In contrast, we propose a plug-in estimation based algorithm which directly estimates the Bayes optimal, without going through surrogate losses. Dembczynski et al. [2] analyze the DTA population optimal classifier for the multilabel rank loss, showing that the Bayes optimal is independent of label correlations in the unweighted case, and construct certain weighted univariate losses which are DTA consistent surrogates in the more general weighted case. Perhaps the work most closely related to ours is by Dembczynski et al. [4] who propose a novel DTA consistent plug-in rule estimation based algorithm for multilabel F-measure. Cheng et al. [1] consider optimizing popular losses in multilabel learning such as Hamming, rank and subset 0/1 loss (which is the multilabel analog of the classical 0-1 loss). They propose a probabilistic version of classifier chains (first introduced by Read et al. [13]) for estimating the Bayes optimal with respect to subset 0/1 loss, though without rigorous theoretical justification.

2 A Framework for Multilabel Classification Metrics

Consider multilabel classification with M labels, where each instance is denoted by $x \in \mathcal{X}$. For convenience, we will focus on the common binary encoding, where the labels are represented by a vector $\mathbf{y} \in \mathcal{Y} = \{0,1\}^M$, so $y_m = 1$ iff the m^{th} label is associated with the instance, and

$y_m = 0$ otherwise. The goal is to learn a multilabel classifier $\mathbf{f} : \mathcal{X} \mapsto \mathcal{Y}$ that optimizes a certain performance metric with respect to $\mathbb{P}$ – a fixed data generating distribution over the domain $\mathcal{X} \times \mathcal{Y}$, using a training set of instance-label pairs $(x^{(n)}, \mathbf{y}^{(n)})$, $n = 1, 2, \ldots, N$ drawn (typically assumed iid.) from $\mathbb{P}$. Let X and Y denote the random variables for instances and labels respectively, and let Ψ denote the performance (utility) metric of interest.

Most classification metrics can be represented as functions of the entries of the confusion matrix. In case of binary classification, the confusion matrix is specified by four numbers, i.e., true positives, true negatives, false positives and false negatives. Similarly, we construct the following primitives for multilabel classification:

$$\begin{aligned} \widehat{\text{TP}}(\mathbf{f})_{m,n} &= [[f_m(x^{(n)}) = 1, y_m^{(n)} = 1]] \qquad & \widehat{\text{TN}}(\mathbf{f})_{m,n} &= [[f_m(x^{(n)}) = 0, y_m^{(n)} = 0]] \\ \widehat{\text{FP}}(\mathbf{f})_{m,n} &= [[f_m(x^{(n)}) = 1, y_m^{(n)} = 0]] \qquad & \widehat{\text{FN}}(\mathbf{f})_{m,n} &= [[f_m(x^{(n)}) = 0, y_m^{(n)} = 1]] \end{aligned} \tag{1}$$

where $[[Z]]$ denotes the indicator function that is 1 if the predicate Z is true or 0 otherwise. It is clear that most multilabel classification metrics considered in the literature can be written as a function of the MN primitives defined in (1).

In the following, we consider a construction which is of sufficient generality to capture all multilabel metrics in common use. Let $A_k(\mathbf{f}) : \{\widehat{\text{TP}}(\mathbf{f})_{m,n}, \widehat{\text{FP}}(\mathbf{f})_{m,n}, \widehat{\text{TN}}(\mathbf{f})_{m,n}, \widehat{\text{FN}}(\mathbf{f})_{m,n}\}_{m=1,n=1}^{M,N} \mapsto \mathbb{R}$, $k = 1, 2, \ldots, K$ represent a set of K functions. Consider *sample* multilabel metrics constructed as functions: $\Psi : \{A_k(\mathbf{f})\}_{k=1}^K \mapsto [0, \infty)$. We note that the metric need *not* decompose over individual instances. Equipped with this definition of a sample performance metric Ψ, consider the population *utility* of a multilabel classifier $\mathbf{f}$ defined as:

$$\mathcal{U}(\mathbf{f}; \Psi, \mathbb{P}) = \Psi(\{\text{E}\left[\, A_k(\mathbf{f}) \,\right]\}_{k=1}^K), \tag{2}$$

where the expectation is over iid draws from the joint distribution $\mathbb{P}$. Note that this can be seen as a multilabel generalization of the so-called Empirical Utility Maximization (EUM) style classifiers studied in binary [9, 10] and multiclass [11] settings.

Our goal is to learn a multilabel classifier that maximizes $\mathcal{U}(\mathbf{f}; \Psi, \mathbb{P})$ for general performance metrics Ψ. Define the (Bayes) optimal multilabel classifier as:

$$\mathbf{f}_\Psi^* = \underset{\mathbf{f}:\mathcal{X} \to \{0,1\}^M}{\text{argmax}} \; \mathcal{U}(\mathbf{f}; \Psi, \mathbb{P}). \tag{3}$$

Let $\mathcal{U}(\mathbf{f}_\Psi^*; \Psi, \mathbb{P}) = \mathcal{U}_\Psi^*$. We say that $\hat{\mathbf{f}}_\Psi$ is a consistent estimator of $\mathbf{f}_\Psi^*$ if $\mathcal{U}(\hat{\mathbf{f}}; \Psi, \mathbb{P}) \xrightarrow{p} \mathcal{U}_\Psi^*$.

Examples. The averaged accuracy (1 - Hamming loss) used in multilabel classification corresponds to simply choosing: $A_1(\mathbf{f}) = \frac{1}{MN} \sum_{m=1}^M \sum_{n=1}^N \widehat{\text{FP}}(\mathbf{f})_{m,n} + \widehat{\text{FN}}(\mathbf{f})_{m,n}$ and $\Psi_{\text{Ham}}(\mathbf{f}) = 1 - A_1(\mathbf{f})$. The measure corresponding to rank loss[2] can be obtained by choosing $A_k(\mathbf{f}) = \frac{1}{M^2} \sum_{m_1=1}^M \sum_{m_2=1}^M \left(\widehat{\text{FP}}(\mathbf{f})_{m_1,k}\right) \left(\widehat{\text{FN}}(\mathbf{f})_{m_2,k}\right)$, for $k = 1, 2, \ldots, N$ and $\Psi_{\text{Rank}} = 1 - \frac{1}{N} \sum_{k=1}^N A_k(\mathbf{f})$. Note that the choice of $\{A_k\}$, and therefore Ψ, is not unique.

Remark 1. *Existing results on multilabel classification have focused on decision-theoretic analysis (DTA) style classifiers, where the utility is defined as:*

$$\mathcal{U}_{\text{DTA}}(\mathbf{f}; \Psi, \mathbb{P}) = \text{E}\left[\, \Psi(\{A_k(\mathbf{f})\}_{k=1}^K) \,\right], \tag{4}$$

and the expectation is over iid samples from $\mathbb{P}$. Furthermore, there are no theoretical results for consistency with respect to general performance metrics Ψ in this setting (See Appendix B.2).

For the remainder of this manuscript, we refer to $\mathcal{U}(\mathbf{f}; \mathbb{P})$ as the utility defined in (2). We will also drop the argument $\mathbf{f}$ (e.g. write $\widehat{\text{TP}}(\mathbf{f})$ as $\widehat{\text{TP}}$) when it is clear from the context.

2.1 A Framework for Averaged Binary Multilabel Classification Metrics

The most popular class of multilabel performance metrics consists of averaged binary performance metrics, that correspond to particular settings of $\{A_k(\mathbf{f})\}$ using certain averages as described in the following. For the remainder of this subsection, the metric $\Psi : [0, 1]^4 \to [0, \infty)$ will refer to a binary classification metric as is typically applied to a binary confusion matrix.

[2] A subtle but important aspect of the definition of rank loss in the existing literature, including [2] and [7], is that the Bayes optimal is allowed to be a real-valued function and may not correspond to a label decision.

Micro-averaging Ψ_{micro}. Micro-averaged multilabel performance metrics are defined by averaging over both labels and examples. Let:

$$\widehat{\text{TP}}(\mathbf{f}) = \frac{1}{MN}\sum_{n=1}^{N}\sum_{m=1}^{M}\widehat{\text{TP}}(\mathbf{f})_{m,n}, \qquad \widehat{\text{FP}}(\mathbf{f}) = \frac{1}{MN}\sum_{n=1}^{N}\sum_{m=1}^{M}\widehat{\text{FP}}(\mathbf{f})_{m,n}, \tag{5}$$

$\widehat{\text{TN}}(\mathbf{f})$ and $\widehat{\text{FN}}(\mathbf{f})$ are defined similarly, then the micro-averaged multilabel performance metrics are given by:

$$\Psi_{\text{micro}}(\{A_k(\mathbf{f})\}_{k=1}^{K}) := \Psi(\widehat{\text{TP}}, \widehat{\text{FP}}, \widehat{\text{TN}}, \widehat{\text{FN}}). \tag{6}$$

In other words, for micro-averaging, one applies a binary performance metric to the confusion matrix defined by the four (averaged) quantities above. The other averaged binary metrics are defined similarly.

Macro-averaging Ψ_{macro}. Macro-averaging measures average classification performance across labels. Define the averaged measures:

$$\widehat{\text{TP}}_m(\mathbf{f}) = \frac{1}{N}\sum_{n=1}^{N}\widehat{\text{TP}}(\mathbf{f})_{m,n}, \qquad \widehat{\text{FP}}_m(\mathbf{f}) = \frac{1}{N}\sum_{n=1}^{N}\widehat{\text{FP}}(\mathbf{f})_{m,n},$$

$\widehat{\text{TN}}_m(\mathbf{f})$ and $\widehat{\text{FN}}_m(\mathbf{f})$ are defined similarly. The macro-averaged performance metric is given by:

$$\Psi_{\text{macro}}(\{A_k(\mathbf{f})\}_{k=1}^{K}) := \frac{1}{M}\sum_{m=1}^{M}\Psi(\widehat{\text{TP}}_m, \widehat{\text{FP}}_m, \widehat{\text{TN}}_m, \widehat{\text{FN}}_m). \tag{7}$$

Instance-averaging Ψ_{instance}. This measures the average classification performance across examples. Define the averaged measures:

$$\widehat{\text{TP}}_n(\mathbf{f}) = \frac{1}{M}\sum_{m=1}^{M}\widehat{\text{TP}}(\mathbf{f})_{m,n}, \qquad \widehat{\text{FP}}_n(\mathbf{f}) = \frac{1}{M}\sum_{m=1}^{M}\widehat{\text{FP}}(\mathbf{f})_{m,n},$$

$\widehat{\text{TN}}_n(\mathbf{f})$ and $\widehat{\text{FN}}_n(\mathbf{f})$ are defined similarly. The instance-averaged performance metric is given by:

$$\Psi_{\text{instance}}(\{A_k(\mathbf{f})\}_{k=1}^{K}) := \frac{1}{N}\sum_{n=1}^{N}\Psi(\widehat{\text{TP}}_n, \widehat{\text{FP}}_n, \widehat{\text{TN}}_n, \widehat{\text{FN}}_n). \tag{8}$$

3 Characterizing the Bayes Optimal Classifier for Multilabel Metrics

We now characterize the optimal multilabel classifier for a large family of multilabel metrics Ψ_{micro}, Ψ_{macro} and Ψ_{instance} as outlined in Section 2.1 with respect to the EUM utility. We begin by observing that while micro-averaging and instance-averaging seem quite different when viewed as sample averages, they are in fact equivalent at the population level. In light of the Proposition, we need only focus on one definition (Ψ_{micro}), to characterize the Bayes optimal for both cases.

Proposition 1. *For a given binary classification metric Ψ, consider the averaged multilabel metrics Ψ_{micro} defined in* (6) *and Ψ_{instance} defined in* (8)*. For any $\mathbf{f}$, $\mathcal{U}(\mathbf{f}; \Psi_{\text{micro}}, \mathbb{P}) \equiv \mathcal{U}(\mathbf{f}; \Psi_{\text{instance}}, \mathbb{P})$. In particular, $\mathbf{f}^*_{\Psi^*_{\text{micro}}} \equiv \mathbf{f}^*_{\Psi^*_{\text{instance}}}$.*

We further restrict our study to metrics Ψ selected from the linear-fractional metric family, recently studied in the context of binary classification [9]. Any Ψ in this family can be written as:

$$\Psi(\widehat{\text{TP}}, \widehat{\text{FP}}, \widehat{\text{FN}}, \widehat{\text{TN}}) = \frac{a_0 + a_{11}\widehat{\text{TP}} + a_{10}\widehat{\text{FP}} + a_{01}\widehat{\text{FN}} + a_{00}\widehat{\text{TN}}}{b_0 + b_{11}\widehat{\text{TP}} + b_{10}\widehat{\text{FP}} + b_{01}\widehat{\text{FN}} + b_{00}\widehat{\text{TN}}},$$

where $a_0, b_0, a_{ij}, b_{ij}, i, j \in \{0, 1\}$ are fixed constants, and $\widehat{\text{TP}}, \widehat{\text{FP}}, \widehat{\text{FN}}, \widehat{\text{TN}}$ are defined as in (5) for Ψ_{micro}. Many popular multilabel metrics can be derived using linear-fractional Ψ. Some examples include[3]:

$$
\begin{aligned}
F_\beta: \quad \Psi_{F_\beta} &= \frac{(1+\beta^2)\widehat{\text{TP}}}{(1+\beta^2)\widehat{\text{TP}}+\beta^2\widehat{\text{FN}}+\widehat{\text{FP}}} &\quad \text{Jaccard}: \quad \Psi_{\text{Jacc}} &= \frac{\widehat{\text{TP}}}{\widehat{\text{TP}}+\widehat{\text{FP}}+\widehat{\text{FN}}} \\
\text{Hamming}: \quad \Psi_{\text{Ham}} &= \widehat{\text{TP}}+\widehat{\text{TN}} &\quad \text{Precision}: \quad \Psi_{\text{Prec}} &= \frac{\widehat{\text{TP}}}{\widehat{\text{TP}}+\widehat{\text{FP}}}
\end{aligned} \tag{9}
$$

Define the population quantities: $\pi = \sum_{m=1}^{M} \mathbb{P}(Y_m = 1)$ and $\gamma(\mathbf{f}) = \sum_{m=1}^{M} \mathbb{P}(f_m(x) = 1)$. Let $\text{TP}(\mathbf{f}) = \mathrm{E}\left[\widehat{\text{TP}}(\mathbf{f})\right]$, where the expectation is over iid draws from $\mathbb{P}$. From (5), it follows that, $\text{FP}(\mathbf{f}) := \mathrm{E}\left[\widehat{\text{FP}}(\mathbf{f})\right] = \gamma(\mathbf{f}) - \text{TP}(\mathbf{f})$, $\text{TN}(\mathbf{f}) = 1 - \pi - \gamma(\mathbf{f}) + \text{TP}(\mathbf{f})$ and $\text{FN}(\mathbf{f}) = \gamma(\mathbf{f}) - \text{TP}(\mathbf{f})$.

Now, the population utility (2) corresponding to Ψ_{micro} can be written succinctly as:

$$
\mathcal{U}(\mathbf{f}; \Psi_{\text{micro}}, \mathbb{P}) = \Psi(\text{TP}(\mathbf{f}), \text{FP}(\mathbf{f}), \text{FN}(\mathbf{f}), \text{TN}(\mathbf{f})) = \frac{c_0 + c_1 \text{TP}(\mathbf{f}) + c_2 \gamma(\mathbf{f})}{d_0 + d_1 \text{TP}(\mathbf{f}) + d_2 \gamma(\mathbf{f})} \tag{10}
$$

with the constants:

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

We assume that the joint $\mathbb{P}$ has a density μ that satisfies $d\mathbb{P} = \mu dx$, and define $\eta_m(x) = \mathbb{P}(Y_m = 1 | X = x)$. Our first main result characterizes the Bayes optimal multilabel classifier $\mathbf{f}^*_{\Psi_{\text{micro}}}$.

Theorem 2. *Given the constants* $\{c_1, c_2, c_0\}$ *and* $\{d_1, d_2, d_0\}$*, define:*

$$
\delta^* = \frac{d_2 \mathcal{U}^*_{\Psi_{micro}} - c_2}{c_1 - d_1 \mathcal{U}^*_{\Psi_{micro}}}. \tag{11}
$$

The optimal Bayes classifier $\mathbf{f}^* := \mathbf{f}^*_{\Psi_{micro}}$ *defined in* (3) *is given by:*

1. *When* $c_1 > d_1 \mathcal{U}^*_{\Psi_{micro}}$, $\mathbf{f}^*$ *takes the form* $f^*_m(x) = [[\eta_m(x) > \delta^*]]$*, for* $m \in [M]$.
2. *When* $c_1 < d_1 \mathcal{U}^*_{\Psi_{micro}}$, $\mathbf{f}^*$ *takes the form* $f^*_m(x) = [[\eta_m(x) \leq \delta^*]]$*, for* $m \in [M]$.

The proof is provided in Appendix A.2, and applies equivalently to instance-averaging. Theorem 2 recovers existing results in binary [9] settings (See Appendix B.1 for details), and is sufficiently general to capture many of the multilabel metrics used in practice. Our proof is closely related to the binary classification case analyzed in Theorem 2 of [9], but differs in the additional averaging across labels. A key observation from Theorem 2 is that the optimal multilabel classifier can be obtained by thresholding the marginal instance-conditional probability for each label $\mathbb{P}(Y_m = 1|x)$ and, importantly, that the optimal classifiers for all the labels share the *same* threshold δ^*. Thus, the effect of the joint distribution is only in the threshold parameter. We emphasize that while the presented results characterize the optimal population classifier, incorporating label correlations into the prediction algorithm may have other benefits with finite samples, such as statistical efficiency when there are known structural similarities between the marginal distributions [3]. Further analysis is left for future work.

The Bayes optimal for the macro-averaged population metric is straightforward to establish. We observe that the threshold is not shared in this case.

Proposition 3. *For a given linear-fractional metric* Ψ*, consider the macro-averaged multilabel metric* Ψ_{macro} *defined in* (7)*. Let* $\mathbf{f}^* = \mathbf{f}^*_{\Psi^*_{\text{macro}}}(x)$*. We have, for* $m = 1, 2, \ldots, M$*:*

$$
f^*_m = [[\eta_m(x) > \delta^*_m]],
$$

where $\delta^*_m \in [0, 1]$ *is a constant that depends on the metric* Ψ *and the label-wise instance-conditional marginals of* $\mathbb{P}$.

Remark 2. *It is clear that micro-, macro- and instance- averaging are equivalent at the population level when the metric* Ψ *is linear. This is a straightforward consequence of the observation that the corresponding sample-based utilities are the same. More generally, micro-, macro- and instance-averaging are equivalent whenever the optimal threshold is a constant independent of* $\mathbb{P}$*, such as for linear metrics, where* $d_1 = d_2 = 0$ *so* $\delta^* = -\frac{c_2}{c_1}$ *(cf. Corollary 4 of Koyejo et al. [9]). Thus, our analysis recovers known results for Hamming loss [3, 7].*

[3]Note that Hamming is typically defined as the loss, given by $1 - \Psi_{\text{Ham}}$.

4 Consistent Plug-in Estimation Algorithm

An important consequence of the characterization of Bayes optimal is that it enables a simple plug-in estimation algorithm that enjoys consistency. The overall procedure to obtain a consistent classifier for the averaged metrics Ψ_{micro} (equiv. Ψ_{instance}) is as follows. First, we obtain an estimate $\hat{\eta}_m(x)$ of the marginal instance-conditional probability $\eta_m(x) = \mathbb{P}(Y_m = 1|x)$ for each label m using validation samples assembled for the corresponding labels $\{(x^{(n)}, y_m^{(n)})\}$ (see Reid and Williamson [14]). Then, the given metric $\Psi_{\text{micro}}(\mathbf{f})$ is maximized on the sample. Note that it suffices to maximize over $\{\mathbf{f}_\delta : f_m(x) = [[\eta_m(x) > \delta]] \; \forall m = 1, 2, \ldots, M\}$ for fixed threshold δ as:

$$\hat{\delta} = \underset{\delta \in (0,1)}{\text{argmax}} \; \Psi_{\text{micro}}(\hat{\mathbf{f}}_\delta), \tag{12}$$

where Ψ_{micro} is the micro-averaged sample metric defined as in (6) (similarly for Ψ_{instance}). Though the threshold search is over a continuous space $\delta \in (0, 1)$ the number of distinct $\Psi_{\text{micro}}(\hat{\mathbf{f}}_\delta)$ values given a training sample of size N is at most NM. Thus (12) can be solved efficiently on a finite sample.

Algorithm 1: Plugin-Estimator for Ψ_{micro} and Ψ_{instance}

Input: Training examples $\mathcal{S} = \{x^{(n)}, \mathbf{y}^{(n)}\}_{n=1}^N$ and metric Ψ_{micro} (or Ψ_{instance}).
for $m = 1, 2, \ldots, M$ **do**
1. Select the training data for label m: $\mathcal{S}_m = \{x^{(n)}, y_m^{(n)}\}_{n=1}^N$.
2. Split the training data $\mathcal{S}_m$ into two sets $\mathcal{S}_{m1}$ and $\mathcal{S}_{m2}$.
3. Estimate $\hat{\eta}_m(x)$ using $\mathcal{S}_{m1}$, define $\hat{f}_m(x) = [[\hat{\eta}_m(x) > \delta]]$.

end for
Obtain $\hat{\delta}$ by solving (12) on $\mathcal{S}_2 = \cup_{m=1}^M \mathcal{S}_{m2}$.
Return: $\hat{\mathbf{f}}_{\hat{\delta}}$.

Consistency of the proposed algorithm. The following theorem shows that the plug-in procedure of Algorithm 1 results in a consistent classifier.

Theorem 4. *Let Ψ_{micro} be a linear-fractional metric. If the estimates $\hat{\eta}_m(x)$ satisfy $\hat{\eta}_m \xrightarrow{p} \eta_m$, $\forall m$, then the output multilabel classifier $\hat{\mathbf{f}}_{\hat{\delta}}$ of Algorithm 1 is consistent.*

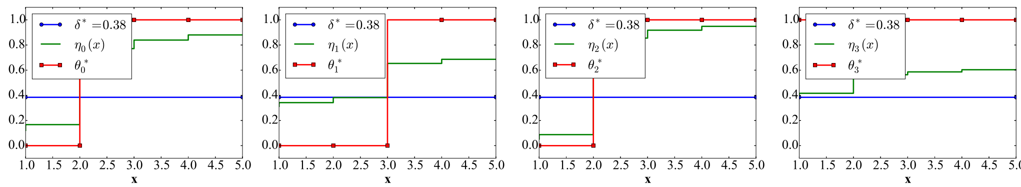
The proof is provided in Appendix A.4. From Proposition 1, it follows that consistency holds for Ψ_{instance} as well. Additionally, in light of Proposition 3, we may apply the learning algorithms proposed by [9] for binary classification independently for each label to obtain a consistent estimator for Ψ_{macro}.

5 Experiments

We present two sets of results. The first is an experimental validation on synthetic data with known ground truth probabilities. The results serve to verify our main result (Theorem 2) characterizing the Bayes optimal for averaged multilabel metrics. The second is an experimental evaluation of the plugin estimator algorithms for micro-, instance-, and macro-averaged multilabel metrics on benchmark datasets.

5.1 Synthetic data: Verification of Bayes optimal

We consider the micro-averaged F_1 metric in (9) for multilabel classification with 4 labels. We sample a set of five 2-dimensional vectors $\mathbf{x} = \{x^{(1)}, x^{(2)}, \ldots, x^{(5)}\}$ from the standard Gaussian. The conditional probability η_m for label m is modeled using a sigmoid function: $\eta_m(x) = \mathbb{P}(Y_m = 1|x) = \frac{1}{1+\exp -w_m^T x}$, using a vector w_m sampled from the standard Gaussian. The Bayes optimal $\mathbf{f}^*(x) \in \{0, 1\}^4$ that maximizes the micro-averaged F_1 population utility is then obtained by exhaustive search over all possible label vectors for each instance. In Figure 1 (a)-(d), we plot the



(a) (b) (c) (d)

Figure 1: Bayes optimal classifier for multilabel F_1 measure on synthetic data with 4 labels, and distribution supported on 5 instances. Plots from left to right show the bayes optimal classifier prediction for instances, for labels 1 through 4. Note that the optimal δ^* at which the label-wise marginal $\eta_m(x)$ is thresholded is shared, conforming to Theorem 2 (larger plots are included in Appendix C).

conditional probabilities (wrt. the sample index n) for each label, the corresponding f_m^* for each x, and the optimal threshold δ^* using (11). We observe that the optimal multilabel classifier indeed thresholds $\mathbb{P}(Y_m|x)$ for each label m, and furthermore, that the threshold is same for all the labels, as stated in Theorem 2.

5.2 Benchmark data: Evaluation of plug-in estimators

We now evaluate the proposed plugin-estimation (Algorithm 1) that is consistent for micro- and instance-averaged multilabel metrics. We focus on two metrics, F_1 and Jaccard, listed in (9). We compare Algorithm 1, designed to optimize micro-averaged (or instance-averaged) multilabel metrics to two related plugin-estimation methods: (i) a separate threshold δ_m^* tuned for each label m individually — this optimizes the utility corresponding to the macro-averaged metric, but is not consistent for micro-averaged or instance-averaged metrics; we refer to this as Macro-Thres (ii) a constant threshold $1/2$ for all the labels — this is known to be optimal for averaged accuracy (equiv. Hamming loss), but not for non-decomposable F_1 or Jaccard metrics. We refer to this as Binary Relevance (BR) [15].

We use four benchmark multilabel datasets[4] in our experiments: (i) SCENE, an image dataset consisting of 6 labels, with 1211 training and 1196 test instances, (ii) BIRDS, an audio dataset consisting of 19 labels, with 323 training and 322 test instances, (iii) EMOTIONS, a music dataset consisting of 6 labels, with 393 training and 202 test instances, and (iv) CAL500, a music dataset consisting of 174 labels, with 400 training and 100 test instances[5]. We perform logistic regression (with L_2 regularization) on a separate validation sample to obtain estimates of $\hat{\eta}_m(x)$ of $\mathbb{P}(Y_m = 1|x)$, for each label m (as described in Section 4). All the methods we evaluate rely on obtaining a good estimator for the conditional probability. So we exclude labels that are associated with very few instances — in particular, we train and evaluate using labels associated with at least 20 instances, in each dataset, for all the methods.

In Table 1, we report the micro-averaged F_1 and Jaccard metrics on the test set for Algorithm 1, Macro-Thres and Binary Relevance. We observe that estimating a fixed threshold for all the labels (Algorithm 1) consistently performs better than estimating thresholds for each label (Macro-Thres) and than using threshold 1/2 for all labels (BR); this conforms to our main result in Theorem 2 and the consistency analysis of Algorithm 1 in Theorem 4. A similar trend is observed for the instance-averaged metrics computed on the test set, shown in Table 2. Proposition 1 shows that maximizing the population utilities of micro-averaged and instance-averaged metrics are equivalent; the result holds in practice as presented in Table 2. Finally, we report macro-averaged metrics computed on test set in Table 3. We observe that Macro-Thres is competitive in 3 out of 4 datasets; this conforms to Proposition 3 which shows that in the case of macro-averaged metrics, it is optimal to tune a threshold specific to each label independently. Beyond consistency, we note that by using more

[4]The datasets were obtained from `http://mulan.sourceforge.net/datasets-mlc.html`.

[5]Original CAL500 dataset does not provide splits; we split the data randomly into train and test sets.

DATASET	BR	Algorithm 1 F_1	Macro-Thres	BR	Algorithm 1 Jaccard	Macro-Thres
SCENE	0.6559	**0.6847** $\pm$ 0.0072	0.6631 $\pm$ 0.0125	0.4878	**0.5151** $\pm$ 0.0084	0.5010 $\pm$ 0.0122
BIRDS	**0.4040**	**0.4088** $\pm$ 0.0130	0.2871 $\pm$ 0.0734	0.2495	**0.2648** $\pm$ 0.0095	0.1942 $\pm$ 0.0401
EMOTIONS	0.5815	**0.6554** $\pm$ 0.0069	0.6419 $\pm$ 0.0174	0.3982	**0.4908** $\pm$ 0.0074	0.4790 $\pm$ 0.0077
CAL500	0.3647	**0.4891** $\pm$ 0.0035	0.4160 $\pm$ 0.0078	0.2229	**0.3225** $\pm$ 0.0024	0.2608 $\pm$ 0.0056

Table 1: Comparison of plugin-estimator methods on multilabel F_1 and Jaccard metrics. Reported values correspond to *micro-averaged* metric (F_1 and Jaccard) computed on test data (with standard deviation, over 10 random validation sets for tuning thresholds). Algorithm 1 is consistent for micro-averaged metrics, and performs the best consistently across datasets.

DATASET	BR	Algorithm 1 F_1	Macro-Thres	BR	Algorithm 1 Jaccard	Macro-Thres
SCENE	0.5695	**0.6422** $\pm$ 0.0206	0.6303 $\pm$ 0.0167	0.5466	**0.5976** $\pm$ 0.0177	0.5902 $\pm$ 0.0176
BIRDS	0.1209	**0.1390** $\pm$ 0.0110	**0.1390** $\pm$ 0.0259	0.1058	**0.1239** $\pm$ 0.0077	0.1195 $\pm$ 0.0096
EMOTIONS	0.4787	**0.6241** $\pm$ 0.0204	0.6156 $\pm$ 0.0170	0.4078	**0.5340** $\pm$ 0.0072	0.5173 $\pm$ 0.0086
CAL500	0.3632	**0.4855** $\pm$ 0.0035	0.4135 $\pm$ 0.0079	0.2268	**0.3252** $\pm$ 0.0024	0.2623 $\pm$ 0.0055

Table 2: Comparison of plugin-estimator methods on multilabel F_1 and Jaccard metrics. Reported values correspond to *instance-averaged* metric (F_1 and Jaccard) computed on test data (with standard deviation, over 10 random validation sets for tuning thresholds). Algorithm 1 is consistent for instance-averaged metrics, and performs the best consistently across datasets.

DATASET	BR	Algorithm 1 F_1	Macro-Thres	BR	Algorithm 1 Jaccard	Macro-Thres
SCENE	0.6601	**0.6941** $\pm$ 0.0205	0.6737 $\pm$ 0.0137	0.5046	**0.5373** $\pm$ 0.0177	0.5260 $\pm$ 0.0176
BIRDS	0.3366	**0.3448** $\pm$ 0.0110	0.2971 $\pm$ 0.0267	0.2178	**0.2341** $\pm$ 0.0077	0.2051 $\pm$ 0.0215
EMOTIONS	0.5440	**0.6450** $\pm$ 0.0204	**0.6440** $\pm$ 0.0164	0.3982	**0.4912** $\pm$ 0.0072	**0.4900** $\pm$ 0.0133
CAL500	0.1293	0.2687 $\pm$ 0.0035	**0.3226** $\pm$ 0.0068	0.0880	0.1834 $\pm$ 0.0024	**0.2146** $\pm$ 0.0036

Table 3: Comparison of plugin-estimator methods on multilabel F_1 and Jaccard metrics. Reported values correspond to the *macro-averaged* metric computed on test data (with standard deviation, over 10 random validation sets for tuning thresholds). Macro-Thres is consistent for macro-averaged metrics, and is competitive in three out of four datasets. Though not consistent for macro-averaged metrics, Algorithm 1 achieves the best performance in three out of four datasets.

samples, joint threshold estimation enjoys additional statistical efficiency, while separate threshold estimation enjoys greater flexibility. This trade-off may explain why Algorithm 1 achieves the best performance in three out of four datasets in Table 3, though it is not consistent for macro-averaged metrics.

6 Conclusions and Future Work

We have proposed a framework for the construction and analysis of multilabel classification metrics and corresponding population optimal classifiers. Our main result is that for a large family of averaged performance metrics, the EUM optimal multilabel classifier can be explicitly characterized by thresholding of label-wise marginal instance-conditional probabilities, with weak label dependence via a shared threshold. We have also proposed efficient and consistent estimators for maximizing such multilabel performance metrics in practice. Our results are a step forward in the direction of extending the state-of-the-art understanding of learning with respect to general metrics in binary and multiclass settings. Our work opens up many interesting research directions, including the potential for further generalization of our results beyond averaged metrics, and generalized results for DTA population optimal classification, which is currently only well-understood for the F-measure.

Acknowledgments: We acknowledge the support of NSF via CCF-1117055, CCF-1320746 and IIS-1320894, 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

[1] Weiwei Cheng, Eyke Hüllermeier, and Krzysztof J Dembczynski. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 279–286, 2010.

[2] Krzysztof Dembczynski, Wojciech Kotlowski, and Eyke Hüllermeier. Consistent multilabel ranking through univariate losses. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1319–1326, 2012.

[3] 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.

[4] 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 International Conference on Machine Learning*, pages 1130–1138, 2013.

[5] 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.

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

[7] Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. *Artificial Intelligence*, 199:22–44, 2013.

[8] Ashish Kapoor, Raajay Viswanathan, and Prateek Jain. Multilabel classification using bayesian compressed sensing. In *Advances in Neural Information Processing Systems*, pages 2645–2653, 2012.

[9] 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.

[10] 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.

[11] Harikrishna Narasimhan, Harish Ramaswamy, Aadirupa Saha, and Shivani Agarwal. Consistent multiclass algorithms for complex performance measures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2398–2407, 2015.

[12] James Petterson and Tibério S Caetano. Submodular multi-label learning. In *Advances in Neural Information Processing Systems*, pages 1512–1520, 2011.

[13] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine learning*, 85(3):333–359, 2011.

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

[15] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer, 2010.

[16] 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.

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

[18] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. Large-scale multi-label learning with missing labels. In *Proceedings of the 31st International Conference on Machine Learning*, pages 593–601, 2014.

A Appendix A: Proofs

A.1 Proof of Proposition 1

Observe that the EUM utility (2) corresponding to micro-averaged metric is simply: $\mathcal{U}(\mathbf{f};\Psi_{\text{micro}},\mathbb{P}) = \Psi(\text{TP}(\mathbf{f}),\text{FP}(\mathbf{f}),\text{FN}(\mathbf{f}),\text{FP}(\mathbf{f}))$ where $\text{TP}(\mathbf{f}) = \text{E}_{\mathbb{P}}\left[\widehat{\text{TP}}(\mathbf{f})\right]$ ($\text{FP}(\mathbf{f}),\text{FN}(\mathbf{f}),\text{TN}(\mathbf{f})$ defined similarly). By linearity of expectation, $\text{E}_{\mathbb{P}}\left[\widehat{\text{TP}}(\mathbf{f})\right] = \frac{1}{MN}\sum_{n=1}^{N}\sum_{m=1}^{M}\text{E}_{\mathbb{P}}\left[\widehat{\text{TP}}(\mathbf{f})_{m,n}\right]$. But $\text{E}_{\mathbb{P}}\left[\widehat{\text{TP}}(\mathbf{f})_{m,n}\right] = \mathbb{P}(f_m(x) = 1, Y_m = 1)$. Similarly, for the instance-averaged metric, we see that the corresponding $\text{TP}(\mathbf{f}) = \text{E}_{\mathbb{P}}\left[\widehat{\text{TP}}(\mathbf{f})\right] = \mathbb{P}(f_m(x) = 1, Y_m = 1)$. Analogously, $\text{FP}(\mathbf{f}),\text{FN}(\mathbf{f}),\text{TN}(\mathbf{f})$ can be seen to be identical for micro- and instance-averaged metrics. Thus, whereas the sample metrics may be different, the population EUM utilities of the micro-averaged and instance-averaged metrics coincide. The second claim that $\mathbf{f}^*_{\Psi^*_{micro}} \equiv \mathbf{f}^*_{\Psi^*_{instance}}$ is immediate.

A.2 Proof of Theorem 2

For simplicity, the proof is presented for the finite domain case. Extension to the continuous case follows directly from the approach in Theorem 2 of [9], which requires a more technical definition of the derivatives. Let $\mathcal{F} = \{\mathbf{f} : \mathcal{X} \to \mathbb{R}^M\}$, and $\mathbf{\Theta} = \{\mathbf{f} : \mathcal{X} \to \{-1,+1\}^M\} \subset \mathcal{F}$ (Note that we use the encoding $\{+1,-1\}$ for ease and it is equivalent to $\{0,1\}$ encoding used in the main text). The derivative of $\mathcal{U}(\mathbf{f};\Psi_{\text{micro}},\mathbb{P})$ w.r.t. $f_m(x)$ is given by:

$$\nabla_{f_m(x)}\mathcal{U}(\mathbf{f}) = \frac{1}{(c_1 - d_1\mathcal{U}(\mathbf{f}))D_r(\mathbf{f})}\left[\eta_m(x) - \frac{d_2\mathcal{U}(\mathbf{f}) - c_2}{c_1 - d_1\mathcal{U}(\mathbf{f})}\right]\mu(x)$$

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

$$\sum_m \sum_{x\in\mathcal{X}} \nabla_{f_m(x)}\mathcal{U}(\mathbf{f}^*)f_m(x)dx \geq \sum_m \sum_{x\in\mathcal{X}} \nabla_{f_m(x)}\mathcal{U}(\mathbf{f}^*)f^*_m(x)dx \quad \forall\, \mathbf{f},\mathbf{f}^* \in \Theta.$$

Thus, when $c_1 \geq d_1\mathcal{U}^*$, a necessary condition for local optimality is that the sign of f^*_m and the sign of $[\nabla_{f_m(x)}\mathcal{U}(\mathbf{f}^*)]$ agree point-wise wrt. x, $\forall\, m$. This is equivalent to the condition that $\text{sign}(f^*_m) = \text{sign}(\eta_m(x) - \delta^*)$. Combining this result with the constraint set $\mathbf{f} \in \mathbf{\Theta}$, we have that $f^*_m = \text{sign}(f^*_m)$, thus $f^*_m = \text{sign}(\eta_m(x) - \delta^*)$ is locally optimal. Finally, we note that $f^*_m = \text{sign}(\eta_m(x) - \delta^*)$ is unique for $\mathbf{f} \in \mathbf{\Theta}$, thus $\mathbf{f}^*$ is globally optimal. The proof for $c_1 < d_1\mathcal{U}^*$ follows using similar arguments. The observation that the threshold is shared between labels follows by definition of the gradient, where we observe that the threshold depends on the optimal utility. We note that despite the close similarity, the above result cannot be derived from Theorem 2 of [9] without significant modification to accommodate a non-iid sampling distribution i.e. different label distributions when viewed as a binary classification problem.

A.3 Proof of Proposition 3

Note that the population utility corresponding to macro-averaged metric Ψ_{macro} can be written as:

$$\mathcal{U}(\mathbf{f};\Psi_{\text{macro}},\mathbb{P}) = \frac{1}{M}\sum_{m=1}^{M}\Psi(\text{TP}_m(\mathbf{f}),\text{FP}_m(\mathbf{f}),\text{FN}_m(\mathbf{f}),\text{TN}_m(\mathbf{f})),$$

where $\text{TP}_m(\mathbf{f}) = \mathbb{P}(f_m(x) = 1, Y_m = 1)$ ($\text{FP}_m(\mathbf{f}),\text{FN}_m(\mathbf{f}),\text{TN}_m(\mathbf{f})$ defined similarly). Let $\mathbb{P}_m$ denote the marginal $\mathbb{P}(X, Y_m = 1)$. Note that $\mathcal{U}(\mathbf{f};\Psi_{\text{macro}},\mathbb{P})$ is maximized by $\mathbf{f}^*$ if for all m, $\Psi(\text{TP}(f),\text{FP}(f),\text{FN}(f),\text{TN}(f))$, where $f : \mathcal{X} \to \{0,1\}$ and $\text{TP}(f) = \mathbb{P}_m(f(x) = 1, Y = 1)$, is optimized by f^*_m. From Theorem 2 of [9], we know that the Bayes optimal $f^*_m(x)$ for linear-fractional Ψ with respect to $\mathbb{P}_m$ is given by $[[\eta_m(x) > \delta^*_m]]$, where δ^*_m depends only on Ψ and marginal $\mathbb{P}_m$.

A.4 Proof of Theorem 4

We first show the following claim:

Claim. For a fixed multilabel classifier $\mathbf{f}$, $\Psi_{\text{micro}}(\mathbf{f}; \{x^{(n)}, \mathbf{y}^{(n)}\}_{n=1}^{N}) \xrightarrow{p} \mathcal{U}(\mathbf{f}; \Psi_{\text{micro}}; \mathbb{P})$, where Ψ_{micro} defined in (6) is the micro-averaged metric computed on iid samples.

Proof. The proof follows closely that of Lemma 8 of [9]. Consider the simplified empirical Ψ_{micro} corresponding to the simplified population utility in (10), obtained by replacing $\text{TP}(\mathbf{f})$ with $\widehat{\text{TP}}(\mathbf{f})$, $\gamma(\mathbf{f})$ with $\hat{\gamma}(\mathbf{f}) := \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}[[f_m^{(n)} = 1]]$ and π (in the definition of c_0 and d_0) with $\hat{\pi} = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}[[y_m^{(n)} = 1]]$. Note that for any given $\epsilon_1 > 0, \rho > 0$, there exists N' such that for any $N > N'$, $\mathbb{P}(|\widehat{\text{TP}}(\mathbf{f}) - E[\widehat{\text{TP}}(\mathbf{f})]| < \epsilon_1) > 1 - \rho/3$, $\mathbb{P}(|\widehat{\gamma}(\mathbf{f}) - \gamma(\mathbf{f})| < \epsilon_1) > 1 - \rho/3$ and $\mathbb{P}(|\hat{\pi} - \pi| < \epsilon_1) > 1 - \rho/3$. Thus by union bound it follows that all the three events hold with probability $1 - \rho$. Write $c_0 = c_o' + c_0''\pi$ and $d_0 = d_0' + d_0''\pi$. Let $\tilde{c}_1 = 1/|c_1|$ if $c_1 \neq 0$ else let $\tilde{c}_1 = 0$. Define $\tilde{c}_2, \tilde{d}_1, \tilde{d}_2$ accordingly. Let $C = \max(\tilde{c}_1, \tilde{c}_2)$ and $D = \max(\tilde{d}_1, \tilde{d}_2)$. Note that for Ψ_{micro} to be valid and bounded, $\max(C, D) > 0$. For a given $\epsilon > 0$, we want $\epsilon_1 \leq \frac{(d_0' + d_0''\pi + d_1 \text{TP}(\mathbf{f}) + d_2\gamma(\mathbf{f}))\epsilon}{D(\mathcal{U}(\mathbf{f};\Psi_{\text{micro}};\mathbb{P}) + \epsilon) + C}$, so that we can guarantee $|\Psi_{\text{micro}}(\mathbf{f}) - \mathcal{U}(\mathbf{f}; \Psi_{\text{micro}}; \mathbb{P})| < \epsilon$. Thus for all $N > N'$, for given ϵ and δ, we have shown that $|\Psi_{\text{micro}}(\mathbf{f}) - \mathcal{U}(\mathbf{f}; \Psi_{\text{micro}}; \mathbb{P})| < \epsilon$ with probability at least $1 - \rho$. This completes the proof of the claim.

Assume the estimated $\hat{\eta}_m(x)$ satisfies $\hat{\eta}_m(x) \xrightarrow{p} \eta_m(x)$ as stated in the Theorem (this can be guaranteed by using a suitable class density estimation algorithm). Consider the multilabel classifier $\mathbf{f}_\delta^* = ([[\eta_m(x) > \delta]])_{m=1}^{M}$. Let δ^* be the optimal threshold corresponding to the Bayes optimal $\mathbf{f}_{\Psi_{\text{micro}}}^*$. Because $\hat{\delta}$ is the empirical minimizer on a finite sample, we have $\Psi_{\text{micro}}(\mathbf{f}_{\hat{\delta}}^*) \geq \Psi_{\text{micro}}(\mathbf{f}_{\delta^*}^*)$ on the sample. So:

$$\begin{aligned}
\mathcal{U}_{\Psi_{\text{micro}}}^* - \mathcal{U}(\mathbf{f}_{\hat{\delta}}^*; \Psi_{\text{micro}}, \mathbb{P}) &= \mathcal{U}_{\Psi_{\text{micro}}}^* - \Psi_{\text{micro}}(\mathbf{f}_{\hat{\delta}}^*) + \Psi_{\text{micro}}(\mathbf{f}_{\hat{\delta}}^*) - \mathcal{U}(\mathbf{f}_{\hat{\delta}}^*; \Psi_{\text{micro}}, \mathbb{P}) \\
&\leq \mathcal{U}_{\Psi_{\text{micro}}}^* - \Psi_{\text{micro}}(\mathbf{f}_{\delta^*}^*) + \Psi_{\text{micro}}(\mathbf{f}_{\hat{\delta}}^*) - \mathcal{U}(\mathbf{f}_{\hat{\delta}}^*; \Psi_{\text{micro}}, \mathbb{P}) \\
&\leq 2\sup_{\delta} |\Psi_{\text{micro}}(\mathbf{f}_\delta^*) - \mathcal{U}(\mathbf{f}_\delta^*; \Psi_{\text{micro}}, \mathbb{P})| \qquad (13)
\end{aligned}$$

Now, to conclude the proof, we argue that the last term in the RHS of the inequality above vanishes as $N \to \infty$. For a given $\mathbf{f} \in \mathbb{R}^M$, let $\mathcal{F}_\delta$ denote the class of multilabel classifiers obtained by thresholding $\mathbf{f}$ for some $\delta \in (0, 1)$. Using Lemma 29.1 in [6], for given ρ and $\epsilon_1 > 0$, $\sup_{\mathbf{f} \in \mathcal{F}_\delta} |\widehat{\text{TP}}(\mathbf{f}) - \text{TP}(\mathbf{f})| < \epsilon_1$ with probability at least $1 - \rho/3$. Following similar arguments in the claim above, given $\epsilon > 0$ and ρ, we can choose ϵ_1 and N large enough such that $\sup_\delta |\Psi_{\text{micro}}(\mathbf{f}_\delta^*) - \mathcal{U}(\mathbf{f}_\delta^*; \Psi_{\text{micro}}, \mathbb{P})| < \epsilon$ with probability at least $1 - \rho$. This shows that the RHS of (13) vanishes as $N \to \infty$. The proof of the theorem is complete.

B Appendix B

B.1 Connection to Existing Results

Our results in Section 3 and Algorithm 1 generalize some of the existing results for learning with general performance metrics in binary classification. In particular, when $M = 1$, our multilabel performance metrics considered in Section 3 reduce to the linear-fractional family binary classification metrics studied by Koyejo et al. [9]. The characterization of Bayes optimal in Theorem 2 of Koyejo et al. [9] can be seen as a special case of our Theorem 2. Similarly, the plugin-estimation algorithm of Koyejo et al. [9] can be derived from our Algorithm 1 for the binary case.

More recently, Narasimhan et al. [11] considered generalized performance metrics for multiclass classification and showed that the Bayes optimal (of the EUM utility) can be characterized as a thresholding of the class-conditional probability. We observe that our framework of metrics introduced in Section 2 readily gives rise to the multiclass performance metrics studied by Narasimhan et al. [11] with the additional constraint of a single label choice for each instance. While they show the Bayes optimal and consistency of learning for a different family than the linear-fractional family we consider in this paper, many of the popular metrics including multiclass F-measure and multiclass Jaccard belong in both the families.

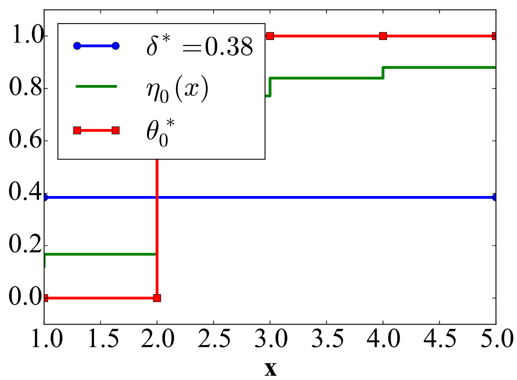
B.2 Multilabel Decision-Theoretic Utility

Existing consistency results and algorithms for multilabel learning focus on the population utility based on decision-theoretic analysis (DTA) defined in (4). Furthermore, we are not aware of consistency results for general multilabel performance metrics. Gao and Zhou [7] define the consistency of multilabel learning with respect to DTA utility. They focus on two specific losses — Hamming and rank loss (the corresponding measures are defined in Section 2) and suggest surrogates for consistent learning with respect to the losses. Note that Hamming loss is a linear metric (i.e. it is linear in the primitives $\widehat{\text{FN}}(\mathbf{f})_{m,n}$ and $\widehat{\text{FP}}(\mathbf{f})_{m,n}$) and hence the Bayes optimal characterizations coincide for both DTA and EUM utility (See Equation (8) of Gao and Zhou [7]). For the Hamming loss, they showed that the Bayes optimal for the DTA utility depends on pairwise conditional distributions, i.e. $\mathbb{P}(Y_m, Y_{m'}|x)$. However, Dembczynski et al. [2] subsequently showed that the Bayes optimal for the un-normalized ranking loss indeed depends only on the label marginals $\mathbb{P}(Y_m|x)$, and in turn showed that minimizing appropriately weighted univariate loss functions (such as exponential and logistic losses) independently for M labels is consistent with respect to rank loss. A subtle but important aspect of the definition of rank loss in the existing literature including [7] and [2] is that the Bayes optimal is allowed to be a real-valued function and may not necessarily correspond to an explicit label decision (note that our definition of Bayes optimal is inherently binary valued). Also, we achieve consistent learning with respect to multilabel EUM utility using a plugin-estimation algorithm. Similar plugin-estimators have been studied in the context of multilabel DTA utility maximization but only for specific metrics. [4] proposed a novel plug-in rule algorithm for estimating the parameters required for a Bayes-optimal prediction for F-measure (i.e., with respect to the DTA utility) via a set of multinomial regression models. Beyond theoretical analysis, further empirical study of the limits of the plug-in approach as compared to modern multilabel algorithms would be illuminating.

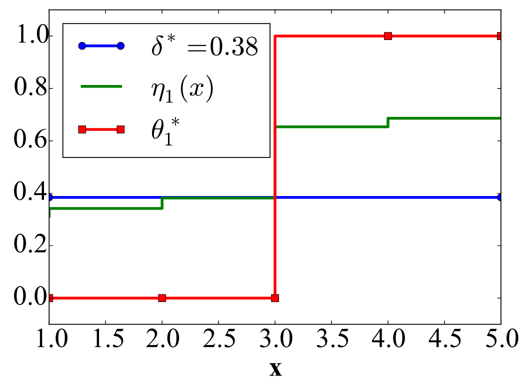
C Appendix C

C.1 Simulated Data Results

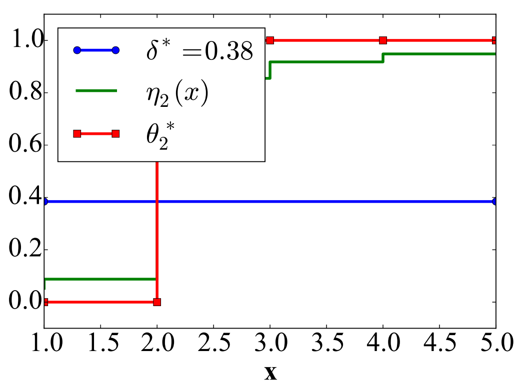
The figure shows larger plots illustrating the Bayes optimal classifier for multilabel F_1 measure on synthetic data with 4 labels, and distribution supported on 5 instances. Plots from left to right show the bayes optimal classifier prediction for instances, for labels 1 through 4. Note that the optimal δ^* at which the label-wise marginal $\eta_m(x)$ is thresholded is shared, conforming to Theorem 2



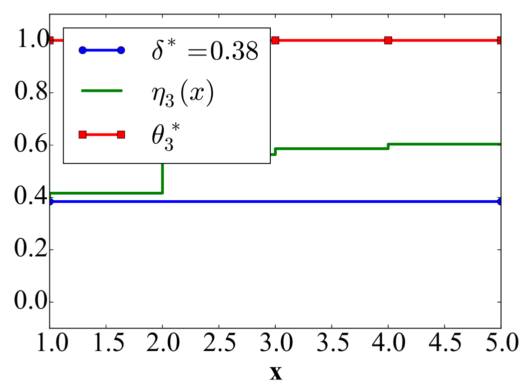
(a)



(b)



(c)



(d)