Calibrated Surrogate Maximization of Linear-fractional Utility in Binary Classification

Han Bao
The University of Tokyo
RIKEN AIP
tsutsumi@ms.k.u-tokyo.ac.jp

Masashi Sugiyama
RIKEN AIP
The University of Tokyo
sugi@k.u-tokyo.ac.jp

Abstract

Complex classification performance metrics such as the F$_\beta$-measure and Jaccard index are often used, in order to handle class-imbalanced cases such as information retrieval and image segmentation. These performance metrics are not decomposable, that is, they cannot be expressed in a per-example manner, which hinders a straightforward application of M-estimation widely used in supervised learning. In this paper, we consider *linear-fractional metrics*, which are a family of classification performance metrics that encompasses many standard ones such as the F$_\beta$-measure and Jaccard index, and propose methods to directly maximize performances under those metrics. A clue to tackle their direct optimization is a *calibrated surrogate utility*, which is a tractable lower bound of the true utility function representing a given metric. We characterize sufficient conditions which make the surrogate maximization coincide with the maximization of the true utility. Simulation results on benchmark datasets validate the effectiveness of our calibrated surrogate maximization especially if the sample sizes are extremely small.

1 Introduction

Binary classification, one of the main focuses in machine learning, is a problem to predict binary responses for input covariates. Classifiers are usually evaluated by the *classification accuracy*, which is the expected proportion of correct predictions. Since the accuracy cannot evaluate classifiers appropriately under class imbalance (Menon et al., 2013) or in the presence of label noises (Menon et al., 2015), alternative performance metrics have been employed such as the F$_\beta$-measure (van Rijsbergen, 1974; Joachims, 2005; Nan et al., 2012; Koyejo et al., 2014), Jaccard index (Koyejo et al., 2014; Berman et al., 2018), and balanced error rate (BER) (Brodersen et al., 2010; Menon et al., 2013, 2015; Charoenphakdee et al., 2019). Once a performance metric is given, it is a natural strategy to optimize the performance of classifiers directly under the given performance metric. However, the alternative performance metrics have difficulty in direct optimization in general, because they are non-decomposable, for which per-example loss decomposition is unavailable. In other words, the M-estimation procedure (van de Geer, 2000) cannot be applied, which makes the optimization of non-decomposable metrics hard.

One of the earliest works tackling the non-traditional metrics (Koyejo et al., 2014) generalized performance metrics into the linear-fractional metrics, which are the linear-fractional form of entries in the confusion matrix, and encompasses the BER, F$_\beta$-measure, Jaccard index, Gower-Legendre index (Gower and Legendre, 1986; Natarajan et al., 2016), and weighted accuracy (Koyejo et al., 2014). Koyejo et al. (2014) formulated the optimization problem in two ways. One is a plug-in rule (Koyejo et al., 2014; Narasimhan et al., 2014; Yan et al., 2018) to estimate the class-posterior probability and its optimal threshold, and the other is an iterative weighted empirical risk minimization approach (Koyejo et al., 2014; Parambath et al., 2014) to find a better cost with which the minimizer of the cost-sensitive risk (Scott, 2012) achieves higher utilities. Although they provide statistically consistent esitmators, the former suffers from high sample complexity due to the class-posterior probability estimation, while the latter is computationally demanding because of iterative classifier training.

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

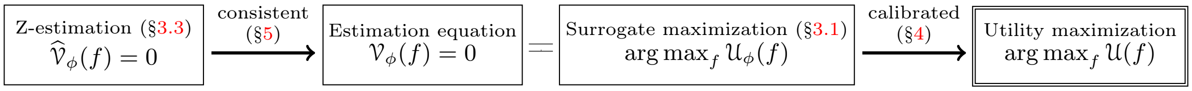
Z-estimation (§3.3)
$\widehat{\mathcal{V}}_\phi(f) = 0$
consistent (§5)
Estimation equation
$\mathcal{V}_\phi(f) = 0$
$=$
Surrogate maximization (§3.1)
$\arg\max_f \mathcal{U}_\phi(f)$
calibrated (§4)
Utility maximization
$\arg\max_f \mathcal{U}(f)$

Figure 1: Overview of this work. Intuitively, we can obtain the utility maximizer by solving $\widehat{\mathcal{V}}_\phi(f) = 0$.

Our goal is to seek for computationally more efficient procedures to directly optimize the linear-fractional metrics, without sacrificing the statistical consistency. Specifically, we provide a novel calibrated surrogate utility which is a tractable lower bound of the true utility representing the metric of our interest. The surrogate maximization is formulated as the combination of concave and quasi-concave programs, which can be optimized efficiently. Then, we derive sufficient conditions on the surrogate calibration, under which the surrogate maximization implies the maximization of the true utility. In addition, we show the consistency of the empirical estimation procedure based on the theory of Z-estimation (van der Vaart, 2000). An overview of our proposed method is illustrated in Fig. 1.

Contributions: (i) *Surrogate calibration* (Sec. 4): We propose a tractable lower bound of the linear-fractional metrics with calibration conditions, which guarantee that the surrogate maximization implies the maximization of the true utility. This approach is model-agnostic differently from many previous approaches (Koyejo et al., 2014; Narasimhan et al., 2014, 2015; Yan et al., 2018). (ii) *Efficient gradient-based optimization* (Secs. 3.2 and 3.3): The surrogate utility has affinity with gradient-based optimization because of its non-vanishing gradient and an unbiased estimator of the gradient direction. Even though the linear-fractional objective does not admit concavity in general, our proposed algorithm is a two-step approach combining concave and quasi-concave programs and hence computationally efficient. (iii) *Consistency analysis* (Sec. 5): The estimator obtained via the surrogate maximization with a finite sample is shown to be consistent to the maximizer of the expected utility.

2 Preliminaries

Throughout this work, we focus on binary classification. Let $[n] \doteq \{1, \dots, n\}$. Let $\mathbb{1}_{\{A\}} \doteq 1$ if the predicate A holds and 0 otherwise. Let $\mathcal{X} \subset \mathbb{R}^d$ be a feature space and $\mathcal{Y} = \{\pm 1\}$ be the label space. We assume that a sample $\mathcal{S} \doteq \{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ independently follows the joint distribution $\mathbb{P}$ with a density p. We often split $\mathcal{S}$ into two independent samples $\mathcal{S}_0 = \{(x_i, y_i)\}_{i=1}^m$ and $\mathcal{S}_1 = \{(x_i, y_i)\}_{i=m+1}^n$. Usually, $m = \lfloor \frac{n}{2} \rfloor$. For a function $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we write $\mathbb{E}[h(X, Y)] = \int_{\mathcal{X}\times\mathcal{Y}} h(X, Y) \mathrm{d}\mathbb{P}$. An expectation with respect to X is written as $\mathbb{E}_X[h(X)] \doteq \int_{\mathcal{X}} h(X) \mathrm{d}\mathbb{P}_{\mathcal{X}}$ for a function $h : \mathcal{X} \to \mathbb{R}$, where $\mathbb{P}_{\mathcal{X}}$ denotes the $\mathcal{X}$-marginal distribution. A classifier is given as a function $f : \mathcal{X} \to \mathbb{R}$, where $\operatorname{sgn}(f(\cdot))$ determines predictions. Here we adopt the convention $\operatorname{sgn}(0) = -1$. Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ be a hypothesis set of classifiers. Let $\pi \doteq \mathbb{P}(Y = +1)$ and $\eta(X) \doteq \mathbb{P}(Y = +1|X)$ be the class-prior/-posterior probabilities of $Y = +1$, respectively. The 0/1-loss is denoted as $\ell(t) \doteq \mathbb{1}_{\{t \le 0\}}$, while $\phi : \mathbb{R} \to \mathbb{R}_{\ge 0}$ denotes a surrogate loss. The norm $\|\cdot\|$ without a subscript is the $\mathbb{L}^2$-norm. For a set $\mathcal{A} \subset \mathcal{F}$, denote $\mathcal{A}^c$ as the complementary set of $\mathcal{A}$, namely, $\mathcal{A}^c \doteq \mathcal{F} \setminus \mathcal{A}$.

The following four quantities are focal targets in binary classification: the true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN).

Definition 1 (Confusion matrix)**.** *Given a classifier $f \in \mathcal{F}$ and a distribution $\mathbb{P}$, its confusion matrix is defined as $C(f, \mathbb{P}) \doteq [\mathsf{TP}, \mathsf{FN}; \mathsf{FP}, \mathsf{TN}]$, where*

$$\begin{aligned}
\mathsf{TP}(f, \mathbb{P}) &\doteq \mathbb{P}(Y = +1, \operatorname{sgn}(f(X)) = +1), \\
\mathsf{FN}(f, \mathbb{P}) &\doteq \mathbb{P}(Y = +1, \operatorname{sgn}(f(X)) = -1), \\
\mathsf{FP}(f, \mathbb{P}) &\doteq \mathbb{P}(Y = -1, \operatorname{sgn}(f(X)) = +1), \\
\mathsf{TN}(f, \mathbb{P}) &\doteq \mathbb{P}(Y = -1, \operatorname{sgn}(f(X)) = -1).
\end{aligned}$$

FN and TP as well as TN and FP can be transformed to each other: $\mathsf{FN}(f, \mathbb{P}) = \pi - \mathsf{TP}(f, \mathbb{P})$ and $\mathsf{TN}(f, \mathbb{P}) = (1 - \pi) - \mathsf{FP}(f, \mathbb{P})$. They can be expressed with ℓ and η, such as $\mathsf{TP}(f, \mathbb{P}) = \mathbb{E}_X[\ell(-f(X))\eta(X)]$. The goal of binary classification is to obtain a classifier that "maximizes" TP and TN, while keeping FP and FN as "low" as possible. Classifiers are evaluated by performance metrics that trade off those four quantities. Performance metrics need to be chosen based on user's preference on the confusion matrix (Sokolova and Lapalme, 2009; Menon et al., 2015). In this work, we focus on the following family of utilities representing the linear-fractional metrics.

Definition 2 (Linear-fractional utility[1])**.** *A linear-fractional utility $\mathcal{U} : \mathcal{F} \to [0, 1]$ is defined as*

$$\mathcal{U}(f) \doteq \frac{\mathbb{E}_X[W_0(f(X), \eta(X))]}{\mathbb{E}_X[W_1(f(X), \eta(X))]}, \tag{1}$$

[1] As mentioned by Dembczyński et al. (2017), there is a dichotomy in the definition of performance metrics: population utility (PU) and expected test utility (ETU). We adopt PU, which is defined as the linear-fractional transform of the population confusion matrix in this context. This is convenient to avoid estimating η compared to ETU.

where $W_0, W_1 : \mathbb{R} \times [0,1] \to \mathbb{R}$ are class-conditional score functions given as

$$W_k(\xi, q) \doteq a_{k,+1}\ell(-\xi)q + a_{k,-1}\ell(-\xi)(1-q) + b_k,$$

and $a_{0,+1} > 0, a_{0,-1} \leq 0, b_0 \in \mathbb{R}, a_{1,+1} \geq 0, a_{1,-1} \geq 0, b_1 \in \mathbb{R}$ are constants such that $0 \leq \mathcal{U}(f) \leq 1$ $(\forall f)$.

The class-conditional score functions correspond to a linear-transformation of TP and FP: $\mathbb{E}_X[W_k(f(X), \eta(X))] = a_{k,+1}\mathsf{TP}(f, \mathbb{P}) + a_{k,-1}\mathsf{FP}(f, \mathbb{P}) + b_k$. Examples of $\mathcal{U}$ are shown in Tab. 1. Given a utility function $\mathcal{U}$, our goal is to obtain a classifier $f^\dagger$ that maximizes $\mathcal{U}$.

$$f^\dagger = \operatorname*{arg\,max}_{f \in \mathcal{F}} \mathcal{U}(f). \tag{2}$$

Traditional Supervised Classification: Here, we briefly review a traditional procedure for supervised classification (Vapnik, 1998). Our aim is to obtain a classifier with high accuracy, which corresponds to minimizing the classification risk $\mathcal{R}(f) \doteq \mathbb{E}[\ell(Yf(X))]$. Since optimizing the 0/1-loss ℓ is a computationally infeasible problem (Ben-David et al., 2003; Feldman et al., 2012), it is a common practice to instead minimize a surrogate risk $\mathcal{R}_\phi(f) \doteq \mathbb{E}[\phi(Yf(X))]$, where $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a surrogate loss. If ϕ is a classification-calibrated loss (Bartlett et al., 2006), it is known that minimizing $\mathcal{R}_\phi$ corresponds to minimizing $\mathcal{R}$. Eventually, what we actually minimize is the empirical (surrogate) risk $\widehat{\mathcal{R}}_\phi(f) \doteq \frac{1}{n}\sum_{i=1}^n \phi(y_i f(x_i))$. The empirical risk $\widehat{\mathcal{R}}_\phi(f)$ is an unbiased estimator of the true risk $\mathcal{R}_\phi(f)$ for a fixed $f \in \mathcal{F}$, and the uniform law of large numbers guarantees that $\widehat{\mathcal{R}}_\phi(f)$ converges to $\mathcal{R}_\phi(f)$ for any $f \in \mathcal{F}$ in probability (Vapnik, 1998; van de Geer, 2000; Mohri et al., 2012). This strategy to minimize $\widehat{\mathcal{R}}_\phi$ is called empirical risk minimization (ERM).

The traditional ERM is devoted to maximizing the accuracy, which is not necessarily suitable when another metric is used for evaluation. Our aim is to give an alternative procedure to maximize $\mathcal{U}$ directly as in Eq. (2). In the next section, we introduce a tractable counterpart of the true utility $\mathcal{U}$ because $\mathcal{U}$ contains the 0/1-loss ℓ and is intractable as $\mathcal{R}_\phi$ above.

3 Surrogate Utility and Optimization

The true utility in Eq. (1) consists of the 0/1-loss ℓ, which is difficult to optimize. In this section, we introduce a surrogate utility in order to make the optimization problem in Eq. (2) easier.

3.1 Lower Bounding True Utility

Assume that we are given a surrogate loss $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$. We hope that the surrogate utility should lower-bound the true utility $\mathcal{U}$, and that TP / FP should become larger / smaller as a result of optimization, respectively. We realize these ideas by constructing surrogate class-conditional score functions $W_{0,\phi}$ and $W_{1,\phi}$ as follows:

$$\begin{aligned} &W_{0,\phi}(\xi, q) \\ &\quad \doteq a_{0,+1}(1 - \phi(\xi))q + a_{0,-1}\phi(-\xi)(1-q) + b_0, \\ &W_{1,\phi}(\xi, q) \\ &\quad \doteq a_{1,+1}(1 + \phi(\xi))q + a_{1,-1}\phi(-\xi)(1-q) + b_1. \end{aligned} \tag{3}$$

We often abbreviate $\mathbb{E}_X[W_{k,\phi}(f(X), \eta(X))]$ as $\mathbb{E}[W_{k,\phi}]$ if it is clear from the context. Given the surrogate class-conditional scores, define the surrogate utility as follows.

$$\mathcal{U}_\phi(f) \doteq \frac{\mathbb{E}_X[W_{0,\phi}(f(X), \eta(X))]}{\mathbb{E}_X[W_{1,\phi}(f(X), \eta(X))]} = \frac{\mathbb{E}[W_{0,\phi}]}{\mathbb{E}[W_{1,\phi}]}. \tag{4}$$

To construct $\mathcal{U}_\phi$, the 0/1-losses appearing in the true utility $\mathcal{U}$ are replaced with the surrogate loss ϕ. The surrogate class-conditional scores in (3) are designed so that the surrogate utility in (4) bounds $\mathcal{U}$ from below.

Lemma 3. *For all f and a surrogate loss $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that $\phi(t) \geq \ell(t)$ for all $t \in \mathbb{R}$, $\mathcal{U}_\phi(f) \leq \mathcal{U}(f)$.*

Proof. Fix $\xi \in \mathbb{R}$ and $q \in [0,1]$. Since $\ell(-\xi) = 1 - \ell(\xi)$, $a_{0,+1}\ell(-\xi) = a_{0,+1}(1-\ell(\xi)) \geq a_{0,+1}(1-\phi(\xi))$ ($\because a_{0,+1} \geq 0$). Together with $a_{0,-1}\ell(-\xi) \geq a_{0,-1}\phi(-\xi)$ ($\because a_{0,-1} \leq 0$), we confirm $W_0(\xi, q) \geq W_{0,\phi}(\xi, q)$. It can be confirmed that $W_1(\xi, q) \leq W_{1,\phi}(\xi, q)$ as well. Hence, $\mathcal{U}(f) \geq \mathcal{U}_\phi(f)$ is easy to see. □

Due to this property, maximizing $\mathcal{U}_\phi$ is at least maximizing a lower bound of $\mathcal{U}$. We will discuss the goodness of this lower bound in Sec. 4, but we can immediately see $\mathcal{U}_\phi(f) (\leq \mathcal{U}(f)) \leq 1$ for any f. In the rest of this paper, we assume that $\mathcal{U}_\phi$ is Fréchet differentiable.

3.2 Tractability of Surrogate Utility

The surrogate utility $\mathcal{U}_\phi$ comes to have a non-vanishing gradient by using a smooth ϕ, and is guaranteed to be a lower bound of $\mathcal{U}$. In this subsection, we discuss how it can be maximized efficiently.

Let us consider an empirical estimator of $\mathcal{U}_\phi$:

$$\widehat{\mathcal{U}}_\phi(f) = \frac{\frac{1}{m}\sum_{i=1}^m \widetilde{W}_{0,\phi}(f(x_i), y_i)}{\frac{1}{n-m}\sum_{i=m+1}^n \widetilde{W}_{1,\phi}(f(x_i), y_i)}, \tag{5}$$

where

$$\widetilde{W}_{0,\phi}(\xi, y) \doteq \begin{cases} a_{0,+1}(1 - \phi(\xi)) + b_0 & \text{if } y = +1, \\ a_{0,-1}\phi(-\xi) + b_0 & \text{if } y = -1, \end{cases}$$

$$\widetilde{W}_{1,\phi}(\xi, y) \doteq \begin{cases} a_{1,+1}(1 + \phi(\xi)) + b_1 & \text{if } y = +1, \\ a_{1,-1}\phi(-\xi) + b_1 & \text{if } y = -1. \end{cases}$$

Table 1: Examples of the linear-fractional performance metrics. $\beta > 0$ is a trade-off parameter for the F_β-measure, while $\alpha \geq 0$ is for the Gower-Legendre index.

Metric	F_β-measure (van Rijsbergen, 1974)	Jaccard index (Jaccard, 1901)	Gower-Legendre index (Gower and Legendre, 1986)
Definition	$\frac{(1+\beta^2)\mathsf{TP}}{(1+\beta^2)\mathsf{TP}+\beta^2\mathsf{FN}+\mathsf{FP}}$	$\frac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FN}+\mathsf{FP}}$	$\frac{\mathsf{TP}+\mathsf{TN}}{\mathsf{TP}+\alpha(\mathsf{FP}+\mathsf{FN})+\mathsf{TN}}$
$(a_{0,+1}, a_{0,-1})$	$(1+\beta^2, 0)$	$(1, 0)$	$(1, -1)$
b_0	0	0	$1-\pi$
$(a_{1,+1}, a_{1,-1})$	$(1, 1)$	$(0, 1)$	$(1-\alpha, \alpha-1)$
b_1	$\beta^2\pi$	π	$1+(\alpha-1)\pi$

A global maximizer of $\widehat{\mathcal{U}}_\phi$ could be efficiently obtained if $\widehat{\mathcal{U}}_\phi$ were concave. However, this is hard to achieve in our case regardless of the choice of ϕ due to its fractional form. Nonetheless, we may formulate our optimization problem as a *quasi-concave* program under a certain condition. First, we introduce the notion of quasi-concavity.

Definition 4 (Quasi-concavity (Boyd and Vandenberghe, 2004))**.** *A function $h : A \to \mathbb{R}$ is said to be quasi-concave if the super-level set $\{x \in A \mid h(x) \geq \alpha\}$ is a convex set for $\forall\alpha \in \mathbb{R}$.*

A quasi-concave function is a generalization of a concave function and has the unimodality though it is not necessarily concave, which ensures the uniqueness of the solution. Next, we have the following result, whose proof is given in App. B. Let $\widehat{\mathcal{U}}^{\mathrm{n}}_\phi(f) \doteq \frac{1}{m}\sum_{i=1}^m \widetilde{W}_{0,\phi}(f(x_i), y_i)$ be the numerator of $\widehat{\mathcal{U}}_\phi$.

Lemma 5. *Let $\bar{\mathcal{F}} \doteq \{f \mid \widehat{\mathcal{U}}^{\mathrm{n}}_\phi(f) \geq 0\} \subset \mathcal{F}$. If ϕ is convex, $\widehat{\mathcal{U}}_\phi$ in Eq. (5) is quasi-concave over $\bar{\mathcal{F}}$ and $\widehat{\mathcal{U}}^{\mathrm{n}}_\phi$ is concave over $\mathcal{F}$.*

From Lemma 5, we observe the following two important facts. First, in the range of $f \notin \bar{\mathcal{F}}$, our objective $\widehat{\mathcal{U}}_\phi$ is generally neither concave nor quasi-concave, but its numerator $\widehat{\mathcal{U}}^{\mathrm{n}}_\phi$ is concave. Second, $\widehat{\mathcal{U}}_\phi$ is quasi-concave over $\bar{\mathcal{F}}$. These observations motivate us to employ Algorithm 1, which first increases the numerator $\widehat{\mathcal{U}}^{\mathrm{n}}_\phi$ only to make it positive and then maximizes the fractional form $\widehat{\mathcal{U}}_\phi$. Since the former is a concave program and the latter is a quasi-concave program within $\bar{\mathcal{F}}$, the entire optimization can be performed computationally efficiently. For quasi-concave optimization, normalized gradient ascent (NGA) (Hazan et al., 2015) is applied, which is guaranteed to find a global solution of quasi-concave objectives. The behavior of Algorithm 1 is illustrated in Figure 2.

```
Algorithm 1: Hybrid Optimization Algorithm
Input  : $\phi$ convex loss, $\theta$ initial classifier
         parameter
while $\widehat{\mathcal{U}}^{\mathrm{n}}_\phi(f_\theta) \leq 0$ do
    $g^{\mathrm{n}} \longleftarrow \nabla_\theta \widehat{\mathcal{U}}^{\mathrm{n}}_\phi(f_\theta)$
    $\theta \longleftarrow$ gradient_based_update($\theta, g^{\mathrm{n}}$)
end
while stopping criterion is not satisfied do
    $g \longleftarrow \nabla_\theta \widehat{\mathcal{U}}_\phi(f_\theta)$, $\widehat{g} = g/\|g\|$
    $\theta \longleftarrow$ gradient_based_update($\theta, \widehat{g}$)
end
Output: maximizer $f_\theta$
```

3.3 Gradient Direction Estimator

The empirical estimator $\widehat{\mathcal{U}}_\phi$ in Eq. (5) is generally biased due to its fractional form. Nevertheless, its gradient $\nabla_f \widehat{\mathcal{U}}_\phi$ is unbiased to the expected gradient $\nabla_f \mathcal{U}_\phi$ up to a positive scalar multiple. Hence, we may safely use $\nabla_f \widehat{\mathcal{U}}_\phi$ as the update direction in NGA.

Below, we state this idea formally. Under the interchangeability of the expectation and derivative, the gradient of the expected utility $\mathcal{U}_\phi$ is expressed as

$$\begin{aligned}
&\nabla_f \mathcal{U}_\phi(f) \\
&= \underbrace{\frac{1}{(\mathbb{E}[W_{1,\phi}])^2}}_{\text{positive scalar}} \underbrace{\mathbb{E}[W_{1,\phi}]\mathbb{E}[\nabla W_{0,\phi}] - \mathbb{E}[W_{0,\phi}]\mathbb{E}[\nabla W_{1,\phi}]}_{\text{gradient direction } (\doteq \mathcal{V}_\phi(f))} \\
&= c\mathcal{V}_\phi(f), \qquad \text{where } c = (\mathbb{E}[W_{1,\phi}])^{-2} > 0,
\end{aligned}$$

from which we can see that its gradient direction is parallel to $\mathcal{V}_\phi$. $\mathcal{V}_\phi$ can be unbiasedly estimated.

Lemma 6. *Denote $\widetilde{W}_{0,\phi}(f(x_i), y_i) = \widetilde{W}_{0,\phi}(z_i)$ for simplicity. Define*

$$\begin{aligned}
\widehat{\mathcal{V}}_\phi(f) \doteq \frac{1}{m(n-m)} \sum_{i=1}^m \sum_{j=m+1}^n \Big\{ &\widetilde{W}_{1,\phi}(z_j)\nabla_f \widetilde{W}_{0,\phi}(z_i) \\
&- \widetilde{W}_{0,\phi}(z_i)\nabla_f \widetilde{W}_{1,\phi}(z_j) \Big\}.
\end{aligned} \tag{6}$$

Then, we have $\mathcal{V}_\phi(f) = \mathbb{E}_{\mathcal{S}}[\widehat{\mathcal{V}}_\phi(f)]$.

Lemma 6 can be confirmed by simple algebra, noting that two samples $\mathcal{S}_0$ and $\mathcal{S}_1$ are independent and identically drawn from $\mathbb{P}$. Since solving $\nabla \widehat{\mathcal{U}}_\phi(f) = 0$ is

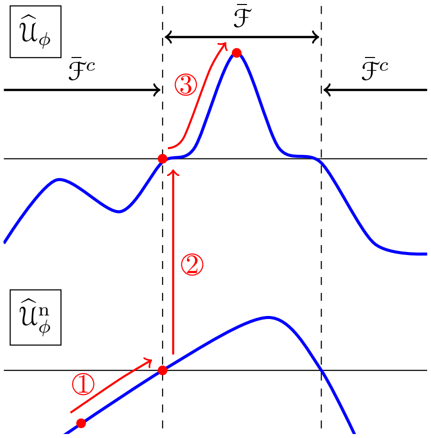
$\widehat{\mathcal{U}}_\phi$
$\bar{\mathcal{F}}$
$\bar{\mathcal{F}}^c$
$\bar{\mathcal{F}}^c$
③
②
$\widehat{\mathcal{U}}^{\mathrm{n}}_\phi$
①

Figure 2: Illustration of our hybrid optimization approach in Algorithm 1. ① maximize the numerator $\widehat{\mathcal{U}}^{\mathrm{n}}_\phi$ (concave), ② once $\widehat{\mathcal{U}}^{\mathrm{n}}_\phi(f) \geq 0$, optimize the fraction $\widehat{\mathcal{U}}_\phi$, ③ maximize the fraction $\widehat{\mathcal{U}}_\phi$ (quasi-concave only in $\bar{\mathcal{F}}$).

Algorithm 2: Normalized Gradient Ascent

```
Input : $\theta$ initial classifier parameter, $\gamma$ learning rate
while stopping criterion is not satisfied do
    $g \longleftarrow \widehat{\mathcal{V}}_\phi(f_\theta)$, $\widehat{g} = g/\|g\|$
    $\theta \longleftarrow \theta + \gamma\widehat{g}$
end
Output: learned classifier parameter $\theta$
```

identical to solving $\widehat{\mathcal{V}}_\phi(f) = 0$, gradient updates using $\nabla\widehat{\mathcal{U}}_\phi$ is aligned to the maximization of $\mathcal{U}_\phi$. Hence, optimization procedures that only need gradients such as gradient ascent and quasi-Newton methods (Boyd and Vandenberghe, 2004), e.g., the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Fletcher, 2013) can be applied to maximize $\mathcal{U}_\phi$. Note that Algorithm 2 can be regarded as an extension of the traditional gradient ascent using $\widehat{\mathcal{V}}_\phi$. We plug either Algorithm 2 or BFGS using the normalized gradient to the second half of Algorithm 1.

4 Calibration Analysis: Bridging Surrogate Utility and True Utility

In Sec. 3, we formulated the tractable surrogate utility. Given the surrogate utility $\mathcal{U}_\phi$, a natural question arises in the same way as the classification calibration in binary classification (Zhang, 2004b; Bartlett et al., 2006): *Does maximizing the surrogate utility $\mathcal{U}_\phi$ imply maximizing the true utility $\mathcal{U}$?* In this section, we study sufficient conditions on the surrogate loss ϕ in order to connect the maximization of $\mathcal{U}_\phi$ and the maximization of $\mathcal{U}$. All proofs in this section are deferred to App. A.

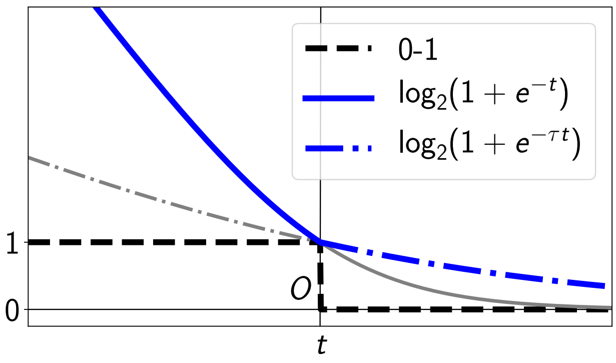

Figure 3: An example of τ-discrepant loss with $\tau > 0$: $\phi(t) = \log_2(1+e^{-t})$ for $t \leq 0$ and $\phi(t) = \log_2(1+e^{-\tau t})$ for $t > 0$.

First, we define the notion of $\mathcal{U}$-*calibration*.

Definition 7 ($\mathcal{U}$-calibration)**.** *The surrogate utility $\mathcal{U}_\phi$ is said to be $\mathcal{U}$-calibrated if for any sequence of measurable functions $\{f_l\}_{l\geq 1}$ and any distribution $\mathbb{P}$, it holds that $\mathcal{U}_\phi(f_l) \to \mathcal{U}^*_\phi \Longrightarrow \mathcal{U}(f_l) \to \mathcal{U}^\dagger$ when $l \to \infty$, where $\mathcal{U}^*_\phi \doteq \sup_f \mathcal{U}_\phi(f)$ and $\mathcal{U}^\dagger \doteq \sup_f \mathcal{U}(f)$ are the suprema taken over all measurable functions.*

This definition is motivated by calibration in other learning problems such as binary classification (Bartlett et al., 2006, Theorem 3), multi-class classification (Zhang, 2004a, Theorem 3), structured prediction (Osokin et al., 2017, Theorem 2), and AUC optimization (Gao and Zhou, 2015, Definition 1). If a surrogate utility is $\mathcal{U}$-calibrated, we may safely optimize the surrogate utility instead of the true utility $\mathcal{U}$. Note that $\mathcal{U}$-calibration is a concept to reduce the surrogate maximization to the maximization of $\mathcal{U}$ *within all measurable functions*. The approximation error of $\mathcal{U}_\phi$ is not the target of our analysis as in Bartlett et al. (2006).

Next, we give a property of loss functions that is needed to guarantee $\mathcal{U}$-calibration.

Definition 8 (τ-discrepant loss)**.** *For a fixed $\tau > 0$, a convex loss function $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is said to be τ-*discrepant *if ϕ satisfies* $\lim_{t\searrow 0} \phi'(t) \geq \tau \lim_{t\nearrow 0} \phi'(t)$.

Intuitively, τ-discrepancy means that the gradient of ϕ around the origin is steeper in the negative domain than the positive domain (see Figure 3). The value τ controls *steepness* of the TP / FP surrogates appearing in the surrogate utility $\mathcal{U}_\phi$. Note that $\phi(\xi)$ and $\phi(-\xi)$ appearing in Eqs. (3) and (4) correspond to TP and FP, respectively, by their constructions.

Below, we see calibration properties for specific linear-fractional metrics, the F_β-measure and Jaccard index. Note that those calibration analyses can be extended to general linear-fractional utilities, which is deferred to App. A.4.

F_β-measure: The F_β-measure has been widely used especially in the field of information retrieval where relevant items are rare (Manning and Schütze, 2008). Since it is defined as the weighted harmonic mean of the precision and recall (see Tab. 1), its optimization is difficult in general. Although much previous work has tried to directly optimize it in the context of the class-posterior probability estimation (Nan et al., 2012; Koyejo et al., 2014; Yan et al., 2018) or the iterative cost-sensitive learning (Koyejo et al., 2014; Parambath et al., 2014), we show that there exists a calibrated surrogate utility that can be used in the direct optimization as well.

For the F_β-measure $\frac{(1+\beta^2)\mathsf{TP}}{(1+\beta^2)\mathsf{TP}+\beta^2\mathsf{FN}+\mathsf{FP}} = \frac{(1+\beta^2)\mathsf{TP}}{\mathsf{TP}+\mathsf{FP}+\beta^2\pi}$, define the true utility $\mathcal{U}^{\mathsf{F}_\beta}$ and the surrogate utility $\mathcal{U}_\phi^{\mathsf{F}_\beta}$ as

$$\mathcal{U}^{\mathsf{F}_\beta}(f) = \frac{\mathbb{E}_X\left[(1+\beta^2)\ell(-f)\eta\right]}{\mathbb{E}_X\left[\ell(-f)\eta + \ell(-f)(1-\eta)+\beta^2\pi\right]},$$
$$\mathcal{U}_\phi^{\mathsf{F}_\beta}(f) = \frac{\mathbb{E}_X\left[(1+\beta^2)(1-\phi(f))\eta\right]}{\mathbb{E}_X\left[(1+\phi(f))\eta + \phi(-f)(1-\eta)+\beta^2\pi\right]}.$$

As for $\mathcal{U}_\phi^{\mathsf{F}_\beta}$, we have the following F_β-calibration guarantee. Denote $(\mathcal{U}_\phi^{\mathsf{F}_\beta})^* \doteq \sup_f \mathcal{U}_\phi^{\mathsf{F}_\beta}(f)$.

Theorem 9 (F_β-calibration)**.** *Assume that a surrogate loss $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is differentiable almost everywhere, convex, and non-increasing, and that $(\mathcal{U}_\phi^{\mathsf{F}_\beta})^* \geq \frac{(1+\beta^2)\tau}{\beta^2-\tau}$ and ϕ is τ-discrepant for some constant $\tau \in (0, \beta^2)$.*[2] *Then, $\mathcal{U}_\phi^{\mathsf{F}_\beta}$ is F_β-calibrated.*

An example of the τ-discrepant surrogate loss is shown in Figure 3. Here τ is a discrepancy hyperparameter. From the fact $(\mathcal{U}_\phi^{\mathsf{F}_\beta})^* \leq 1$, τ ranges over $(0, \frac{\beta^2}{2+\beta^2}]$. We may determine τ by cross-validation, or fix it at $\tau = \frac{\beta^2}{2+\beta^2}$ by assuming $(\mathcal{U}_\phi^{\mathsf{F}_\beta})^* \approx 1$.

Jaccard Index: The Jaccard index, also referred to as the *intersection over union (IoU)*, is a metric of similarity between two sets: For two sets A and B, it is defined as $\frac{|A\cap B|}{|A\cup B|} \in [0,1]$ (Jaccard, 1901). The Jaccard index between the sets of examples predicted as positives and labeled as positives becomes $\frac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FN}+\mathsf{FP}}$, as is shown in Tab. 1. This measure is not only used for measuring the performance of binary classification (Koyejo et al., 2014; Narasimhan et al., 2015), but also for semantic segmentation (Everingham et al., 2010; Csurka et al., 2013; Ahmed et al., 2015; Berman et al., 2018).

For the Jaccard index $\frac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FN}+\mathsf{FP}} = \frac{\mathsf{TP}}{\mathsf{FP}+\pi}$, define the true utility $\mathcal{U}^{\mathsf{Jac}}$ and the surrogate utility $\mathcal{U}_\phi^{\mathsf{Jac}}$ as

$$\mathcal{U}^{\mathsf{Jac}}(f) = \frac{\mathbb{E}_X[\ell(-f)\eta]}{\mathbb{E}_X[\ell(-f)(1-\eta)+\pi]},$$
$$\mathcal{U}_\phi^{\mathsf{Jac}}(f) = \frac{\mathbb{E}_X[(1-\phi(f))\eta]}{\mathbb{E}_X[\phi(-f)(1-\eta)+\pi]}.$$

As for $\mathcal{U}_\phi^{\mathsf{Jac}}$, we have the following Jaccard-calibration. Denote $(\mathcal{U}_\phi^{\mathsf{Jac}})^* \doteq \sup_f \mathcal{U}_\phi^{\mathsf{Jac}}(f)$.

Theorem 10 (Jaccard-calibration)**.** *Assume that a surrogate loss $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is differentiable almost everywhere, convex, and non-increasing, and that $(\mathcal{U}_\phi^{\mathsf{Jac}})^* \geq \tau$ and ϕ is τ-discrepant for some constant $\tau \in (0,1)$. Then, $\mathcal{U}_\phi^{\mathsf{Jac}}$ is Jaccard-calibrated.*

Theorem 10 also relies on the τ-discrepancy as in Theorem 9. Thus, the loss shown in Figure 3 can also be used in the Jaccard case with a certain range of τ. In the same manner as the F_β-measure, a hyperparameter τ ranges over $(0,1)$, which we may either determine by cross-validation or fix to a certain value.

Remark: The τ-discrepancy is a technical assumption making stationary points of $\mathcal{U}_\phi$ lie in the Bayes optimal set of $\mathcal{U}$. This is a mere sufficient condition for $\mathcal{U}$-calibration, while the classification-calibration (Bartlett et al., 2006) is the necessary and sufficient condition for the accuracy.[3] It is left as an open problem to seek for necessary conditions.

5 Consistency Analysis: Bridging Finite Sample and Asymptotics

In this section, we analyze statistical properties of the estimator $\widehat{\mathcal{V}}_\phi$ in Eq. (6). To make our analysis simple, the linear-in-input model $f_\theta(x) = \theta^\top x$ is considered throughout this section, where $\theta \in \Theta$ is a classifier parameter and $\Theta \subset \mathbb{R}^d$ is a compact parameter space. The maximization procedure introduced above can be naturally seen as *Z-estimation* (van der Vaart, 2000), which is an estimation procedure to solve an estimation equation. In our case, the maximization of $\mathcal{U}_\phi$ is reduced to a Z-estimation problem to solve the system $\widehat{\mathcal{V}}_\phi(f) = 0$. The first lemma shows that the derivative estimator $\widehat{\mathcal{V}}_\phi$ admits the uniform convergence. Its proof is deferred to App. C.

Lemma 11 (Uniform convergence)**.** *For simplicity, assume that $m = n/2$. For $k = 0, 1$, let $c_k \doteq \sup_{\xi\in\mathbb{R}, y\in\mathcal{Y}} |W_{k,\phi}(\xi, y)| < +\infty$. Assume that $W_k(\cdot, y)$ for $y \in \mathcal{Y}$ are ρ_k-Lipschitz continuous for some $0 < \rho_k < \infty$, and that $\|x\| < c_\mathcal{X}$ $(\forall x \in \mathcal{X})$ and $\|\theta\| < c_\Theta$*

[2] Note that $(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*$ is non-negative and therefore such τ always exists. The non-negativity is discussed in App. A.5.

[3] We give the surrogate calibration conditions for the accuracy in App. A.3.

($\forall \theta \in \Theta$) for some $0 < c_{\mathcal{X}}, c_{\Theta} < \infty$. Then,

$$\sup_{\theta \in \Theta} \left\| \widehat{\mathcal{V}}_\phi(f_\theta) - \mathcal{V}_\phi(f_\theta) \right\| = \mathcal{O}_p(n^{-\frac{1}{2}}), \tag{7}$$

where $\mathcal{O}_p$ denotes the order in probability.

The Lipschitz continuity and smoothness assumptions in Lemma 11 can be satisfied if the surrogate loss ϕ satisfies a certain Lipschitzness and smoothness. Note that Lemma 11 still holds for τ-discrepant surrogates since we allow surrogates to have different smoothness parameters for both positive and negative domains. Lemma 11 is the basis for showing the consistency. Let $\theta^* \doteq \arg\max_{\theta \in \Theta} \mathcal{U}_\phi(f_\theta)$ and $\widehat{\theta}_n = \arg\max_{\theta \in \Theta} \widehat{\mathcal{U}}_\phi(f_\theta)$. Under the identifiability described below, f_{θ^*} and $f_{\widehat{\theta}_n}$ are roots of $\mathcal{V}_\phi$ and $\widehat{\mathcal{V}}_\phi$, respectively. Then, we can show the consistency of $\widehat{\theta}_n$.

Theorem 12 (Consistency)**.** *Assume that θ^* is* iden*tifiable, that is,* $\inf\{\|\mathcal{V}_\phi(f_\theta)\| \mid \|\theta - \theta^*\| \geq \epsilon\} > \|\mathcal{V}_\phi(f_{\theta^*})\| = 0$ *for all $\epsilon > 0$, and that Eq. (7) holds for $\widehat{\mathcal{V}}_\phi$. Then, $\widehat{\theta}_n \xrightarrow{p} \theta^*$.*

Theorem 12 is an immediate result of van der Vaart (2000, Theorem 5.9), given the uniform convergence (Lemma 11) and the identifiability assumption. Note that the identifiability assumes that $\mathcal{V}_\phi$ has a unique zero f_{θ^*}, which is also usual in the M-estimation: The global optimizer is identifiable. Since Algorithm 1 is a combination of concave and quasi-concave programs, the identifiability would be reasonable to assume.

6 Related Work

In this section, we summarize the existing lines of research on the optimization of generalized performance metrics, which elucidates advantages of our approach.

(i) Surrogate optimization: One of the earliest attempts to optimize non-decomposable performance metrics dates back to Joachims (2005), formulating the structured SVM as a surrogate objective. However, Dembczyński et al. (2013) showed that this surrogate is inconsistent, which means that the surrogate maximization does not necessarily imply the maximization of the true metric. Kar et al. (2014) showed the sublinear regret for the structural surrogate by Joachims (2005) in online setting. Later, Yu and Blaschko (2015), Eban et al. (2017), and Berman et al. (2018) have tried different surrogates, but their calibration has not been studied yet.

(ii) Plug-in rule: Instead of the surrogate optimization, Dembczyński et al. (2013) mentioned that a plug-in rule is consistent, where η and a threshold parameter are estimated independently. We can estimate η by minimizing strictly proper losses (Reid and Williamson, 2009). The plug-in rule has been investigated in many settings (Nan et al., 2012; Dembczyński et al., 2013; Koyejo et al., 2014; Narasimhan et al., 2014; Busa-Fekete et al., 2015; Yan et al., 2018). However, one of the weaknesses of the plug-in rule is that it requires an accurate estimate of η, which is less sample-efficient than the usual classification with convex surrogates (Bousquet et al., 2004; Tsybakov, 2008). Moreover, estimation of the threshold parameter heavily relies on an estimate of η.

Table 2: Comparison of related work.

Method	Consistency	Avoids to estimate η	Efficient optimization
ours	✓	✓	✓
(i)	✗	✓	✓
(ii)	✓	✗	✓
(iii)	✓	✓	✗

(iii) Cost-sensitive risk minimization: On the other hand, Parambath et al. (2014) is a pioneering work to focus on the *pseudo-linearity* of the metrics, which reduces their maximization to an alternative optimization with respect to a classifier and the sublevel. This can be formulated as an iterative cost-sensitive risk minimization (Koyejo et al., 2014; Narasimhan et al., 2015, 2016; Sanyal et al., 2018). Though these methods are blessed with the consistency, they need to train classifiers many times, which may lead to high computational costs, especially for complex hypothesis sets.

Remark: Our proposed methods can be considered to belong to the family (i), while one of the crucial differences is the fact that we have calibration guarantee. We do not need to estimate the class-posterior probability as in (ii), or train classifiers many times as in (iii). This comparison is summarized in Tab. 2.

7 Experiments

In this section, we investigate empirical performances of the surrogate optimizations (Algorithm 1 with NGA and normalized BFGS). Details of datasets, baselines, and full experimental results are shown in App. D.

Implementation Details of Proposed Methods: The linear-in-input model $f_\theta(x) = \theta^\top x$ was used for the hypothesis set $\mathcal{F}$. As the initializer of θ, the ERM minimizer trained by SVM was used. For both NGA and BFGS, gradient updates were iterated 300 times. NGA and normalized BFGS are referred to as U-GD and U-BFGS below, respectively. The surrogate loss shown in Fig. 3 was used: $\phi(m) = \log_2(1+e^{-m})$ when $m \leq 0$ and $\phi(m) = \log_2(1+e^{-\tau m})$ when $m > 0$, where τ was set to 0.33 in the F$_1$-measure case and 0.75 in

Table 3: Benchmark results: 50 trials are conducted for each pair of a method and dataset. Standard errors (multiplied by 10^4) are shown in parentheses. Bold-faces indicate outperforming methods, chosen by one-sided t-test with the significant level 5%. We emphasize that the same number of gradient updates are executed for both U-GD and U-BFGS.

(F$_1$-measure)	Proposed		Baselines		
Dataset	U-GD	U-BFGS	ERM	W-ERM	Plug-in
adult	0.617 (101)	0.660 (11)	0.639 (51)	0.676 (18)	**0.681 (9)**
breast-cancer	**0.963 (31)**	**0.960 (32)**	0.950 (37)	0.948 (44)	0.953 (40)
diabetes	**0.834 (32)**	**0.828 (31)**	0.817 (50)	0.821 (40)	0.820 (42)
sonar	**0.735 (95)**	**0.740 (91)**	0.706 (121)	0.655 (189)	**0.721 (113)**

(Jaccard index)	Proposed		Baselines		
Dataset	U-GD	U-BFGS	ERM	W-ERM	Plug-in
adult	0.499 (44)	0.498 (11)	0.471 (51)	0.510 (20)	**0.516 (10)**
breast-cancer	**0.921 (54)**	**0.918 (55)**	0.905 (66)	0.903 (78)	**0.913 (69)**
diabetes	**0.714 (44)**	0.702 (50)	0.692 (70)	0.698 (56)	0.695 (60)
sonar	**0.600 (125)**	**0.600 (111)**	0.552 (147)	0.495 (202)	**0.572 (134)**

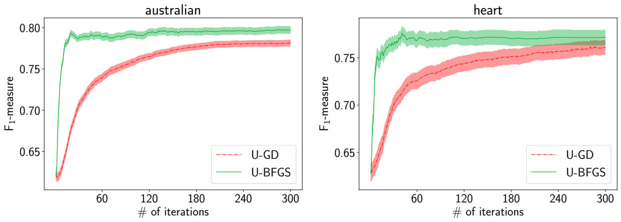

Figure 4: Convergence comparison of the F$_1$-measure (left two figures) and Jaccard index (right two figures). Standard errors of 50 trials are shown as shaded areas.

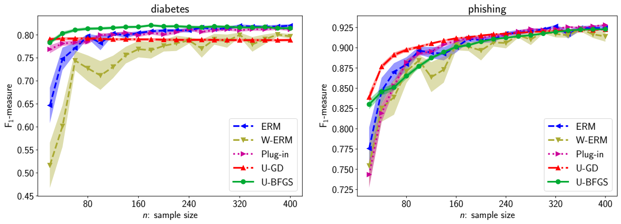

Figure 5: The relationship of the test F$_1$-measure (left two figures) / Jaccard index (right two figures) and sample size (horizontal axes). Standard errors of 50 trials are shown as shaded areas.

the Jaccard index case.[4] The training set was divided into 4 to 1 and the latter set was used for validation. We used a common learning rate in Algorithm 1, which was chosen from $\{10^1, 10^{-1}, 10^{-3}, 10^{-5}\}$ by cross validation.

Convergence Comparison: We compare convergence behaviors of U-GD and U-BFGS. In this experiment, we ran them 300 iterations from randomly initialized parameters drawn from $\mathcal{N}(0_d, I_d)$. The results are summarized in Fig. 4. As we expected, U-BFGS converges much faster than U-GD in most of the cases, up to 30 iterations. Note that U-BFGS and U-GD are in the trade-off relationship in that the former converges within fewer steps while the latter can update the solution faster in each step.

Performance Comparison in Benchmark: We compared the proposed methods with baselines. The results of the F$_1$-measure and Jaccard index are summarized in Tab. 3, respectively, from which we can see the better or at least comparable performances of the proposed methods.

Sample Complexity: We empirically study the relationship between the performance and the sample size. We randomly subsample each original dataset to reduce the sample sizes to $\{20, 40, \ldots, 400\}$, and train all methods on the reduced samples. The experimental results are shown in Fig. 5. Overall, U-GD and U-BFGS outperform, which is especially significant when the sample sizes are quite small. It is worth noting that U-GD works even better than U-BFGS in some cases, though U-GD does not behave significantly better in Tab. 3. This can happen because the Hessian approximation in BFGS might not work well when the sample sizes are extremely small.

8 Conclusion

In this work, we gave a new insight into the calibrated surrogate for the linear-fractional metrics. Sufficient conditions for the surrogate calibration were stated, which is the first calibration result for the linear-fractional metrics to the best of our knowledge. The surrogate maximization can be performed by the combination of concave and quasi-concave programs, and its performance is validated via simulations.

[4] The discrepancy parameter τ should be chosen within $(0, \frac{1}{3})$ and $(0, 1)$ for the F$_1$-measure and Jaccard index, respectively. Here, we fix them to the slightly small values than the upper limits of their ranges. In App. D.6, we study the relationship between performance sensitivity on τ.

Acknowledgement

We would like to thank Nontawat Charoenphakdee, Junya Honda, and Akiko Takeda for fruitful discussions. HB was supported by JST, ACT-I, Japan, Grant Number JPMJPR18UI. MS was supported by JST CREST Grant Number JPMJCR18A2.

References

Ahmed, F., Tarlow, D., and Batra, D. (2015). Optimizing expected Intersection-over-Union with candidate-constrained CRFs. In *CVPR*, pages 1850–1858.

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

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Ben-David, S., Eiron, N., and Long, P. M. (2003). On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514.

Berman, M., Triki, A. R., and Blaschko, M. B. (2018). The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*.

Bousquet, O., Boucheron, S., and Lugosi, G. (2004). Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pages 169–207. Springer.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *ICPR*, pages 3121–3124.

Busa-Fekete, R., Szörényi, B., Dembczyński, K., and Hüllermeier, E. (2015). Online F-measure optimization. In *NeurIPS*, pages 595–603.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Charoenphakdee, N., Lee, J., and Sugiyama, M. (2019). On symmetric losses for learning from corrupted labels. In *ICML*.

Csurka, G., Larlus, D., Perronnin, F., and Meylan, F. (2013). What is a good evaluation measure for semantic segmentation? In *BMVC*, pages 1–11.

Dembczyński, K., Jachnik, A., Kotłowski, W., Waegeman, W., and Hüllermeier, E. (2013). Optimizing the F-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. In *ICML*, pages 1130–1138.

Dembczyński, K., Kotłowski, W., Koyejo, O., and Natarajan, N. (2017). Consistency analysis for binary classification revisited. In *ICML*, pages 961–969.

Eban, E., Schain, M., Mackey, A., Gordon, A., Saurous, R. A., and Elidan, G. (2017). Scalable learning of non-decomposable objectives. In *AISTATS*, pages 832–840.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338.

Feldman, V., Guruswami, V., Raghavendra, P., and Wu, Y. (2012). Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, 41(6):1558–1590.

Fletcher, R. (2013). *Practical Methods of Optimization*. John Wiley & Sons.

Gao, W. and Zhou, Z.-H. (2015). On the consistency of AUC pairwise optimization. In *IJCAI*, pages 939–945.

Gower, J. C. and Legendre, P. (1986). Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3(1):5–48.

Hazan, E., Levy, K., and Shalev-Shwartz, S. (2015). Beyond convexity: Stochastic quasi-convex optimization. In *NeurIPS*, pages 1594–1602.

Jaccard, P. (1901). Étude de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Joachims, T. (2005). A support vector method for multivariate performance measures. In *ICML*, pages 377–384.

Kar, P., Narasimhan, H., and Jain, P. (2014). Online and stochastic gradient methods for non-decomposable loss functions. In *NeurIPS*, pages 694–702.

Koyejo, O. O., Natarajan, N., Ravikumar, P. K., and Dhillon, I. S. (2014). Consistent binary classification with generalized performance metrics. In *NeurIPS*, pages 2744–2752.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes.* Springer.

Lichman, M. (2013). UCI machine learning repository.

Manning, R. P. C. and Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

McDiarmid, C. (1989). On the method of bounded differences. *Surveys in Combinatorics*, 141(1):148–188.

Menon, A., Narasimhan, H., Agarwal, S., and Chawla, S. (2013). On the statistical consistency of algorithms for binary classification under class imbalance. In *ICML*, pages 603–611.

Menon, A., van Rooyen, B., Ong, C. S., and Williamson, R. (2015). Learning from corrupted binary labels via class-probability estimation. In *ICML*, pages 125–134.

Mohri, M., Rostamizadeh, A., and Talkwalkar, A. (2012). *Foundation of Machine learning.* MIT Press.

Nan, Y., Chai, K. M., Lee, W. S., and Chieu, H. L. (2012). Optimizing F-measure: A tale of two approaches. In *ICML*, pages 289–296.

Narasimhan, H., Kar, P., and Jain, P. (2015). Optimizing non-decomposable performance measures: a tale of two classes. In *ICML*, pages 199–208.

Narasimhan, H., Pan, W., Kar, P., Protopapas, P., and Ramaswamy, H. G. (2016). Optimizing the multiclass F-measure via biconcave programming. In *ICDM*, pages 1101–1106.

Narasimhan, H., Vaish, R., and Agarwal, S. (2014). On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In *NeurIPS*, pages 1493–1501.

Natarajan, N., Koyejo, O., Ravikumar, P., and Dhillon, I. (2016). Optimal classification with multivariate losses. In *NeurIPS*, pages 1530–1538.

Osokin, A., Bach, F., and Lacoste-Julien, S. (2017). On structured prediction theory with calibrated convex surrogate losses. In *NeurIPS*, pages 302–313.

Parambath, S. P., Usunier, N., and Grandvalet, Y. (2014). Optimizing F-measures by cost-sensitive classification. In *NeurIPS*, pages 2123–2131.

Reid, M. D. and Williamson, R. C. (2009). Surrogate regret bounds for proper losses. In *ICML*, pages 897–904.

Sanyal, A., Kumar, P., Kar, P., Chawla, S., and Sebastiani, F. (2018). Optimizing non-decomposable measures with deep networks. *Machine Learning*, 107(8-10):1597–1620.

Scott, C. (2012). Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6:958–992.

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

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

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation.* Springer.

van de Geer, S. (2000). *Empirical Processes in M-estimation.* Cambridge University Press.

van der Vaart, A. W. (2000). *Asymptotic Statistics.* Cambridge University Press.

van Rijsbergen, C. J. (1974). *Foundation of Evaluation.* Number 4.

Vapnik, V. (1998). *Statistical Learning Theory.* Wiley, New York.

Yan, B., Koyejo, O., Zhong, K., and Ravikumar, P. (2018). Binary classification with Karmic, threshold-quasi-concave metrics. In *ICML*, pages 5531–5540.

Yu, J. and Blaschko, M. (2015). Learning submodular losses with the Lovász hinge. In *ICML*, pages 1623–1631.

Zhang, T. (2004a). Statistical analysis of some multicategory large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251.

Zhang, T. (2004b). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85.