JMLR: Workshop and Conference Proceedings 45:301–316, 2015 ACML 2015

Surrogate regret bounds for generalized classification performance metrics

Wojciech Kotłowski WKOTLOWSKI@CS.PUT.POZNAN.PL
Poznań University of Technology, Poland

Krzysztof Dembczyński KDEMBCZYNSKI@CS.PUT.POZNAN.PL
Poznań University of Technology, Poland

Editor: Geoffrey Holmes and Tie-Yan Liu

Abstract

We consider optimization of generalized performance metrics for binary classification by means of surrogate loss. We focus on a class of metrics, which are linear-fractional functions of the false positive and false negative rates (examples of which include F_β-measure, Jaccard similarity coefficient, AM measure, and many others). Our analysis concerns the following two-step procedure. First, a real-valued function f is learned by minimizing a surrogate loss for binary classification on the training sample. It is assumed that the surrogate loss is a strongly proper composite loss function (examples of which include logistic loss, squared-error loss, exponential loss, etc.). Then, given f, a threshold $\hat{\theta}$ is tuned on a separate validation sample, by direct optimization of the target performance measure. We show that the regret of the resulting classifier (obtained from thresholding f on $\hat{\theta}$) measured with respect to the target metric is upperbounded by the regret of f measured with respect to the surrogate loss. Our finding is further analyzed in a computational study on both synthetic and real data sets.

Keywords: Generalized performance metric, regret bound, surrogate loss function, binary classification, F-measure, Jaccard similarity, AM measure.

1. Introduction

In binary classification, misclassification error is not necessarily an adequate evaluation metric, and one often resorts to more complex metrics, better suited for the problem. For instance, when the classes are imbalanced, F_β-measure (Lewis, 1995; Jansche, 2005; Nan et al., 2012) and AM measure (balanced eror rate) (Menon et al., 2013) are frequently used. Optimizing such generalized performance metrics poses computational and statistical challenges, as they cannot be decomposed into losses on individual observations.

In this paper, we consider optimization of generalized performance metrics by means of surrogate loss. We restrict our attention to a family of performance metrics which are ratios of linear functions of false positives (FP) and false negatives (FN). Such functions are called linear-fractional, and include the aforementioned F_β and AM measures, as well as Jaccard similarity coefficient, weighted accuracy, and many others. We focus on the most popular approach to optimizing generalized performance metrics in practice, based on the following two-step procedure. First, a real-valued function f is learned by minimizing a surrogate loss for binary classification on the training sample. Then, given f, a threshold $\hat{\theta}$ is tuned

on a separate validation sample, by direct optimization of the target performance measure with respect to a classifier obtained from f by thresholding at $\hat{\theta}$. This approach can be motivated by the asymptotic analysis: minimization of appropriate surrogate loss results in estimation of conditional ("posterior") class probabilities, and many performance metrics are maximized by a classifier which predicts by thresholding on the scale of conditional probabilities (Nan et al., 2012; Zhao et al., 2013; Koyejo et al., 2014). However, it is unclear what can be said about the behavior of this procedure on finite samples.

In this paper, we are interested in theoretical analysis and justification of this approach for any sample size, and for any, not necessarily perfect, classification function. To this end, we use the notion of *regret* with respect to some evaluation metric, which is a difference between the performance of a given classifier and the performance of the optimal classifier with respect to this metric. We show that the regret of the resulting classifier (obtained from thresholding f on $\hat{\theta}$) measured with respect to the target measure is upperbounded by the regret of f measured with respect to the surrogate loss. Our result holds for any surrogate loss function, which is *strongly proper composite* loss function (Agarwal, 2014), examples of which include logistic loss, squared-error loss, exponential loss, etc. Interestingly, the proof of our result goes by an intermediate bound of the regret with respect to the target measure by a cost-sensitive classification regret. As a byproduct, we get a bound on the cost-sensitive classification regret by a surrogate regret of a real-valued function which holds *simultaneously* for *all* misclassification costs: the misclassification costs only influence the threshold, but not: the function, the surrogate loss, or the regret bound. Our finding is further analyzed in a computational study on both synthetic and real data sets.

We note that the goal of this paper is not to propose a new learning algorithm, but rather to provide a deeper statistical understanding of an existing method. The two-stage procedure is commonly used in the binary classification with generalized performance metrics, but this is exactly the reason why we think it is important to study this method in more depth from a theoretical point of view.

Related work. Existing theoretical work on generalized performance metrics is mainly concerned with *statistical consistency* also known as *calibration*, which determines whether convergence to the minimizer of a surrogate loss implies convergence to the minimizer of the task performance measure as sample size goes to infinity (Nan et al., 2012; Zhao et al., 2013; Narasimhan et al., 2014; Koyejo et al., 2014). Here we give a stronger result which bounds the regret with respect to the performance metric by the regret with respect to the surrogate loss. Our result is valid for all finite sample sizes and informs about the rates of convergence.

Parambath et al. (2014) present an alternative approach to maximizing linear-fractional measures by learning a sequence of binary classification problems with varying misclassification costs. While we were inspired by their theoretical analysis, their approach is, however, more complicated than the two-step approach analyzed here, which requires solving an ordinary binary classification problem only once. Moreover, as part of our proof, we show that by minimizing a strongly proper composite loss, we are *implicitly* minimizing cost-sensitive classification error for any misclassification costs without any overhead. Hence, the costs need not be known during learning, and can only be determined later on a separate validation sample by optimizing the threshold.

Outline. The paper is organized as follows. In Section 2 we introduce basic concepts, definitions and notation. The main result is presented in Section 3 and proved in Section 4. The theoretical contribution of the paper is complemented by computational experiments in Section 5, prior to concluding with a summary in Section 6.

2. Problem setting

Binary classifier. In binary classification, the goal is, given an input (feature vector) $x \in X$, to accurately predict the output (label) $y \in \{-1, 1\}$. We assume input-output pairs (x, y) are generated i.i.d. according to $\Pr(x, y)$. A *classifier* is a mapping $h\colon X \to \{-1, 1\}$. Given h, we define the following four quantities:

$$\begin{aligned}
\mathrm{TP}(h) &= \Pr(h(x) = 1 \wedge y = 1),\\
\mathrm{FP}(h) &= \Pr(h(x) = 1 \wedge y = -1),\\
\mathrm{TN}(h) &= \Pr(h(x) = -1 \wedge y = -1),\\
\mathrm{FN}(h) &= \Pr(h(x) = -1 \wedge y = 1),
\end{aligned}$$

which are known as *true positives*, *false positives*, *true negatives* and *false negatives*, respectively. Note that for any h, $\mathrm{FP}(h) + \mathrm{TN}(h) = \Pr(y = -1)$ and $\mathrm{TP}(h) + \mathrm{FN}(h) = \Pr(y = 1)$, so out of the four quantities above, only two are independent. In this paper, we use the convention to parameterize all measures by means of $\mathrm{FP}(h)$ and $\mathrm{FN}(h)$.

Generalized classification performance metrics. We call a two-argument function $\Psi = \Psi(\mathrm{FP}, \mathrm{FN})$ a *(generalized) classification performance metric*. Given a classifier h, we define $\Psi(h) = \Psi(\mathrm{FP}(h), \mathrm{FN}(h))$. Throughout the paper we assume that $\Psi(\mathrm{FP}, \mathrm{FN})$ is *linear-fractional*, i.e. is a ratio of linear functions:

$$\Psi(\mathrm{FP}, \mathrm{FN}) = \frac{a_0 + a_1 \mathrm{FP} + a_2 \mathrm{FN}}{b_0 + b_1 \mathrm{FP} + b_2 \mathrm{FN}}, \tag{1}$$

where we allow coefficients a_i, b_i to depend on the distribution $\Pr(x, y)$.[1] We also assume $\Psi(\mathrm{FP}, \mathrm{FN})$ is non-increasing in FP and non-increasing in FN, a property that is possessed by virtually all performance measures used in practice. Table 1 lists three popular examples of linear-fractional performance metrics.

Let h^*_Ψ be the maximizer of $\Psi(h)$ over all classifiers:[2]

$$h^*_\Psi = \operatorname*{arg\,max}_{h \in \{-1,1\}^X} \Psi(h).$$

Given any classifier h, we define its Ψ*-regret* as a distance of h from the optimal h^*_Ψ measured by means of Ψ:

$$\mathrm{Reg}_\Psi(h) = \Psi(h^*_\Psi) - \Psi(h).$$

1. Note that $\Psi(\mathrm{FP}, \mathrm{FN})$ can be reparameterized to be a function of $(\mathrm{FP}, \mathrm{TN})$, $(\mathrm{TP}, \mathrm{FN})$, or $(\mathrm{TP}, \mathrm{TN})$, and will remain linear-fractional in all these parameterizations.
2. If h^*_Ψ is not unique, take any maximizer.

metric	expression
F_β-measure	$F_\beta = \frac{(1+\beta^2)(P-\mathrm{FN})}{(1+\beta^2)P-\mathrm{FN}+\mathrm{FP}}$
Jaccard similarity	$J = \frac{P-\mathrm{FN}}{P+\mathrm{FP}}$
AM measure	$\mathrm{AM} = \frac{2P(1-P)-(1-P)\mathrm{FN}-P\,\mathrm{FP}}{2P(1-P)}$

Table 1: Some popular linear-fractional performance measures expressed as functions of FN and FP. P abbreviates $\Pr(y=1)$. See (Koyejo et al., 2014) for a more detailed description.

Strongly proper composite losses. Here we briefly outline the theory of strongly proper composite loss functions. See (Agarwal, 2014) for a more detailed description.

Define a *binary class probability estimation (CPE) loss function* (Reid and Williamson, 2010, 2011) as a function $c\colon \{-1,1\}\times[0,1]\to\mathbb{R}_+$, where $c(y,\widehat{\eta})$ assigns penalty to prediction $\widehat{\eta}$, when the observed label is y. Define the *conditional c-risk* as:[3]

$$\mathrm{risk}_c(\eta,\widehat{\eta}) = \eta c(1,\widehat{\eta}) + (1-\eta)c(-1,\widehat{\eta}),$$

the expected loss of prediction $\widehat{\eta}$ when the label is drawn from a distribution with $\Pr(y=1)=\eta$. We say CPE loss is *proper* if for any $\eta\in[0,1]$, $\eta\in\arg\min_{\widehat{\eta}\in[0,1]}\mathrm{risk}_c(\eta,\widehat{\eta})$. In other words, proper losses are minimized by taking the true class probability distribution as a prediction; hence $\widehat{\eta}$ can be interpreted as probability estimate of η. Define the *conditional c-regret* as:

$$\begin{aligned}\mathrm{reg}_c(\eta,\widehat{\eta}) &= \mathrm{risk}_c(\eta,\widehat{\eta}) - \inf_{\widehat{\eta}'}\mathrm{risk}_c(\eta,\widehat{\eta}')\\ &= \mathrm{risk}_c(\eta,\widehat{\eta}) - \mathrm{risk}_c(\eta,\eta),\end{aligned}$$

the difference between the conditional c-risk of $\widehat{\eta}$ and the optimal c-risk. We say a CPE loss c is *λ-strongly proper* if for any $\eta,\widehat{\eta}$:

$$\mathrm{reg}_c(\eta,\widehat{\eta}) \geq \frac{\lambda}{2}(\eta-\widehat{\eta})^2,$$

i.e. the conditional c-regret is everywhere lowerbounded by a squared difference of its arguments. It can be shown (Agarwal, 2014) that under mild regularity assumption a proper CPE loss c is λ-strongly proper if and only if the function $H_c(\eta) := \mathrm{risk}_c(\eta,\eta)$ is λ-strongly concave. This fact lets us easily verify whether a given loss function is λ-strongly proper.

It is often more convenient to reparameterize the loss function from $\widehat{\eta}\in[0,1]$ to a real-valued $f\in\mathbb{R}$ through a strictly increasing (and therefore invertible) *link function* $\psi\colon[0,1]\to\mathbb{R}$:

$$\ell(y,f) = c\big(y,\psi^{-1}(f)\big).$$

3. Throughout the paper, we follow the convention that all conditional quantities are lowercase (regret, risk), while all unconditional quantities are uppercase (Regret, Risk).

If c is λ-strongly proper, we call function $\ell\colon \{-1,1\}\times\mathbb{R}\to\mathbb{R}_+$ *λ-strongly proper composite loss function*. The notions of conditional ℓ-risk $\mathrm{risk}_\ell(\eta,f)$ and conditional ℓ-regret $\mathrm{reg}_\ell(\eta,f)$ extend naturally to the case of composite losses:

$$\begin{aligned}\mathrm{risk}_\ell(\eta,f) &= \eta\ell(1,f)+(1-\eta)\ell(-1,f)\\ \mathrm{reg}_\ell(\eta,f) &= \mathrm{risk}_\ell(\eta,f)-\inf_{f'\in\mathbb{R}}\mathrm{risk}_\ell(\eta,f')\\ &= \mathrm{risk}_\ell(\eta,f)-\mathrm{risk}_\ell(\eta,\psi(\eta)).\end{aligned}$$

and the strong properness of underlying CPE loss implies:

$$\mathrm{reg}_\ell(\eta,f)\geq\frac{\lambda}{2}\left(\eta-\psi^{-1}(f)\right)^2 \tag{2}$$

As an example, consider a *logarithmic scoring rule*[4]: $c(y,\widehat{\eta}) = -[\![y=1]\!]\log\widehat{\eta}-[\![y=-1]\!]\log(1-\widehat{\eta})$. Its conditional risk is given by:

$$\mathrm{risk}_c(\eta,\widehat{\eta}) = -\eta\log\widehat{\eta}-(1-\eta)\log(1-\widehat{\eta}),$$

the *cross-entropy* between η and $\widehat{\eta}$. The conditional c-regret is the binary *Kullback-Leibler divergence* between η and $\widehat{\eta}$:

$$\mathrm{reg}_c(\eta,\widehat{\eta}) = \eta\log\frac{\eta}{\widehat{\eta}}+(1-\eta)\log\frac{1-\eta}{1-\widehat{\eta}}.$$

Note that since $H(\eta)=\mathrm{risk}_c(\eta,\eta)$ is the binary entropy function, and $\left|\frac{\mathrm{d}^2H}{\mathrm{d}\eta^2}\right| = \frac{1}{\eta(1-\eta)}\geq\frac{1}{4}$, c is 4-strongly proper loss. Using the *logit* link function $\psi(\widehat{\eta})=\log\frac{\widehat{\eta}}{1-\widehat{\eta}}$, we end up with the logistic loss function:

$$\ell(y,f)=\log\left(1+e^{-yf}\right),$$

which is 4-strongly proper composite from the definition.

Table 2 presents some of the commonly used losses which are strongly proper composite. Note that the *hinge loss* $\ell(y,f)=(1-yf)_+$, used e.g. in support vector machines (Hastie et al., 2009), is *not* strongly proper composite (even not proper composite).

3. Main result

Given a real-valued function $f\colon X\to\mathbb{R}$, and a λ-strongly proper composite loss $\ell(y,f)$, define the *ℓ-risk* of f as the expected loss of $f(x)$ with respect to the data distribution:

$$\begin{aligned}\mathrm{Risk}_\ell(f) &= \mathbb{E}_{(x,y)}\left[\ell(y,f(x))\right]\\ &= \mathbb{E}_x\left[\mathrm{risk}_\ell(\eta(x),f(x))\right],\end{aligned}$$

where $\eta(x)=\Pr(y=1|x)$. Let f_ℓ^* be the minimizer $\mathrm{Risk}_\ell(f)$ over all functions, $f_\ell^*=\arg\min_f\mathrm{Risk}_\ell(f)$. Since ℓ is proper composite:

$$f_\ell^*(x)=\psi\big(\eta(x)\big).$$

4. $[\![Q]\!]$ is the indicator function, equal to 1 if Q holds, and to 0 otherwise.

loss function	squared-error	logistic	exponential
$\ell(y,f)$	$(y-f)^2$	$\log\left(1+e^{-fy}\right)$	e^{-yf}
$c(1,\widehat{\eta})$	$4(1-\widehat{\eta})^2$	$-\log\widehat{\eta}$	$\sqrt{\frac{1-\widehat{\eta}}{\widehat{\eta}}}$
$c(-1,\widehat{\eta})$	$4\widehat{\eta}^2$	$-\log(1-\widehat{\eta})$	$\sqrt{\frac{\widehat{\eta}}{1-\widehat{\eta}}}$
$\psi(\widehat{\eta})$	$2\widehat{\eta}-1$	$\log\frac{\widehat{\eta}}{1-\widehat{\eta}}$	$\frac{1}{2}\log\frac{\widehat{\eta}}{1-\widehat{\eta}}$
λ	8	4	4

Table 2: Three popular strongly proper composite losses: squared-error, logistic and exponential losses. Shown are the formula $\ell(y,f)$, the underlying CPE loss $c(y,\widehat{\eta})$ with the link function $\psi(\widehat{\eta})$, as well as the strong properness constant λ. See (Agarwal, 2014) for more details and examples.

Define the ℓ-regret of f as:

$$\begin{aligned}\mathrm{Reg}_\ell(f) &= \mathrm{Risk}_\ell(f) - \mathrm{Risk}_\ell(f_\ell^*)\\ &= \mathbb{E}_x\left[\mathrm{risk}_\ell(\eta(x), f(x)) - \mathrm{risk}_\ell(\eta(x), f_\ell^*(x))\right].\end{aligned}$$

Any real-valued function $f\colon X\to\mathbb{R}$ can be turned into a classifier $h_{f,\theta}\colon X\to\{-1,1\}$, by thresholding at some value θ:

$$h_{f,\theta}(x) = \mathrm{sgn}(f(x)-\theta).$$

The purpose of this paper is to address the following problem: given a function f with ℓ-regret $\mathrm{Reg}_\ell(f)$, and a threshold θ, what can we say about Ψ-regret of $h_{f,\theta}$? For instance, can we bound $\mathrm{Reg}_\Psi(h_{f,\theta})$ in terms of $\mathrm{Reg}_\ell(f)$? We give a positive answer to this question, which is based on the following regret bound:

Lemma 1 *Let $\Psi(\mathrm{FP},\mathrm{FN})$ be a linear-fractional function of the form (1), which is non-increasing in* FP *and* FN*. Assume that there exists $\gamma>0$, such that for any classifier $h\colon X\to\{-1,1\}$:*

$$b_0 + b_1\mathrm{FP}(h) + b_2\mathrm{FN}(h) \geq \gamma,$$

i.e. the denominator of Ψ is positive and bounded away from zero. Let ℓ be a λ-strongly proper composite loss function. Then, there exists a threshold θ^*, such that for any real-valued function $f\colon X\to\mathbb{R}$,

$$\mathrm{Reg}_\Psi(h_{f,\theta^*}) \leq C\sqrt{\frac{2}{\lambda}}\sqrt{\mathrm{Reg}_\ell(f)},$$

where $C=\frac{1}{\gamma}\left(\Psi(h_\Psi^*)(b_1+b_2)-(a_1+a_2)\right)>0$.

metric	γ	C
F_β-measure	$\beta^2 P$	$\frac{1+\beta^2}{\beta^2 P}$
Jaccard similarity	P	$\frac{J^*+1}{P}$
AM measure	$2P(1-P)$	$\frac{1}{2P(1-P)}$

Table 3: Constants which appear in the bound of Lemma 1 for several performance metrics.

The proof is quite long and hence is postponed to Section 4. Interestingly, the proof goes by an intermediate bound of the Ψ-regret by a cost-sensitive classification regret. We note that the bound in Lemma 1 is in general unimprovable, in the sense that it is easy to find f, Ψ, ℓ, and distribution $\Pr(x, y)$, for which the bound holds with equality (see proof for details). We split the constant in front of the bound into C and λ, because C depends only on Ψ, while λ depends only on ℓ. Table 3 lists these constants for some popular metrics.

Lemma 1 has the following interpretation. If we are able to find a function f with small ℓ-regret, we are guaranteed that there exists a threshold θ^* such that h_{f,θ^*} has small Ψ-regret. Note that the same threshold θ^* will work for any f, and the right hand side of the bound is *independent* of θ^*. Hence, to minimize the right hand side we only need to minimize ℓ-regret, and we can deal with the threshold afterwards.

Lemma 1 also reveals the form of the optimal classifier h^*_Ψ: take $f = f^*_\ell$ in the lemma and note that $\mathrm{Reg}_\ell(f^*_\ell) = 0$, so that $\mathrm{Reg}_\Psi(h_{f^*_\ell,\theta^*}) = 0$, which means that $h_{f^*_\ell,\theta^*}$ is the minimizer of Ψ:

$$h^*_\Psi(x) = \mathrm{sgn}(f^*_\ell(x) - \theta^*) = \mathrm{sgn}(\eta(x) - \psi^{-1}(\theta^*)),$$

where the second equality is due to $f^*_\ell = \psi(\eta)$ and strict monotonicity of ψ. Hence, h^*_Ψ is a threshold function on η. The proof of Lemma 1 (see Section 4) actually specifies the exact value of the threshold θ^*:

$$\psi^{-1}(\theta^*) = \frac{\Psi(h^*_\Psi)b_1 - a_1}{\Psi(h^*_\Psi)(b_1 + b_2) - (a_1 + a_2)}, \tag{3}$$

which is in agreement with the result obtained by Koyejo et al. (2014).[5]

To make Lemma 1 easier to grasp, consider a special case when $\Psi = \mathrm{FP} + \mathrm{FN}$ is the classification accuracy. In this case, (3) gives $\Psi^{-1}(\theta^*) = 1/2$. Hence, we obtained the well-known result that the classifier maximizing the accuracy is a threshold function on η at $1/2$. Then, Lemma 1 states that given a real-valued f, we should take a classifier h_{f,θ^*} which thresholds f at $\theta^* = \psi(1/2)$ (one can verify that $\theta^* = 0$ for logistic, squared-error and exponential losses). The bounds from the lemma are in this case identical (up to a multiplicative constant) to the bounds by Bartlett et al. (2006).

Unfortunately, in general the optimal threshold θ^* is unknown, as (3) contains an unknown quantity $\Psi(h^*_\Psi)$. The solution in this case is to, given f, directly search for a threshold which maximizes $\Psi(h_{f,\theta})$. This is the main result of the paper:

5. Koyejo et al. (2014) required some continuity assumptions to prove (3). Our analysis shows that these assumptions are not necessary.

Theorem 2 *Given a real-valued function f, let $\hat{\theta} = \arg\max_\theta \Psi(h_{f,\theta})$. Then, under the assumptions and notation from Lemma 1:*

$$\mathrm{Reg}_\Psi(h_{f,\hat{\theta}}) \leq C\sqrt{\frac{2}{\lambda}}\sqrt{\mathrm{Reg}_\ell(f)}.$$

Proof The result follows immediately from Lemma 1: Solving $\max_\theta \Psi(h_{f,\theta})$ is equivalent to solving $\min_\theta \mathrm{Reg}_\Psi(h_{f,\theta})$, and $\min_\theta \mathrm{Reg}_\Psi(h_{f,\theta}) \leq \mathrm{Reg}_\Psi(h_{f,\theta^*})$, where θ^* is the threshold given by Lemma 1. ■

Theorem 2 motivates the following procedure for maximization of Ψ:

1. Find f with small ℓ-regret, e.g. by using a learning algorithm minimizing ℓ-risk on the training sample.
2. Given f, solve $\hat{\theta} = \arg\max_\theta \Psi(h_{f,\theta})$.

Theorem 2 states that the Ψ-regret of the classifier obtained by this procedure is upper-bounded by the ℓ-regret of the underlying real-valued function.

We now shortly discuss how to approach step 2 of the procedure in practice. In principle, this step requires maximizing Ψ defined through FP and FN, which are expectations over an unknown distribution $\Pr(x, y)$. However, as long as Ψ does not change too rapidly (e.g. Ψ has bounded derivatives), it is sufficient to optimize θ on the empirical counterpart of Ψ calculated on a separate validation sample. Step 2 involves optimization within a class of threshold functions (since f is fixed), which has VC-dimension equal to 2 (Devroye et al., 1996). If Ψ has bounded derivatives, there exist constants G_1, G_2 such that:

$$\Psi(\mathrm{FP}, \mathrm{FN}) - \Psi(\widehat{\mathrm{FP}}, \widehat{\mathrm{FN}}) \leq G_1|\mathrm{FP} - \widehat{\mathrm{FP}}| + G_2|\mathrm{FN} - \widehat{\mathrm{FN}}|, \tag{4}$$

where $\widehat{\mathrm{FP}}$ and $\widehat{\mathrm{FN}}$ are empirical counterparts of FP and FN for any given threshold θ. By VC theory, the deviations of $\widehat{\mathrm{FP}}$ from FP, and of $\widehat{\mathrm{FN}}$ from FN can be upperbounded with high probability *uniformly* over the class of all threshold functions by $O(1/\sqrt{m})$, where m is the validation sample size. This and (4) imply the same uniform bound on the deviation of $\Psi(\widehat{\mathrm{FP}}, \widehat{\mathrm{FN}})$ from $\Psi(\mathrm{FP}, \mathrm{FN})$, which in turn implies that the empirical maximizer of Ψ is $O(1/\sqrt{m})$ close to $\max_\theta \Psi(h_{f,\theta})$. Hence, step 2 can be performed within $O(1/\sqrt{m})$ accuracy on a validation sample independent from the training sample.

4. Proof of Lemma 1

The proof can be skipped without affecting the flow of later sections. The proof consists of two steps. First, we bound the Ψ-regret of any classifier h by its cost-sensitive classification regret (introduced below). Next, we show that there exists a threshold θ^*, such that for any f, the cost-sensitive classification regret of h_{f,θ^*} is upperbounded by the ℓ-regret of f.

Bounding Ψ-regret by cost-sensitive regret. Given a real number $\alpha \in [0,1]$, define a *cost-sensitive classification loss* $\ell_\alpha \colon \{-1,1\} \times \{-1,1\} \to \mathbb{R}_+$ as:

$$\ell_\alpha(y, \widehat{y}) = \alpha[\![y = -1]\!][\![\widehat{y} = 1]\!] + (1-\alpha)[\![y = 1]\!][\![\widehat{y} = -1]\!].$$

The cost-sensitive loss assigns different costs of misclassification for positive and negative labels. Given classifier h, the *cost-sensitive risk* of h is:

$$\begin{aligned}\mathrm{Risk}_\alpha(h) &= \mathbb{E}_{(x,y)}[\ell_\alpha(y, h(x))] \\ &= \alpha \mathrm{FP}(h) + (1-\alpha)\mathrm{FN}(h),\end{aligned}$$

and the *cost-sensitive regret* is:

$$\mathrm{Reg}_\alpha(h) = \mathrm{Risk}_\alpha(h) - \mathrm{Risk}_\alpha(h^*_\alpha),$$

where $h^*_\alpha = \arg\min_h \mathrm{Risk}_\alpha(h)$. We now show that there exists α such that for any h,

$$\mathrm{Reg}_\Psi(h) \leq C \mathrm{Reg}_\alpha(h), \tag{5}$$

where C is defined as in the content of Lemma 1. For the sake of clarity, we use a shorthand notation $\Psi = \Psi(h)$, $\Psi^* = \Psi(h^*_\Psi)$, $\mathrm{FP} = \mathrm{FP}(h)$, $\mathrm{FN} = \mathrm{FN}(h)$, $A = a_0 + a_1\mathrm{FP} + a_2\mathrm{FN}$, $B = b_0 + b_1\mathrm{FP} + b_2\mathrm{FN}$ for the numerator and denominator of $\Psi(h)$, and analogously FP^*, FN^*, A^* and B^* for $\Psi(h^*_\Psi)$. In this notation:

$$\begin{aligned}\mathrm{Reg}_\Psi(h) = \Psi^* - \Psi &= \frac{\Psi^* B - A}{B} \\ &= \frac{\Psi^* B - A - \overbrace{(\Psi^* B^* - A^*)}^{=0}}{B} \\ &= \frac{\Psi^*(B - B^*) - (A - A^*)}{B} \\ &= \frac{(\Psi^* b_1 - a_1)\,(\mathrm{FP} - \mathrm{FP}^*) + (\Psi^* b_2 - a_2)\,(\mathrm{FN} - \mathrm{FN}^*)}{B} \\ &\leq \frac{(\Psi^* b_1 - a_1)\,(\mathrm{FP} - \mathrm{FP}^*) + (\Psi^* b_2 - a_2)\,(\mathrm{FN} - \mathrm{FN}^*)}{\gamma},\end{aligned} \tag{6}$$

where the last inequality follows from $B \geq \gamma$ (assumption) and the fact that $\mathrm{Reg}_\Psi(h) \geq 0$ for any h. Since Ψ is non-increasing in TP and FP, we have

$$\frac{\partial \Psi^*}{\partial \mathrm{FP}^*} = \frac{a_1 B^* - b_1 A^*}{(B^*)^2} = \frac{a_1 - b_1 \Psi^*}{B^*} \leq 0,$$

and similarly $\frac{\partial \Psi^*}{\partial \mathrm{FN}^*} = \frac{a_2 - b_2 \Psi^*}{B^*} \leq 0$. This and the assumption $B^* \geq \gamma$ implies that both $\Psi^* b_1 - a_1$ and $\Psi^* b_2 - a_2$ are non-negative, so can be interpreted as misclassification costs. If we normalize the costs by defining:

$$\alpha = \frac{\Psi^* b_1 - a_1}{\Psi^*(b_1 + b_2) - (a_1 + a_2)}, \tag{7}$$

then (6) implies:

$$\begin{aligned}\mathrm{Reg}_\Psi(h) &\leq C\left(\mathrm{Risk}_\alpha(h) - \mathrm{Risk}_\alpha(h^*_\Psi)\right) \\ &\leq C\left(\mathrm{Risk}_\alpha(h) - \mathrm{Risk}_\alpha(h^*_\alpha)\right) = C\mathrm{Reg}_\alpha(h).\end{aligned}$$

Bounding cost-sensitive regret by ℓ-regret. We will show that there exists threshold θ^* such that:

$$\mathrm{Reg}_\alpha(h_{f,\theta^*}) \leq \sqrt{\frac{2}{\lambda}}\sqrt{\mathrm{Reg}_\ell(f)}. \tag{8}$$

This, along with (5) implies Lemma 1. A part of the proof follows the steps of Lemma 4 and Theorem 5 in (Menon et al., 2013), and Theorem 13 in (Agarwal, 2014). First, we will show that (8) holds *conditionally* for every x. To this end, we fix x and deal with $h(x) \in \{-1, 1\}$, $f(x) \in \mathbb{R}$ and $\eta(x) \in [0, 1]$, using a shorthand notation h, f, η.

Given $\eta \in [0, 1]$ and $h \in \{-1, 1\}$, define the *conditional cost-sensitive risk* as:

$$\mathrm{risk}_\alpha(\eta, h) = \alpha(1-\eta)[\![h = 1]\!] + (1-\alpha)\eta[\![h = -1]\!].$$

Let $h^*_\alpha = \arg\min_h \mathrm{risk}_\alpha(\eta, h)$. It can be easily verified that:

$$h^*_\alpha = \mathrm{sgn}(\eta - \alpha). \tag{9}$$

Define the *conditional cost-sensitive regret* as

$$\mathrm{reg}_\alpha(\eta, h) = \mathrm{risk}_\alpha(\eta, h) - \mathrm{risk}_\alpha(\eta, h^*_\alpha).$$

Note that if $h = h^*_\alpha$, then $\mathrm{reg}_\alpha(\eta, h) = 0$. Otherwise, $\mathrm{reg}_\alpha(\eta, h) = |\eta - \alpha|$, so that:

$$\mathrm{reg}_\alpha(\eta, h) = [\![h \neq h^*_\alpha]\!]|\eta - \alpha|.$$

Now assume $h = \mathrm{sgn}(\widehat{\eta} - \alpha)$ for some $\widehat{\eta}$, i.e. h is of the same form as h^*_α in (9), with η replaced by $\widehat{\eta}$. We show that for such h,

$$\mathrm{reg}_\alpha(\eta, h) \leq |\eta - \widehat{\eta}|. \tag{10}$$

This statement trivially holds when $h = h^*_\alpha$. If $h \neq h^*_\alpha$, then η and $\widehat{\eta}$ are on the opposite sides of α (i.e. either $\eta \geq \alpha$ and $\widehat{\eta} < \alpha$ or $\eta < \alpha$ and $\widehat{\eta} \geq \alpha$), hence $|\eta - \alpha| \leq |\eta - \widehat{\eta}|$, which proves (10).

Now, we set the threshold to $\theta^* = \psi(\alpha)$, so that given $f \in \mathbb{R}$,

$$h_{f,\theta^*} = \mathrm{sgn}(f - \theta^*) = \mathrm{sgn}(f - \psi(\alpha)) = \mathrm{sgn}(\psi^{-1}(f) - \alpha),$$

due to strict monotonicity of ψ. Using (10) with $h = h_{f,\theta^*}$ and $\widehat{\eta} = \psi^{-1}(f)$ gives:

$$\begin{aligned}\mathrm{reg}_\alpha(\eta, h_{f,\theta^*}) &\leq |\eta - \psi^{-1}(f)| = \sqrt{(\eta - \psi^{-1}(f))^2} \\ &\leq \sqrt{\frac{2}{\lambda}}\sqrt{\mathrm{reg}_\ell(\eta, f)},\end{aligned} \tag{11}$$

and the last inequality follows from strong properness (2).

To prove the unconditional statement (8), we take expectation with respect to x on both sides of (11):

$$\begin{aligned}\mathrm{Reg}_\alpha(\eta, h_{f,\theta^*}) &= \mathbb{E}_x\left[\mathrm{reg}_\alpha(\eta, h_{f,\theta^*}(x))\right] \\ \text{(by (11))} \qquad &\leq \sqrt{\frac{2}{\lambda}}\mathbb{E}_x\left[\sqrt{\mathrm{reg}_\ell(\eta(x), f(x))}\right] \\ &\leq \sqrt{\frac{2}{\lambda}}\sqrt{\mathbb{E}_x\left[\mathrm{reg}_\ell(\eta(x), f(x))\right]} \\ &= \sqrt{\frac{2}{\lambda}}\sqrt{\mathrm{Reg}_\ell(f)},\end{aligned}$$

where the second inequality is from Jensen's inequality applied to the concave function $x \mapsto \sqrt{x}$. Note that the proof actually specifies the exact value of the universal threshold, $\theta^* = \psi(\alpha)$, where α is given by (7).

The bound in Lemma 1 is unimprovable in a sense that there exist f, Ψ, ℓ, and distribution $\Pr(x, y)$ for which the bound is tight. To see this, take, for instance, squared error loss $\ell(y, f) = (y - f)^2$ and classification accuracy metric $\Psi(\mathrm{FP}, \mathrm{FN}) = 1 - \mathrm{FP} - \mathrm{FN}$. The constants in Lemma 1 are equal to $\gamma = 1$, $C = 2$, and $\lambda = 8$ (see Table 1), while the optimal threshold is $\theta^* = 0$. The bound then simplifies to

$$\mathrm{Reg}_{0/1}(\mathrm{sgn}(f)) \leq \sqrt{\mathrm{Reg}_{\mathrm{sqr}}(f)},$$

which is known to be tight (Bartlett et al., 2006).

5. Empirical results

We perform experiments on synthetic and benchmark data to empirically study the two-step procedure analyzed in the previous sections. We use logistic loss in this procedure as a surrogate loss. Recall that logistic loss is 4-strongly proper composite (see Table 2). We compare its performance with *hinge loss*, which is even *not* a proper composite function. As our task performance metrics, we take the F-measure (F_β-measure with $\beta = 1$) and the AM measure. We could also use the Jaccard similarity coefficient; it turns out, however, that the threshold optimized for the F-measure coincides with the optimal threshold for the Jaccard similarity coefficient, so the latter measure does not give anything substantially different than the F-measure.

The purpose of this study is *not* about comparing the two-step approach with alternative methods; this has already been done in the previous work on the subject, see, e.g., (Nan et al., 2012; Parambath et al., 2014). We also note that similar experiments have been performed in the cited papers on the statistical consistency of generalized performance metrics (Koyejo et al., 2014; Narasimhan et al., 2014; Parambath et al., 2014). Therefore, we unavoidably repeat some of the results obtained therein, but the main novelty of the experiments reported here is that we emphasize the difference between proper composite losses and non-proper losses.

5.1. Synthetic data

We performed two experiments on synthetic data. The first experiment deals with a discrete domain in which we learn within a class of all possible classifiers. The second experiment concerns continuous domain in which we learn within a restricted class of linear functions.

First experiment. We let the input domain X to be a finite set, $X = \{1, 2, \ldots, 25\}$, and take $\Pr(x)$ to be uniform over X. For each $x \in X$, we randomly draw a value of $\eta(x)$ from the uniform distribution on the interval $[0, 1]$. In the first step, we take an algorithm which minimizes a given surrogate loss ℓ within the class of *all* function. Hence, given the training data of size n, the algorithm computes the empirical minimizer of loss ℓ independently for each x. As surrogate losses, we use logistic and hinge loss. In the second step, we tune the threshold $\hat{\theta}$ on a separate validation set, also of size n. For each n, we repeat the procedure

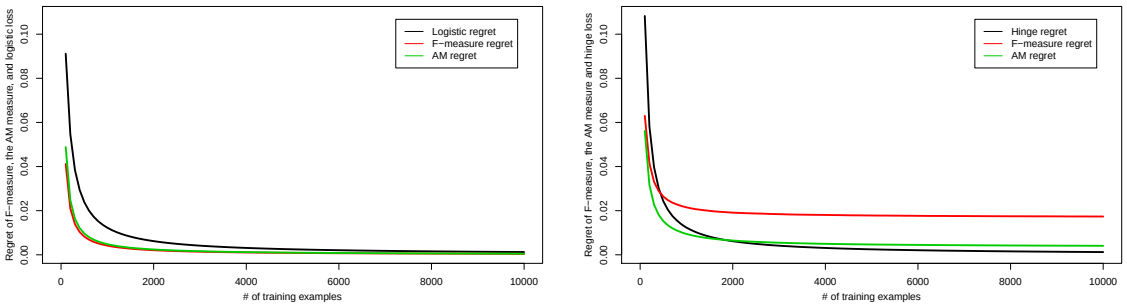

Figure 1: Regret (averaged over 100,000 repetitions) on the discrete synthetic model as a function of the number of training examples. Left panel: logistic loss is used as a surrogate loss. Right panel: hinge loss is used as surrogate loss.

100,000 times, averaging over samples and over models (different random choices of $\eta(x)$). We start with $n = 100$ and increase the number of training examples up to $n = 10,000$. The ℓ-regret and Ψ-regret can be easily computed, as the distribution is known and X is discrete.

The results are given in Fig. 1. The ℓ-regret goes down to zero for both surrogate losses, which is expected, since this is the objective function minimized by the algorithm. Minimization of logistic loss (left plot) gives vanishing Ψ-regret for both the F-measure and the AM measure, as predicted by Theorem 2. In contrast, minimization of the hinge loss (right plot) is suboptimal for both task metrics and gives non-zero Ψ-regret even in the limit $n \to \infty$. This behavior can easily be explained by the fact that hinge loss is not a proper (composite) loss: the risk minimizer for hinge loss is given by $f^*_\ell(x) = \mathrm{sgn}(\eta(x) - 1/2)$ (Bartlett et al., 2006). Hence, the hinge loss minimizer is already a threshold function on $\eta(x)$, with the threshold value set to $1/2$. If, for a given performance metric Ψ, the optimal threshold θ^* is different than $1/2$, the hinge loss minimizer will necessarily have suboptimal Ψ-risk. This is clearly visible for the F-measure. The better result on the AM measure is explained by the fact that the average optimal threshold over all models is 0.5 for this measure, so the minimizer of hinge loss is not that far from the minimizer of AM measure.

Second experiment. We take $X = \mathbb{R}^2$ and generate $x \in X$ from a standard Gaussian distribution. We use a logistic model of the form $\eta(x) = \frac{1}{1+\exp(-a_0 - a^\top x)}$. The weights $a = (a_1, a_2)$ and a_0 are also drawn from a standard Gaussian. For a given model (set of weights), we take training sets of increasing size from $n = 100$ up to $n = 3000$, using 20 different sets for each n. We also generate one test set of size 100,000. For each n, we use $2/3$ of the training data to learn a linear model $f(x) = w_0 + w^\top x$, using either support vector machines (SVM, with linear kernel) or logistic regression (LR). We use implementation of these algorithms from the LibLinear package Fan et al. (2008).[6] The remaining $1/3$ of the training data is used for tuning the threshold. We average the results over 20 different models.

6. Software available at http://www.csie.ntu.edu.tw/~cjlin/liblinear

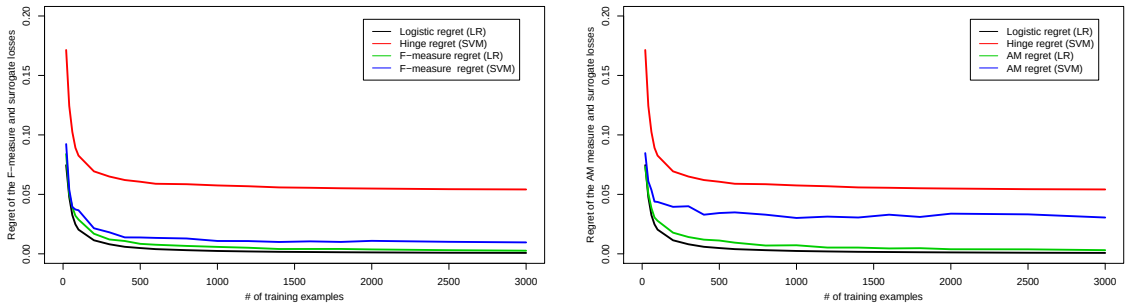

Figure 2: Regret (averaged over 20 x 20 = 400 repetitions) on the logistic model as a function of the number of training examples. Left panel: regret with respect to the F-measure and surrogate losses. Right panel: regret with respect to the AM measure and surrogate losses.

The results are given in Fig. 2. The results obtained for LR (logistic loss minimizer) agree with our theoretical analysis: the ℓ-regret and Ψ-regret with respect to both F-measure and AM measure go to zero. This is expected, as the data generating model is a linear logistic model, and thus coincides with a class of functions over which we optimize. The situation is different for SVM (hinge loss minimizer). Firstly, the ℓ-regret for hinge loss does not converge to zero. This is because the risk minimizer for hinge loss is a threshold function $\text{sgn}(\eta(x) - 1/2)$, and it is not possible to approximate such a function with linear model $f(x) = w_0 + w^\top x$. Hence, even when $n \to \infty$, the empirical hinge loss minimizer (SVM) does not converge to the risk minimizer. This behavior, however, can be *advantageous* for SVM in terms of the task performance measures. This is because the risk minimizer for hinge loss, a threshold function on $\eta(x)$ with the threshold value $1/2$, will perform poorly, for example, in terms of the F-measure and AM measure, for which the optimal threshold θ^* is usually very different from $1/2$. In turn, the linear model constraint will prevent convergence to the risk minimizer, and the resulting linear function $f(x) = w_0 + w^\top x$ will often be close to some reversible function of $\eta(x)$; hence after tuning the threshold, we will often end up close to the minimizer of a given task performance measure. This is seen for the F-measure on the left panel in Fig. 2. In this case, the F-regret of SVM gets quite close to zero, but is still worse than LR. The non-vanishing regret is mainly caused by the fact that for some models with imbalanced class priors, SVM reduce weights w to zero and sets the intercept w_0 to 1 or -1, predicting the same value for all $x \in X$ (this is not caused by a software problem, it is how the empirical loss minimizer behaves). Interestingly, the F-measure is only slightly affected by this pathological behavior of empirical hinge loss minimizer. In turn, the AM measure, for which the plots are drawn in the right panel in Fig. 2, is not robust against this failure of SVM and the model gets the highest possible regret in this case.

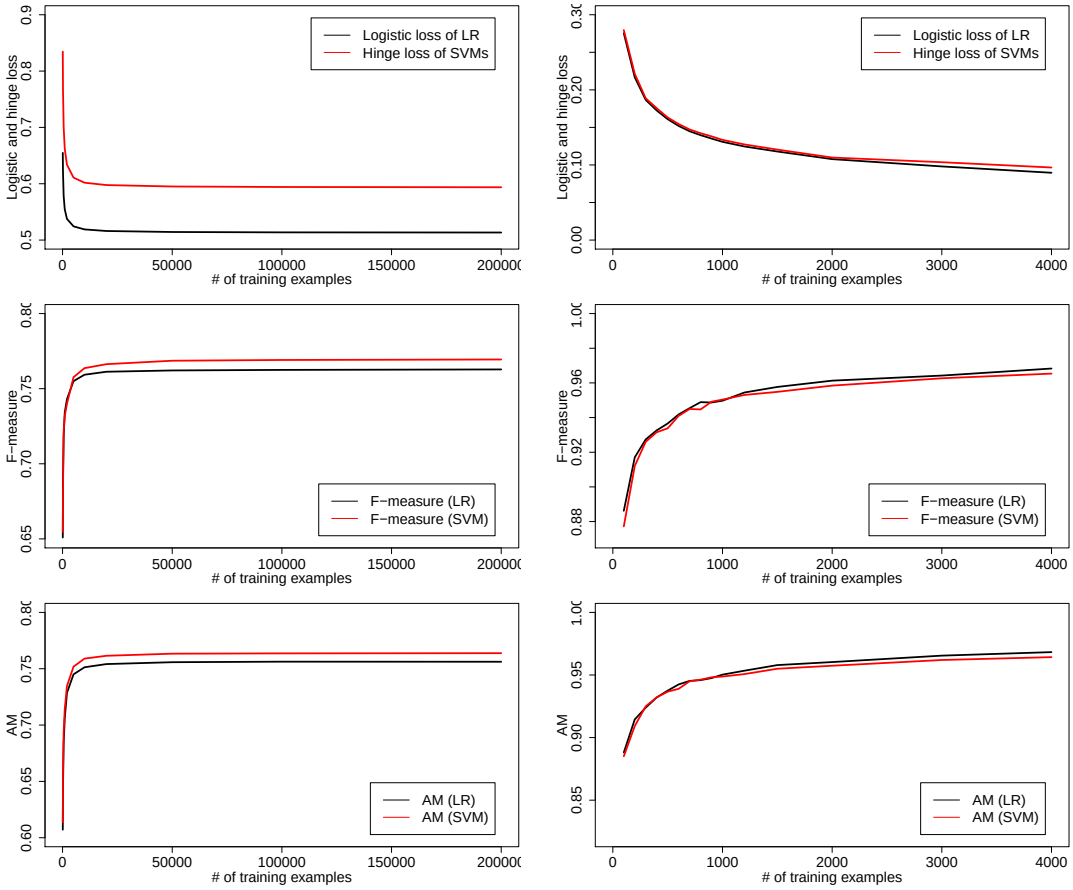

Figure 3: Average test set performance on benchmark data sets. Left panel: `covtype` dataset. Right panel: the `gisette` dataset. The top plots show logistic and hinge loss, the center plots show the F-measure, the bottom plots show the AM measure.

dataset	#examples	#features
covtype.binary	581,012	54
gisette	7,000	5,000

Table 4: Basic statistics for benchmark datasets

5.2. Benchmark data

We also performed a similar experiment on two binary benchmark datasets,[7] described in Table 5.2. We randomly take out a test set of size 181,012 for `covtype`, and of size 3,000

7. Datasets taken from LibSVM repository: http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets

for `gisette`. We use the remaining examples for training. As before, we incrementally increase the size of the training set. We use 2/3 of training examples for learning linear model with SVM or LR, and the rest for tuning the threshold. We repeat the experiment (random train/validation/test split) 20 times. The results are plotted in Fig 3. Since the data distribution is unknown, we are unable to compute the risk minimizers, hence we plot the average loss/metric on the test set rather than the regret. The results show that SVM perform better on the `covtype` dataset, while LR performs better on the `gisette` dataset. However, there is very little difference in performance of SVM and LR in terms of the F-measure and the AM measure on these data sets. We suspect this is due to the fact that $\eta(x)$ function is very different from linear for these problems, so that neither LR nor SVM converge to the ℓ-risk minimizer, and Theorem 2 does not apply. Further studies would be required to understand the behavior of surrogate losses in this case.

6. Summary

We presented a theoretical analysis of a two-step approach to optimize classification performance metrics, which first learns a real-valued function f on a training sample by minimizing a surrogate loss, and then tunes the threshold on f by optimizing the target performance metric on a separate validation sample. We showed that if the metric is a linear-fractional function, and the surrogate loss is strongly proper composite, then the regret of the resulting classifier (obtained from thresholding real-valued f) measured with respect to the target metric is upperbounded by the regret of f measured with respect to the surrogate loss. The proof of our result goes by an intermediate bound of the regret with respect to the target measure by a cost-sensitive classification regret. As a byproduct, we get a bound on the cost-sensitive classification regret by a surrogate regret of a real-valued function which holds simultaneously for all misclassification costs. Our finding is back in a computational study on both synthetic and real data sets.

A natural question is whether our results can be generalized to other classification performance metrics, not necessarily of the linear-fractional form (1).

Acknowledgments

Wojciech Kotłowski has been supported by the Polish National Science Centre under grant no. 2013/11/D/ST6/03050. Krzysztof Dembczyński has been supported by the Polish National Science Centre under grant no. 2013/09/D/ST6/03917.

References

Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15:1653–1674, 2014.

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

Luc Devroye, Laszlo Györfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1st edition, 1996.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

Martin Jansche. Maximum expected F-measure training of logistic regression models. In *HLT/EMNLP 2005*, pages 736–743, 2005.

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

David Lewis. Evaluating and optimizing autonomous text classification systems. In *SIGIR 1995*, pages 246–254, 1995.

Aditya K. Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning (ICML)*, 2013.

Ye Nan, Kian Ming Adam Chai, Wee Sun Lee, and Hai Leong Chieu. Optimizing F-measure: A tale of two approaches. In *International Conference on Machine Learning (ICML)*, 2012.

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

Shameem Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet. Optimizing F-measures by cost-sensitive classification. In *Neural Information Processing Systems (NIPS)*, 2014.

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

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

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