On Loss Functions and Regret Bounds for Multi-Category Classification

Zhiqiang Tan and Xinwei Zhang

***Abstract*— We develop new approaches in multi-class settings for constructing loss functions and establishing corresponding regret bounds with respect to the zero-one or cost-weighted classification loss. We provide new general representations of losses by deriving inverse mappings from a concave generalized entropy to a loss through a convex dissimilarity function related to the multi-distribution f-divergence. This approach is then applied to study both hinge-like losses and proper scoring rules. In the first case, we derive new hinge-like convex losses, which are tighter extensions outside the probability simplex than related hinge-like losses and geometrically simpler with fewer non-differentiable edges. We also establish a classification regret bound in general for all losses with the same generalized entropy as the zero-one loss, thereby substantially extending and improving existing results. In the second case, we identify new sets of multi-class proper scoring rules through different types of dissimilarity functions and reveal interesting relationships between various composite losses currently in use. We also establish new classification regret bounds in general for multi-class proper scoring rules and, as applications, provide simple meaningful regret bounds for two specific sets of proper scoring rules. These results generalize, for the first time, previous two-class regret bounds to multi-class settings.**

***Index Terms*— Boosting, Bregman divergence, composite loss, exponential loss, f-divergence, generalized entropy, hinge loss, proper scoring rule, surrogate regret bounds.**

I. Introduction

MULTI-CATEGORY classification has been extensively studied in machine learning and statistics. For concreteness, let $\{(X_i, Y_i) : i = 1, \ldots, n\}$ be training data generated from a certain probability distribution on (X, Y), where X is a covariate or feature vector and Y is a class label, with possible values from 1 to $m\,(\geq 2)$. Various learning methods are developed in the form of minimizing an empirical risk function,

$$\hat{R}_L(\alpha) = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, \alpha(X_i)), \tag{1}$$

where $L(y, \alpha(x))$ is a loss function, and $\alpha(x)$ is a vector-valued function of covariates, taken from a potentially rich family of functions, for example, reproducing kernel Hilbert spaces or neural networks. For convenience, $\alpha(x)$ is called an action function, following the terminology of decision theory [1], [2]. The performance of $\alpha(x)$ is typically evaluated by the zero-one risk on test data,

$$E\big\{L^{\text{zo}}(Y_0, \tilde{\alpha}(X_0))\big\}, \tag{2}$$

where (X_0, Y_0) is a new observation, independent of training data, and $\tilde{\alpha}(x)$ is m-dimensional, either $\alpha(x)$ itself or converted from $\alpha(x)$, depending on whether $\alpha(x)$ is m-dimensional or not, and $L^{\text{zo}}(y, \tilde{\alpha}(x))$ is the zero-one loss, defined as 0 if the yth component of $\tilde{\alpha}(x)$ is a maximum and 1 otherwise. Due to discontinuity, using L^{zo} directly as L in (1) is computationally intractable. Hence the loss L used in (1) is also referred to as a surrogate loss for L^{zo}.

It is helpful to distinguish two types of loss functions L commonly used for training in (1). One type of losses, called scoring rules, involves an action defined as a probability vector $q : \mathcal{X} \to \Delta_m$, where $\mathcal{X}$ is the covariate space and Δ_m is the probability simplex with m categories [3], [4]. The elements of $q(x)$ can be interpreted as class probabilities. Typically, the probability vector $q(x)$ is parameterized in terms of a real vector $h(x)$ as $q^h(x)$, via an invertible link such as the multinomial logistic (or softmax) link. The resulting loss $L(y, q^h(x))$ is then called a composite loss, with $h(x)$ as an action function [5]. It is often desired to combine a *proper* scoring rule, which ensures infinite-sample consistency of probability estimation (see Section II-B), with a link function q^h such that $L(y, q^h(x))$ is convex in h. In multi-class settings, composite losses satisfying these properties include the standard multinomial likelihood loss and two variants of exponential losses related to boosting [6], [7], all combined with the multinomial logistic link.

Another type of losses involves an action function, $\alpha : \mathcal{X} \to \mathbb{R}^m$, allowed to take unrestricted values in $\mathbb{R}^m$. The elements of $\alpha(x)$, loosely called margins, can be interpreted as relative measures of association of x with the m classes. Although a composite loss $L(y, q^h(x))$ based on a scoring rule can be considered with $h(x)$ as a margin vector, it is mainly of interest to include in this type hinge-like losses, where the margins are designed not to be directly mapped to probability vectors. The hinge loss is originally related to support vector machines in two-class settings. This loss, $L^{\text{hin}}(y, \tau(x))$, is known to be convex in its action τ, and achieve classification calibration (or infinite-sample classification consistency), which means that a minimizer of the hinge loss in the population version leads to

Manuscript received 15 May 2021; revised 25 January 2022; accepted 3 April 2022. Date of publication 18 April 2022; date of current version 13 July 2022. *(Corresponding author: Zhiqiang Tan.)*

The authors are with the Department of Statistics, Rutgers University, Piscataway, NJ 08854 USA (e-mail: ztan@stat.rutgers.edu; xinwei.zhang@rutgers.edu).

Communicated by V. Y. F. Tan, Associate Editor for Machine Learning.

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/3167635.

Digital Object Identifier 10.1109/TIT.2022.3167635

a Bayes rule minimizing the zero-one risk (2) [8]–[10]. There are various extensions of the hinge loss to multi-class settings. The hinge-like losses in [11] and [12] are shown to achieve classification calibration, whereas those in [13] and [14] fail to achieve such a property [15], [16].

Classification calibration is also called Fisher consistency, although it is appropriate to distinguish two types of Fisher consistency, in parallel to the two types of losses above: Fisher probability consistency as satisfied by a proper scoring rule or Fisher classification consistency as achieved by a hinge-like loss. In general, Fisher probability consistency (or properness) implies classification consistency, but not vice versa [5]. On the other hand, there are interesting results indicating that only hinge-like losses are classification consistent with respect to the zero-one loss when both an action function and a data quantizer are estimated [12], [17].

The purpose of this article is two-fold: first constructing new multi-class losses while studying existing ones, and second establishing corresponding classification regret bounds. Such a regret bound compares the regret of the loss under study with that of the zero-one or cost-weighted classification loss and implies that classification calibration is achieved with a quantitative guarantee [10]. Our development in both directions is facilitated by the concept of a generalized entropy, defined as the minimum Bayes risk for a loss [2]. In the following, we give an overview of the main results and related work.

A. Main Results

The main results from our work can be split into three groups.

First, in Section III, we provide new general representations of multi-class loss functions depending on a (concave) generalized entropy through a (convex) dissimilarity function f, which is related to the multi-distribution f-divergence [18], [19]. These results are complementary to previous representations directly based on the generalized entropy [3], [12], [20]. While the generalized entropy is defined on m-dimensional probability vectors, the dissimilarity function is defined on $(m-1)$-dimensional free-varying vectors of probability ratios. To demonstrate advantages, this approach is applied in the subsequent sections to construct new specific hinge-like losses and proper scoring rules.

Second, in Section IV, we investigate hinge-like losses in two directions.

- We derive two new hinge-like losses, related to [11] and [12] respectively (Section IV-A). In each case, our new loss and the existing one admit the same generalized entropy and coincide with each other for actions restricted to the probability simplex Δ_m, but our loss is uniformly lower (hence a tighter extension outside the probability simplex) and geometrically simpler with fewer non-differentiable edges.
- We establish classification regret bounds for our new hinge-like losses (Section IV-B) and more broadly for *all* losses with the same generalized entropy as the zero-one loss (Section IV-C). These results represent a substantial extension and improvement over existing ones [12], [16].

Third, in Section V, we investigate proper scoring rules in two directions.

- We derive two new sets of proper scoring rules: multi-class pairwise losses corresponding a univariately additive dissimilarity function f and multi-class simultaneous losses with non-additive dissimilarity functions (Section V-A). These sets of losses not only reveal interesting relationships between the likelihood and exponential losses mentioned earlier, but also lead to new specific losses including a pairwise likelihood loss distinct from the standard multinomial likelihood loss.
- We establish classification regret bounds for multi-class proper scoring rules in general with respect to the zero-one or cost-weighted classification loss, and then derive simple meaningful regret bounds for two specific sets of proper scoring rules including the multinomial likelihood loss and the pairwise likelihood and exponential losses (Section V-B). These results appear to generalize, for the first time, previous two-class regret bounds to multi-class settings [10], [21].

B. Related Work

There is an extensive literature on multi-category classification including and beyond the special case of binary classification. We discuss directly related work to ours, in addition to those mentioned above. An inverse mapping from a generalized entropy to a proper scoring rule can be seen in the canonical representation of proper scoring rules [3], [20]. This and related representations are extensively used in the design and study of composite binary losses [4], [21] and composite multi-class losses [5].

Recently, an inverse mapping is constructed by [12] from a generalized entropy to a convex loss with actions in $\mathbb{R}^m$, hence different from the canonical representation of proper scoring rules. Our construction of losses is in a similar spirit as [12], but operates explicitly through a dissimilarity function f. Our approach is applicable to handling both hinge-like losses and proper scoring rules and leads to interesting new findings. For example, our inverse mapping in terms of f are applied to discover new hinge-like losses, by first identifying a hinge-like loss on the probability simplex and then constructing a convex extension. The hinge-like losses in [11] and [12] are also such convex extensions. This point of view enriches our understanding of multi-class hinge-like losses. For another example, using an additive function f provides a convenient, general extension of two-class proper scoring rules to multi-class settings. By comparison, using an additive generalized entropy does not seem to achieve a similar effect.

Our regret bounds for the new hinge-like losses are similar to those in [12]. However, we also establish in general that all losses with the same generalized entropy as the zero-one loss achieve a regret bound which ensures classification calibration. Compared with [16], our result provides a more concrete sufficient condition for achieving classification calibration, in addition to a quantitative guarantee.

Our new regret bounds for proper scoring rules generalize two-class results in [21], Section 7.1, to multi-class settings,

by carefully exploiting the Bregman representation for the regret of a proper scoring rule together with a novel bound on the regret of the zero-one or cost-weighted classification loss (Lemmas 5–6). Such a generalization seems to be previously unnoticed (cf. [5]).

As noted earlier, classification regret bounds provide a quantitative guarantee on classification calibration, a qualitative property studied in [15], [22], and [16] among others. Although two-class regret bounds can be obtained for all margin-based losses including the hinge loss and proper scoring rules [9], [10], [23], such results seem to rely on simplification due to two classes.

Notation: Denote $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$, $\mathbb{R}_+ = \{b \in \mathbb{R} : b \geq 0\}$, and $\overline{\mathbb{R}}_+ = \{b \in \overline{\mathbb{R}} : b \geq 0\}$. For $m \geq 2$, denote $[m]$ as the set $\{1, \ldots, m\}$, 1_m as the $m \times 1$ vector of all ones, I_m as the $m \times m$ identity matrix, and Δ_m as the probability simplex $\{q \in \mathbb{R}_+^m : 1_m^{\mathrm{T}} q = 1\}$. For $j \in [m]$, a basis vector $e_j \in \Delta_m$ is defined such that its jth element is 1 and the remaining elements are 0. The indicator function $1\{\cdot\}$ is defined as 1 if the argument is true or 0 otherwise.

II. Background

We provide a selective review of basic concepts and results which are instrumental to our subsequent development. See [2], [4], [20], and [12] among others for more information.

A. Losses, Risks and Entropies

Consider the population version of the multi-category classification problem. Let $X \in \mathcal{X}$ be a vector of observed covariates or features, but $Y \in [m]$ an unobserved class label, where (X, Y) are generated from some joint probability distribution which can be assumed to be known unless otherwise noted. It is of interest to predict the value of Y based on X (i.e., assign X to one of the m classes). The prediction can be performed using an action function $\alpha : \mathcal{X} \to \mathcal{A}$, and evaluated through a loss function $L(y, \alpha(x))$ when the true label of x is y. Typically, an action in the space $\mathcal{A}$ is a vector whose components, as probabilities or margins, measure the strengths of association with the m classes. The risk (or expected loss) of the action function $\alpha(x)$ is

$$R_L(\alpha) = E(L(Y, \alpha(X)) = E\left\{\sum_{j=1}^{m} \pi_j(X) L(j, \alpha(X))\right\}, \tag{3}$$

where $\pi_j(x) = P(Y = j | X = x)$, the conditional probability of class j given covariates x, and the second expectation is taken over the marginal distribution of X only.

From another perspective, the preceding problem can also be formulated as a Bayesian experiment with m probability distributions $(P_1, \ldots, P_m)$ on $\mathcal{X}$, corresponding to the within-class distributions of covariates [1]. Denote by $p_j(x)$ the density function of P_j with respect to a baseline measure μ. Given a label $Y = j$ (regarded as an m-valued parameter), the random variable X is drawn from the distribution P_j. Let $\pi^0 = (\pi_1^0, \ldots, \pi_m^0)^{\mathrm{T}} \in \Delta_m$ be the prior probabilities of Y, corresponding to the marginal class probabilities. Then the posterior probabilities of Y given $X = x$ are $\pi_j(x) = \pi_j^0 p_j(x) / \{\sum_{k=1}^m \pi_k^0 p_k(x)\}$, the same as the conditional class probabilities given covariates mentioned above. In this context, $R_L(\alpha)$ is also called the Bayes risk of $\alpha(x)$. By standard Bayes theory [19, Eq. (5)], the minimum Bayes risk, or even shortened as the Bayes risk, can be obtained as

$$\inf_{\alpha : \mathcal{X} \to \mathcal{A}} R_L(\alpha) = E\{H_L(\pi(X))\}, \tag{4}$$

where $\alpha : \mathcal{X} \to \mathcal{A}$ can be any measurable function, $\pi(x) = (\pi_1(x), \ldots, \pi_m(x))^{\mathrm{T}}$, and H_L is a function defined on Δ_m such that for $\eta = (\eta_1, \ldots, \eta_m)^{\mathrm{T}} \in \Delta_m$,

$$H_L(\eta) = \inf_{\gamma \in \mathcal{A}} \left\{\sum_{j=1}^{m} \eta_j L(j, \gamma)\right\}, \tag{5}$$

The function H_L, which is concave on Δ_m, is called an uncertainty function [1] or a generalized entropy associated with the loss L [2].

A subtle point is that minimization in (4) is over all measurable functions $\alpha : \mathcal{X} \to \mathcal{A}$, whereas minimization in (5) is over all elements $\gamma \in \mathcal{A}$. The generalized entropy H_L is merely a function on Δ_m, induced by the loss $L(j, \gamma)$ on $[m] \times \mathcal{A}$, where the covariate vector X is conditioned on (or lifted out). Similarly, the risk of an action $\gamma \in \mathcal{A}$ is defined as $R_L(\eta, \gamma) = \sum_{j=1}^m \eta_j L(j, \gamma)$, and the regret (or excess risk) of the action is defined as

$$B_L(\eta, \gamma) = R_L(\eta, \gamma) - H_L(\eta), \tag{6}$$

where $H_L(\eta) = \inf_{\gamma' \in \mathcal{A}} R_L(\eta, \gamma')$ by (5). This simplification where the covariate vector X is lifted out is often useful when studying losses and regrets.

B. Scoring Rules

A scoring rule is a particular type of loss $L(j, q)$, where its action q is a probability vector in Δ_m, interpreted as the predicted class probabilities [2]. Sometimes, the expected loss, $R_L(\eta, q) = \sum_{j=1}^m \eta_j L(j, q)$, is also referred to as a scoring rule, for measuring the discrepancy between underlying and predicted probability vectors, η and q [20].

A scoring rule $L(j, q)$ is said to be proper if $H_L(\eta) = R_L(\eta, \eta)$, i.e.,

$$R_L(\eta, \eta) \leq R_L(\eta, q), \quad \eta, q \in \Delta_m.$$

The rule is strictly proper if the inequality is strict for $q \neq \eta$. Hence for a proper scoring rule, the expected loss $R_L(\eta, q)$ is minimized over $q \in \Delta_m$ when $q = \eta$, the predicted probability vector coincides with the underlying probability vector. This condition is typically required for establishing large-sample consistency of (conditional) probability estimators (e.g., [4], [9]).

As shown in [3] and [20], a proper scoring rule $L(j, q)$ in general admits the following representation:

$$R_L(\eta, q) = H_L(q) \ - (q - \eta)^{\mathrm{T}} \partial H_L(q), \quad \eta, q \in \Delta_m, \tag{7}$$

where $-\partial H_L$ is a sub-gradient of the convex function $-H_L$ on $\mathbb{R}^m$. Note that the generalized entropy H_L is evaluated at q, not η, in (7). Then the regret in (6) becomes

$$B_L(\eta, q) = H_L(q) \ - H_L(\eta) - (q-\eta)^{\mathrm{T}} \partial H_L(q), \quad (8)$$

which is the Bregman divergence from q to η associated with the convex function $-H_L$.

An important example of proper scoring rules is the logarithmic scoring rule [24], $L(j,q) = -\log q_j$. The corresponding expected loss is $R_L(\eta, q) = -\sum_{j=1}^m \eta_j \log q_j$, which is, up to scaling, the negative expected log-likelihood of the predicted probability vector q with the underlying probability vector η for multinomial data. The generalized entropy is $H_L(\eta) = -\sum_{j=1}^m \eta_j \log \eta_j$, the negative Shannon entropy. The regret is $B_L(\eta, q) = \sum_{j=1}^m \eta_j \log(\eta_j/q_j)$, the Kullback–Liebler divergence.

C. Classification Losses

Consider the zero-one loss, formally defined as

$$L^{\mathrm{zo}}(j, \gamma) = 1\{j \neq \mathrm{argmax}_{k\in[m]} \gamma_k\}, \quad j \in [m], \gamma \in \mathbb{R}^m,$$

where, if not unique, $\mathrm{argmax}_{k\in[m]}\gamma_k$ can be fixed as the index of any maximum component of γ. As mentioned below (2), L^{zo} is typically used to evaluate performance, but not as the loss L for training, and the action γ can be transformed from the action of L. Nevertheless, the generalized entropy defined by (5) with $L = L^{\mathrm{zo}}$ is

$$H^{\mathrm{zo}}(\eta) = 1 - \max_{k\in[m]} \eta_k, \quad \eta \in \Delta_m. \quad (9)$$

This function is concave and continuous, but not everywhere differentiable.

In practice, there can be different costs of misclassification, depending on which classes are involved. For example, the cost of classifying a cancerous tumor as benign can be greater than in the other direction. Let $C = (c_{jk})_{j,k\in[m]}$ be a cost matrix, where $c_{jk} \geq 0$ indicates the cost of classifying class j as class k. For each $j \in [m]$, assume that $c_{jj} = 0$ and $c_{jk} > 0$ for some $k \neq j$. Consider the cost-weighted classification loss

$$L^{\mathrm{cw}}(j, \gamma) = c_{jk} \text{ if } k = \mathrm{argmax}_{l\in[m]} \gamma_l, \quad j \in [m], \gamma \in \mathbb{R}^m.$$

As shown in [12], the generalized entropy defined by (5) with $L = L^{\mathrm{cw}}$ is

$$H^{\mathrm{cw}}(\eta) = \min_{k\in[m]} \eta^{\mathrm{T}} C_k, \quad \eta \in \Delta_m, \quad (10)$$

where $C = (C_1, \ldots, C_m)$ is the column representation of C. The standard zero-one loss corresponds to the special choice $C = 1_m 1_m^{\mathrm{T}} - I_m$.

An intermediate case is the class weighted classification loss,

$$L^{\mathrm{cw0}}(j, \gamma) = c_{j0} 1\{j \neq \mathrm{argmax}_{k\in[m]} \gamma_k\}, \quad j \in [m], \gamma \in \mathbb{R}^m,$$

where $c_{j0} > 0$ is the cost associated with misclassification of class j. This loss is more general than the standard zero-one loss L^{zo}, although a special case of the cost-weighted loss L^{cw} with $C = C_0 1_m^{\mathrm{T}} - \mathrm{diag}(C_0)$, where $C_0 = (c_{10}, \ldots, c_{m0})^{\mathrm{T}}$. The generalized entropy associated with L^{cw0} is $H^{\mathrm{cw0}}(\eta) = \eta^{\mathrm{T}} C_0 - \max_{k\in[m]} \eta_k c_{k0}$.

D. Entropies and Divergences

In DeGroot's theory [1], any concave function H on Δ_m can be used as an uncertainty function. The information of X about label ("parameter") Y is defined as the reduction of uncertainty (or entropy) from the prior to the posterior:

$$I_H(X; \pi^0) = H(\pi^0) - E\{H(\pi(X))\},$$

which is nonnegative by the concavity of H. The information $I_H(X; \pi^0)$ is closely related to the f-divergence between the multiple distributions $(P_1, \ldots, P_m)$, which is a generalization of the f-divergence between two distributions [25], [26]. Heuristically, the more dissimilar $(P_1, \ldots, P_m)$ are from each other, the more information about Y is obtained after observing X.

For a convex function f on $\overline{\mathbb{R}}_+^{m-1}$ with $f(1_{m-1}) = 0$, the f-divergence between $(P_1, \ldots, P_{m-1})$ and P_m with densities $(p_1, \ldots, p_{m-1})$ and p_m is

$$D_f(P_{1:(m-1)} \| P_m) = \frac{1}{m} \int f\left(\frac{p_1(x)}{p_m(x)}, \ldots, \frac{p_{m-1}(x)}{p_m(x)}\right) p_m(x) \, \mathrm{d}\mu(x),$$

which is nonnegative by the convexity of f. Compared with the standard definition of multi-way f-divergences [12], [18], our definition above involves a rescaling factor m^{-1}, for notational simplicity in the later discussion; otherwise, for example, rescaling would be needed in Eqs. (14) and (15).

There is a one-to-one correspondence between the statistical information $I_H(X; \pi^0)$ and multi-way f-divergences, as discussed in [19]. For any prior probability $\pi^0 \in \Delta_m$ and probability distributions $(P_1, \ldots, P_m)$, if a convex function f on $\overline{\mathbb{R}}_+^{m-1}$ with $f(1_{m-1}) = 0$ and a concave function H on Δ_m are related such that for $\eta = (\eta_1, \ldots, \eta_m)^{\mathrm{T}} \in \Delta_n$,

$$H(\eta) = -\frac{\eta_m}{m\pi_m^0} f\left(\frac{\pi_m^0}{\pi_1^0}\frac{\eta_1}{\eta_m}, \ldots, \frac{\pi_m^0}{\pi_{m-1}^0}\frac{\eta_{m-1}}{\eta_m}\right), \quad (11)$$

then $I_H(X; \pi^0) = D_f(P_{1:(m-1)} \| P_m)$ or, because $H(\pi^0) = -f(1_{m-1}) = 0$ here,

$$-E\{H(\pi(X))\} = D_f(P_{1:(m-1)} \| P_m), \quad (12)$$

where the expectation is taken over $X \sim \sum_{j=1}^m \pi_j^0 P_j$.

III. General Construction of Losses

In practice, a learning method for classification involves minimization of (1), an empirical version of the risk (3) based on training data, with specific choices of a loss function $L(y, \alpha)$ and a potentially rich family of action functions $\alpha(x)$. As suggested in Section II-A, we study construction of the loss $L(y, \alpha)$ as a function of a label y and a freely-varying action α, with the dependency on covariates (or features) lifted out. As a result, we not only derive new general classes of losses, but also improve understanding of various existing losses as shown in Sections IV–V. Nevertheless, the interplay between losses and function classes remains important, but challenging to study, for further research.

Equation (5) is a mapping from a loss L to a generalized entropy H_L, which is in general many-to-one (i.e., different

losses can lead to the same generalized entropy). [12] constructed an inverse mapping from a generalized entropy to a convex loss. For a closed, concave function H on Δ_m, define a loss with action space $\mathcal{A} = \mathbb{R}^m$ such that for $\gamma = (\gamma_1, \ldots, \gamma_m)^{\mathrm{T}} \in \mathbb{R}^m$,

$$L_H(j, \gamma) = -\gamma_j + (-H)^*(\gamma), \tag{13}$$

where $(-H)^*(\gamma) = \sup_{\eta \in \Delta_m} \{\gamma^{\mathrm{T}} \eta + H(\eta)\}$, the conjugate of $-H$. Then $L_H(j, \gamma)$ is convex in γ and (5) is satisfied with $H_{L_H} = H$, by [12], Proposition 3. Hence a convex loss is obtained for a concave function on Δ_m to be the generalized entropy. Note that the loss L_H is over-parameterized, because $(-H)^*(\gamma - b1_m) = -b + (-H)^*(\gamma)$ and hence $L_H(j, \gamma - b1_m) = L_H(j, \gamma)$ for any constant $b \in \mathbb{R}$.

We derive a new mapping from generalized entropies to convex losses, by working with perspective-like functions related to multi-distribution f-divergences. First, there exists a one-to-one correspondence between concave functions H on Δ_m and convex functions f on $\overline{\mathbb{R}}_+^{m-1}$. For a convex function f on $\overline{\mathbb{R}}_+^{m-1}$, define a function on Δ_m:

$$H_f(\eta) = -\eta_m f\left(\frac{\eta_1}{\eta_m}, \ldots, \frac{\eta_{m-1}}{\eta_m}\right). \tag{14}$$

Conversely, for a concave function H on Δ_m, define a function on $\overline{\mathbb{R}}_+^{m-1}$:

$$f_H(t) = -t_\bullet H\left(\frac{t_1}{t_\bullet}, \ldots, \frac{t_{m-1}}{t_\bullet}, \frac{1}{t_\bullet}\right), \tag{15}$$

where $t = (t_1, \ldots, t_{m-1})^{\mathrm{T}}$ and $t_\bullet = 1 + \sum_{j=1}^{m-1} t_j$. The mappings H_f and f_H are of a similar form to perspective functions associated with f and H respectively, although neither fits the standard definition of perspective functions [27].

Lemma 1 (García-García and Williamson [19]): For a convex function f on $\overline{\mathbb{R}}_+^{m-1}$, the function H_f defined by (14) is concave on Δ_m such that (15) is satisfied with $f_{H_f} = f$. Conversely, for a concave function H on Δ_m, the function f_H defined by (15) is convex on $\overline{\mathbb{R}}_+^{m-1}$ such that (14) is satisfied with $H_{f_H} = H$. Moreover, it is preserved that $H(1_m/m) = -m^{-1} f(1_{m-1})$ under (14) and (15).

Remark 1: Equations (14) and (15) can be obtained as a special case of (11) with $\pi^0 = 1_m/m$, from [19]. As mentioned in Section II-D, (11) is originally determined such that identity (12) holds for linking the expected entropy and multi-way f-divergences, which are, by definition, concerned with the covariates and within-class distributions. Nevertheless, our subsequent development is technically independent of this connection, because covariates are lifted out in our study. In other words, we merely use (14) and (15) as convenient mappings between H and f. The usual restriction $f(1_{m-1}) = 0$ used in f-divergences does not need to be imposed.

Our first main result shows a mapping from a convex function f to a convex loss L such that the concave function H_f is the generalized entropy associated with L.

Proposition 1: For a closed, convex function f on $\overline{\mathbb{R}}_+^{m-1}$, define a loss with action space $\mathcal{A} = \mathrm{dom}(f^*)$ such that for $s = (s_1, \ldots, s_{m-1})^{\mathrm{T}} \in \mathrm{dom}(f^*)$,

$$L_f(j, s) = \begin{cases} -s_j, & j \in [m-1], \\ f^*(s), & j = m, \end{cases} \tag{16}$$

where $f^*(s) = \sup_{t \in \overline{\mathbb{R}}_+^{m-1}} \{s^{\mathrm{T}} t - f(t)\}$ and $\mathrm{dom}(f^*) = \{s \in \mathbb{R}_+^{m-1} : f^*(s) < \infty\}$. Then $L_f(j, s)$ is convex in s. Moreover, the concave function H_f defined by (14) is the generalized entropy associated with L_f, that is, (5) is satisfied with $H_{L_f} = H_f$.

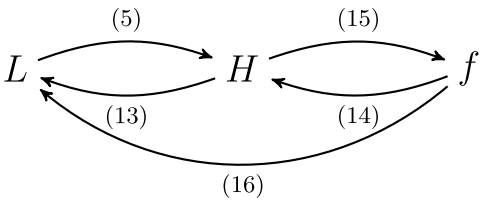


Fig. 1. Relations between loss L, generalized entropy H, and dissimilarity function f.

Proof: For $\eta \in \Delta_m$ and $s \in \mathrm{dom}(f^*)$, the definition of L_f implies that $\sum_{j=1}^m \eta_j L_f(j, s) = -\sum_{j=1}^{m-1} \eta_j s_j + \eta_m f^*(s)$. Hence

$$\begin{aligned} \inf_{s \in \mathcal{A}} \left\{ \sum_{j=1}^m \eta_j L_f(j, s) \right\} &= -\sup_{s \in \mathcal{A}} \left\{ \sum_{j=1}^{m-1} \eta_j s_j - \eta_m f^*(s) \right\} \\ &= -\eta_m \sup_{s \in \mathrm{dom}(f^*)} \left\{ \sum_{j=1}^{m-1} \frac{\eta_j}{\eta_m} s_j - f^*(s) \right\} \\ &= -\eta_m f\left(\frac{\eta_1}{\eta_m}, \ldots, \frac{\eta_{m-1}}{\eta_m}\right) = H_f(\eta), \end{aligned}$$

where the second last equality holds by Fenchel's conjugacy relationship. ■

Compared with (13), Eq. (16) together with (15) presents an alternative approach for determining a convex loss L from a generalized entropy H through a dissimilarity function f. See Figure 1 which illustrates various relationships discussed. For ease of interpretation, a convex function f on $\overline{\mathbb{R}}_+^{m-1}$ can be called a dissimilarity function, similarly as a concave function H on Δ_m can be a generalized entropy.

In spite of the one-to-one correspondence between entropy and dissimilarity functions H and f by (14) and (15), we stress that the new loss (16) is in general distinct from (13). An immediate difference, which is further discussed in Section IV, is that the action space for loss (16), $\mathrm{dom}(f^*)$, can be a strict subset of $\mathbb{R}^{m-1}$, whereas the action space for loss (13) is either $\mathbb{R}^m$ with over-parametrization as noted above or $\mathbb{R}^{m-1}$ with, for example, $\gamma_m = 0$ fixed to remove over-parametrization. Moreover, loss (16) can also be used to derive a new class of closed-form losses based on arbitrary convex functions f as shown in the following result and, as discussed in Section IV, to find novel multi-class, hinge-like losses related to the zero-one or cost-weighted classification loss.

Proposition 2: For a closed, convex function f on $\overline{\mathbb{R}}_+^{m-1}$, define a loss with action space $\mathcal{A} = \overline{\mathbb{R}}_+^{m-1}$ such that for

$u = (u_1, \ldots, u_{m-1})^{\mathrm{T}} \in \overline{\mathbb{R}}_+^{m-1}$,

$$L_{f2}(j,u) = \begin{cases} -\partial_j f(u), & j \in [m-1], \\ u^{\mathrm{T}} \partial f(u) - f(u), & j = m, \end{cases} \tag{17}$$

where $\partial f = (\partial_1 f, \ldots, \partial_{m-1} f)^{\mathrm{T}}$ is a sub-gradient of f, arbitrarily fixed (if needed). Then the concave function H_f defined by (14) is the generalized entropy associated with L_{f2}, that is, (5) is satisfied with $H_{L_{f2}} = H_f$.

Proof (Outline): A basic idea is to use the parametrization $s = \partial f(u)$ and Fenchel's conjugacy property $f^*(s) = u^{\mathrm{T}} s - f(u)$, and then obtain the loss L_{f2} from L_f in Proposition 1. This argument gives a one-sided inequality for the desired equality (5). A complete proof is provided in the Supplement. ∎

Compared with loss (16), the preceding loss (17) is of a closed form without involving the conjugate f^*, which can be nontrivial to calculate. On the other hand, loss (17) may not be convex in its action u. Nevertheless, it is often possible to choose a link function, for example, $u^h = (u_1^h, \ldots, u_{m-1}^h)^{\mathrm{T}}$ with $u_j^h = \exp(h_j)$ such that $L_{f2}(j, u^h)$ becomes convex in $(h_1, \ldots, h_{m-1})$. This link can be easily identified as the multinomial logistic link after reparameterizing $(u_1, \ldots, u_{m-1})$ as probability ratios below.

The following result shows that a reparametrization of loss (17) with actions defined as probability vectors in Δ_m automatically yields a proper scoring rule. See Section II-B for the related background on scoring rules. Together with the relationship between f and H by (14) and (15), our construction gives a mapping from a dissimilarity function f or equivalently a generalized entropy H to a proper scoring rule.

Proposition 3: For a closed, convex function f on $\overline{\mathbb{R}}_+^{m-1}$, define a loss with action space $\mathcal{A} = \Delta_m$ such that for $q = (q_1, \ldots, q_m)^{\mathrm{T}} \in \Delta_m$,

$$L_{f3}(j,q) = \begin{cases} -\partial_j f(u^q), & j \in [m-1], \\ u^{q\mathrm{T}} \partial f(u^q) - f(u^q), & j = m, \end{cases} \tag{18}$$

where $u^q = (q_1/q_m, \ldots, q_{m-1}/q_m)^{\mathrm{T}}$. Then L_{f3} is a proper scoring rule, with H_f defined by (14) as the generalized entropy, satisfying

$$\inf_{q \in \Delta_m} \left\{ \sum_{j=1}^m \eta_j L_{f3}(j,q) \right\} = H_f(\eta) = \sum_{j=1}^m \eta_j L_{f3}(j,\eta).$$

Proof: The generalized entropy from L_{f3} is H_f, due to Proposition 2 and the one-to-one mapping $u^q = (q_1/q_m, \ldots, q_{m-1}/q_m)^{\mathrm{T}}$. Then direct calculation shows that

$$\begin{aligned} &\sum_{j=1}^m \eta_j L_{f3}(j,\eta) \\ &= -\sum_{j=1}^{m-1} \eta_j \partial_j f(u^\eta) + \eta_m \left\{ \sum_{j=1}^{m-1} \frac{\eta_j}{\eta_m} \partial_j f(u^\eta) - f(u^\eta) \right\} \\ &= -\eta_m f(u^\eta) = H_f(\eta) = \inf_{q \in \Delta_m} \left\{ \sum_{j=1}^m \eta_j L_{f3}(j,q) \right\}. \end{aligned}$$

Hence L_{f3} is a proper scoring rule. ∎

For completeness, the expected loss associated with L_{f_3} can be shown to satisfy the canonical representation (7) with H_f defined by (14):

$$\begin{aligned} &\sum_{j=1}^m \eta_j L_{f3}(j,q) \\ &= -\sum_{j=1}^{m-1} \eta_j \partial_j f(u^q) + \eta_m \left\{ \sum_{j=1}^{m-1} \frac{q_j}{q_m} \partial_j f(u^q) - f(u^q) \right\} \\ &= H_f(q) - \sum_{j=1}^m (q_j - \eta_j) \partial_j H_f(q), \end{aligned} \tag{19}$$

where $-\partial H_f = (-\partial_1 H_f, \ldots, -\partial_m H_f)^{\mathrm{T}}$ is the sub-gradient of $-H_f$. See the Supplement for a proof. Conversely, the loss L_{f3} can also be obtained by calculating the canonical representation (7) for the concave function H_f in (14) and then taking η to be a basis vector, $e_1, \ldots, e_m$, one by one in the resulting expression, which is on the left of the second equality in (19). Moreover, by the necessity of the representation (7), we see that L_{f_3} in (18) is the only proper scoring rule with the generalized entropy H_f.

Corollary 1: For a closed, convex function f on $\overline{\mathbb{R}}_+^{m-1}$, any proper scoring rule with H_f in (14) as the generalized entropy can be expressed as L_{f3} in (18), up to possible choices of sub-gradients of f, $\{\partial_j f : j \in [m-1]\}$.

While the preceding use of the canonical representation (7) seems straightforward, our development from Propositions 1 to 3 remains worthwhile. The proper scoring rule L_{f3} in (18) is of simple form, depending explicitly on a dissimilarity function f. Moreover, as shown in Section IV, Proposition 1 can be further exploited to derive new convex losses which are related to classification losses but are not proper scoring rules.

IV. Hinge-Like Losses

The purpose of this section is three-fold. We derive novel hinge-like, convex losses which induce the same generalized entropy as the zero-one, or more generally, cost-weighted classification loss in multi-class settings. Our hinge-like losses are uniformly lower (after suitable alignment) and geometrically simpler (with fewer non-differentiable ridges) than related hinge-like losses in [11] and [12]. Moreover, we show that similar classification regret bounds are achieved by our hinge-like losses and those in [11] and [12]. These regret bounds give a quantitative guarantee on classification calibration as studied in [15] and [16] among others. Finally, we provide a general characterization of losses with the same generalized entropy as the zero-one loss and establish a general classification regret bound for all such losses, beyond the hinge-like losses specifically constructed.

A. Construction of Hinge-Like Losses

We propose a novel approach for constructing hinge-like, convex losses in multi-class settings: we first derive (using Proposition 1) a new loss with actions restricted to the probability simplex Δ_m and its generalized entropy identical to that

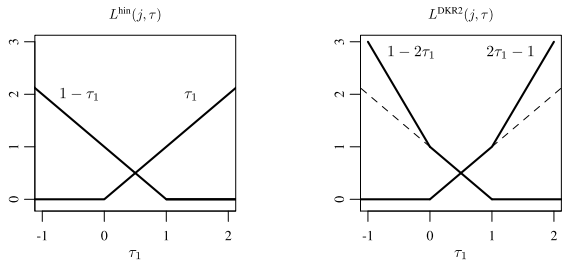


Fig. 2. Two-class hinge loss (left) and hinge-like loss in [12].

of the zero-one or cost-weighted classification loss, and then we find a convex extension of the loss such that its actions are defined on $\mathbb{R}^m$ and its generalized entropy remains unchanged.

As a prologue, we discuss why Proposition 1 is used here instead of Proposition 3 (which is used in Section V for constructing proper scoring rules). It is helpful to consider the two-class setting. The generalized entropy for the zero-one loss is $H^{\text{zo}}(\eta) = \min(\eta_1, \eta_2)$ and the dissimilarity function is $f^{\text{zo}}(t_1) = -\min(1, t_1)$. For this choice of f, it remains valid to apply Proposition 3. With $\partial f^{\text{zo}}(t_1) = -1\{t_1 \le 1\}$, the resulting loss can be shown to be $L_{f3}(1,q) = 1\{q_1 \le q_2\}$ and $L_{f3}(2,q) = 1\{q_2 > q_1\}$, which is just the zero-one loss with action $q \in \Delta_2$. Such a discontinuous loss is computationally intractable for training, even though it is a proper but not strictly proper scoring rule (consistently with Proposition 3). In contrast, as illustrated in Figure 2, the popular hinge loss can be defined, for notational consistency with our later results, such that for $\tau \in \mathbb{R}$,

$$L^{\text{hin}}(1,\tau) = \max(0, 1-\tau), \quad L^{\text{hin}}(2,\tau) = \max(0,\tau). \quad (20)$$

which is continuous and convex in τ and known to yield the same generalized entropy H^{zo} as the zero-one loss (Nguyen et al. 2009). In the following, we show that Proposition 1 can be leveraged to develop convex, hinge-like losses in multi-class settings.

1) Application of Proposition 1: Our application of Proposition 1 is facilitated by the following lemma based on [28]. The lemma gives the conjugate function of the dissimilarity function f^{cw}, corresponding to the generalized entropy H^{cw} in (10) for the cost-weighted classification loss L^{cw}. By definition (15), the dissimilarity function f^{cw} can be calculated as

$$f^{\text{cw}}(t) = -\min_{k\in[m]} C_k^{\text{T}}\tilde{t},$$

where $\tilde{t} = (t^{\text{T}}, 1)^{\text{T}} = (t_1, \ldots, t_{m-1}, 1)^{\text{T}}$ and, as before, $C = (C_1, \ldots, C_m)$ is a column representation of the cost matrix for the cost-weighted classification loss L^{cw}.

Lemma 2: The conjugate of the convex function f^{cw} is

$$f^{\text{cw}*}(s) = \min_{\{\lambda\in\Delta_m : s_j \le -(C\lambda)_j, j\in[m-1]\}} (C\lambda)_m,$$

where $(C\lambda)_j$ denotes the jth component of $C\lambda$ for $j \in [m]$, and the minimum over an empty set is defined as ∞.

From Lemma 2, the domain of $f^{\text{cw}*}$ is a strict subset of $\mathbb{R}^{m-1}$, a phenomenon mentioned earlier in the discussion of Proposition 1:

$$\text{dom}(f^{\text{cw}*}) = \{s \in \mathbb{R}^{m-1} : s_j \le -(C\lambda)_j, j \in [m-1] \text{ for some } \lambda \in \Delta_m\}.$$

The following loss can be obtained from Proposition 1 with the convex function $f = f^{\text{cw}}$ and further simplification with a reparametrization $s_j = -(C\lambda)_j$ for $j \in [m-1]$.

Lemma 3: Define a loss with action space $\mathcal{A} = \Delta_m$ such that for $\lambda \in \Delta_m$,

$$L^{\text{cw2}}(j,\lambda) = \begin{cases} (C\lambda)_j, & j \in [m-1], \\ (C\lambda)_m, & j = m, \end{cases} \quad (21)$$

Then the loss L^{cw2} induces the same generalized entropy H^{cw} in (10) as does the cost-weighted classification loss L^{cw}.

It is interesting that the loss L^{cw2} is defined with actions restricted to the probability simplex Δ_m. But L^{cw2} is not a proper scoring rule, because in general

$$\inf_{\lambda\in\Delta_m}\left\{\sum_{j=1}^m \eta_j L^{\text{cw2}}(j,\lambda)\right\} = \inf_{\lambda\in\Delta_m} \eta^{\text{T}} C\lambda = \min_{k\in[m]} \eta^{\text{T}} C_k$$
$$\neq \eta^{\text{T}} C\eta = \sum_{j=1}^m \eta_j L^{\text{cw2}}(j,\eta).$$

by Lemma 3. In fact, the minimum risk in the first line is achieved by λ equal to a basis vector $e_l \in \Delta_m$ such that $\eta^{\text{T}} C_l = \min_{k\in[m]} \eta^{\text{T}} C_k$.

2) Extension Beyond the Probability Simplex: The loss $L^{\text{cw2}}(j,\lambda)$ is convex (more precisely, linear!) in its action λ when restricted to Δ_m. To handle this restriction, there are several possible approaches. One is to introduce a link function such as the multinomial logistic link $\lambda^h = (\lambda_1^h, \ldots, \lambda_m^h)^{\text{T}}$, where $\lambda_j^h = \exp(h_j)/\sum_{k=1}^m \exp(h_k)$ with $h = (h_1, \ldots, h_{m-1})^{\text{T}} \in \mathbb{R}^{m-1}$ unrestricted and $h_m = 0$ fixed. But the resulting loss $L^{\text{cw2}}(j, \lambda^h)$ would be non-convex in h. Another approach is to define a trivial extension of L^{cw2} such that $L^{\text{cw2}}(j,\lambda) = \infty$ whenever λ lies outside the restricted set Δ_m. But for numerical implementation with this extension, either a link function such as the multinomial logistic link would still be needed, or the predicted action for a new observation is likely to lie outside the probability simplex Δ_m, which then requires additional treatment. By comparison, our approach is to carefully construct an extension of L^{cw2} which remains convex in its action and induces the same generalized entropy H^{cw}, while avoiding the infinity value outside the restricted set Δ_m.

The version of L^{cw2} in (21) with $C = 1_m 1_m^{\text{T}} - I_m$ as in the zero-one loss is

$$L^{\text{zo2}}(j,\lambda) = 1 - \lambda_j, \quad j \in [m], \lambda \in \Delta_m. \quad (22)$$

In the two-class setting, the hinge loss L^{hin} can be shown to be a desired convex extension of the loss L^{zo2}, considered a function of j and λ_1:

$$L^{\text{zo2}}(1,\lambda) = 1 - \lambda_1, \quad L^{\text{zo2}}(2,\lambda) = \lambda_1, \quad \lambda_1 \in [0,1].$$

See Figure 2 for an illustration. In multi-class settings, our first extension of the loss L^{cw2} is as follows, related to the multi-class hinge-like loss in [11].

Proposition 4: Define a loss with action space $\mathcal{A} = \mathbb{R}^{m-1}$ such that for $\tau \in \mathbb{R}^{m-1}$,

$$L^{\mathrm{cw3}}(j,\tau) = \begin{cases} c_{jm}(\tau_m^{(j)})_+ + \sum_{k\in[m-1],k\neq j} c_{jk}\tau_{k+}, \\ \qquad \text{if } j \in [m-1], \\ \sum_{k\in[m-1]} c_{mk}\tau_{k+}, \\ \qquad \text{if } j = m, \end{cases} \tag{23}$$

where $b_+ = \max(0, b)$ for $b \in \mathbb{R}$, and

$$\tau_m^{(j)} = 1 - \tau_j - \textstyle\sum_{k\in[m-1],k\neq j} \tau_{k+}, \quad j \in [m-1]. \tag{24}$$

Then $L^{\mathrm{cw3}}(j,\tau)$ is convex in τ, and coincides with $L^{\mathrm{cw2}}(j,\tilde{\tau})$ provided $\tilde{\tau} \in \Delta_m$, where $\tilde{\tau} = (\tau_1, \ldots, \tau_{m-1}, 1 - \sum_{k=1}^{m-1} \tau_k)^{\mathrm{T}}$. Moreover, L^{cw3} induces the same generalized entropy H^{cw} in (10) as does the cost-weighted classification loss L^{cw}.

A special case of the loss L^{cw3} with the cost matrix $C = 1_m 1_m^{\mathrm{T}} - I_m$ as in the zero-one loss can be expressed such that for $\tau = (\tau_1, \ldots, \tau_{m-1})^{\mathrm{T}} \in \mathbb{R}^{m-1}$,

$$L^{\mathrm{zo3}}(j,\tau) = \begin{cases} \max\left(1 - \tau_j, \sum_{k\in[m-1],k\neq j} \tau_{k+}\right), \\ \qquad \text{if } j \in [m-1], \\ \sum_{k\in[m-1]} \tau_{k+}, \\ \qquad \text{if } j = m, \end{cases}$$

where the summation over an empty set is defined as 0. Then L^{zo3} induces the same generalized entropy H^{zo} in (9) as does the zero-one loss L^{zo}. In the two-class setting, the loss L^{zo3} can be easily seen to coincide with the hinge loss (20).

We compare the new loss with the hinge-like loss in [11] corresponding to the zero-one loss with $C = 1_m 1_m^{\mathrm{T}} - I_m$, which is defined such that for $\gamma \in \mathbb{R}^m$,

$$L^{\mathrm{LLW}}(j,\gamma) = \sum_{k\in[m],k\neq j} (1 + \gamma_k)_+, \quad j \in [m],$$

subject to the restriction that $\sum_{k=1}^m \gamma_k = 0$. The general case of cost-weighted classification can be similarly discussed. To facilitate comparison, a reparametrization of the loss L^{LLW} can be obtained such that for $\tau \in \mathbb{R}^{m-1}$,

$$L^{\mathrm{LLW2}}(j,\tau) = \begin{cases} \sum_{k\in[m-1],k\neq j} \tau_{k+} + \left(1 - \sum_{k\in[m-1]} \tau_k\right)_+, \\ \qquad \text{if } j \in [m-1], \\ \sum_{k\in[m-1]} \tau_{k+}, \\ \qquad \text{if } j = m. \end{cases}$$

In the Supplement, it is shown that $L^{\mathrm{LLW2}}(j,\tau) = L^{\mathrm{LLW}}(j,\gamma)/m$ for $j \in [m]$, provided that $\tau_k = (1 + \gamma_k)/m$ for $k \in [m-1]$. Figure 3 illustrates the two losses L^{zo3} and L^{LLW2} in the three-class setting. The loss L^{zo3} is a tighter convex extension than L^{LLW2} from L^{zo2} in (22), and $L^{\mathrm{zo3}}(j,\tau)$ is geometrically simpler with fewer non-differentiable ridges than $L^{\mathrm{LLW2}}(j,\tau)$ for $j \in [m-1]$. See the Supplement for further discussion of the comparison.

There are various ways in which the loss L^{cw2} can be extended from the probability simplex Δ_m to $\mathbb{R}^m$. We describe another extension, related to the multi-class hinge-like loss in [12] associated with the zero-one loss. The general case of cost-weighted classification can be handled through the transformation (53) in Section V-B, although such a general construction is not discussed in [12].

Proposition 5: Define a loss with action space $\mathcal{A} = \mathbb{R}^{m-1}$ such that for $\tau \in \mathbb{R}^{m-1}$,

$$L^{\mathrm{zo4}}(j,\tau) = 1 - \tilde{\tau}_j + S_\tau^{(j)}, \quad j \in [m], \tag{25}$$

where $(\tilde{\tau}_1, \ldots, \tilde{\tau}_{m-1}) = (\tau_1, \ldots, \tau_{m-1})$, $\tilde{\tau}_m = 1 - \sum_{k=1}^{m-1} \tau_k$, and for $j \in [m]$,

$$S_\tau^{(j)} = \max\left\{0, \tilde{\tau}_j - 1, \frac{\tilde{\tau}_j + \tilde{\tau}_{j(1)} - 1}{2}, \ldots, \frac{\tilde{\tau}_j + \tilde{\tau}_{j(1)} + \cdots + \tilde{\tau}_{j(m-2)} - 1}{m-1}\right\},$$

with $\tilde{\tau}_{j(1)} \geq \ldots \geq \tilde{\tau}_{j(m-1)}$ the sorted components of $\tilde{\tau} = (\tilde{\tau}_1, \ldots, \tilde{\tau}_m)^{\mathrm{T}}$ excluding $\tilde{\tau}_j$. Then $L^{\mathrm{zo4}}(j,\tau)$ is convex in τ, and coincides with $L^{\mathrm{zo2}}(j,\tilde{\tau})$ provided $\tilde{\tau} \in \Delta_m$. Moreover, L^{zo4} induces the same generalized entropy H^{zo} in (9) as does the zero-one loss L^{zo}.

The hinge-like loss in [12] is defined such that for $\gamma \in \mathbb{R}^m$,

$$L^{\mathrm{DKR}}(j,\gamma) = 1 - \gamma_j + S_\gamma, \quad j \in [m],$$

where $S_\gamma = \max\{\gamma_{(1)} - 1, \frac{\gamma_{(1)} + \gamma_{(2)} - 1}{2}, \ldots, \frac{\gamma_{(1)} + \cdots + \gamma_{(m)} - 1}{m}\}$, and $\gamma_{(1)} \geq \ldots \geq \gamma_{(m)}$ are the sorted components of $\gamma \in \mathbb{R}^m$. This loss is invariant to any translation in γ, that is, $L^{\mathrm{DKR}}(j, \gamma - b1_m) = L^{\mathrm{DKR}}(j,\gamma)$ for any $b \in \mathbb{R}$. It suffices to consider $L^{\mathrm{DKR}}(j,\gamma)$ subject to the restriction that $\sum_{k=1}^m \gamma_k = 1$, or equivalently consider the loss

$$L^{\mathrm{DKR2}}(j,\tau) = 1 - \tilde{\tau}_j + S_\tau, \quad j \in [m],$$

where $(\tilde{\tau}_1, \ldots, \tilde{\tau}_{m-1}) = (\tau_1, \ldots, \tau_{m-1})$ and $\tilde{\tau}_m = 1 - \sum_{k=1}^{m-1} \tau_k$ as in (25), and

$$S_\tau = \max\left\{0, \tilde{\tau}_{(1)} - 1, \frac{\tilde{\tau}_{(1)} + \tilde{\tau}_{(2)} - 1}{2}, \ldots, \frac{\tilde{\tau}_{(1)} + \cdots + \tilde{\tau}_{(m-1)} - 1}{m-1}\right\},$$

with $\tilde{\tau}_{(1)} \geq \ldots \geq \tilde{\tau}_{(m)}$ the sorted components of $\tilde{\tau} = (\tilde{\tau}_1, \ldots, \tilde{\tau}_m)^{\mathrm{T}}$. There does not seem to be a direct transformation between the two losses L^{zo4} and L^{DKR2}, in spite of their similar expressions. An illustration is provided by Figures 2 and 4 in two- and three-class settings. The loss L^{zo4} is a tighter convex extension than L^{DKR2} from L^{zo2} in (22), and $L^{\mathrm{zo4}}(j,\tau)$ is geometrically simpler with fewer non-differentiable ridges than $L^{\mathrm{LLW2}}(j,\tau)$ for $j \in [m]$. See the Supplement for further discussion of the comparison.

For motivation, we discuss two possible reasons why the new hinge-like losses can be preferred over their previous counterparts, for example, L^{zo4} over L^{DKR2}. The first reason is based on classification regret bounds. From Proposition 6 and [12, Proposition 5] as discussed in Section IV-B, we have, for L to be either L^{zo4} or L^{DKR2} (which induces the same generalized entropy H^{zo} as the zero-one loss, i.e., $H_{L^{\mathrm{zo4}}} = H_{L^{\mathrm{DKR2}}} = H^{\mathrm{zo}}$),

$$\frac{1}{m} E\{L^{\mathrm{zo}}(Y_0, \tilde{\tau}(X_0)) - H^{\mathrm{zo}}(\pi(X_0))\} \leq E\{L(Y_0, \tau(X_0)) - H^{\mathrm{zo}}(\pi(X_0))\}, \tag{26}$$

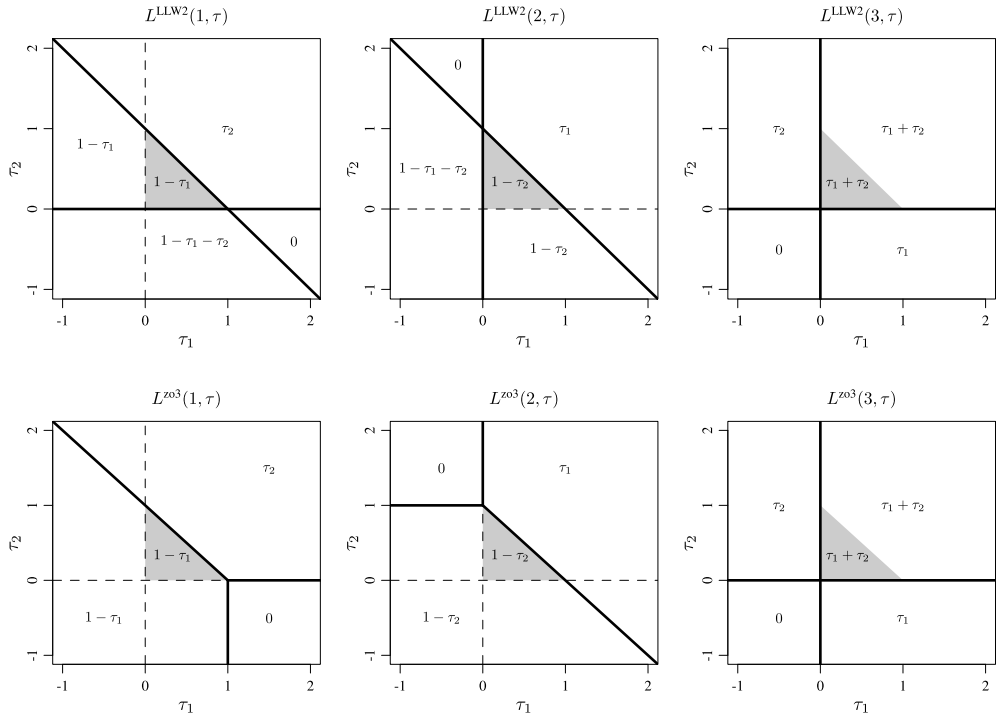


Fig. 3. Three-class hinge-like losses L^{LLW2} (top) and L^{zo3} (bottom). Regions separated by solid lines are associated with the function values indicated.

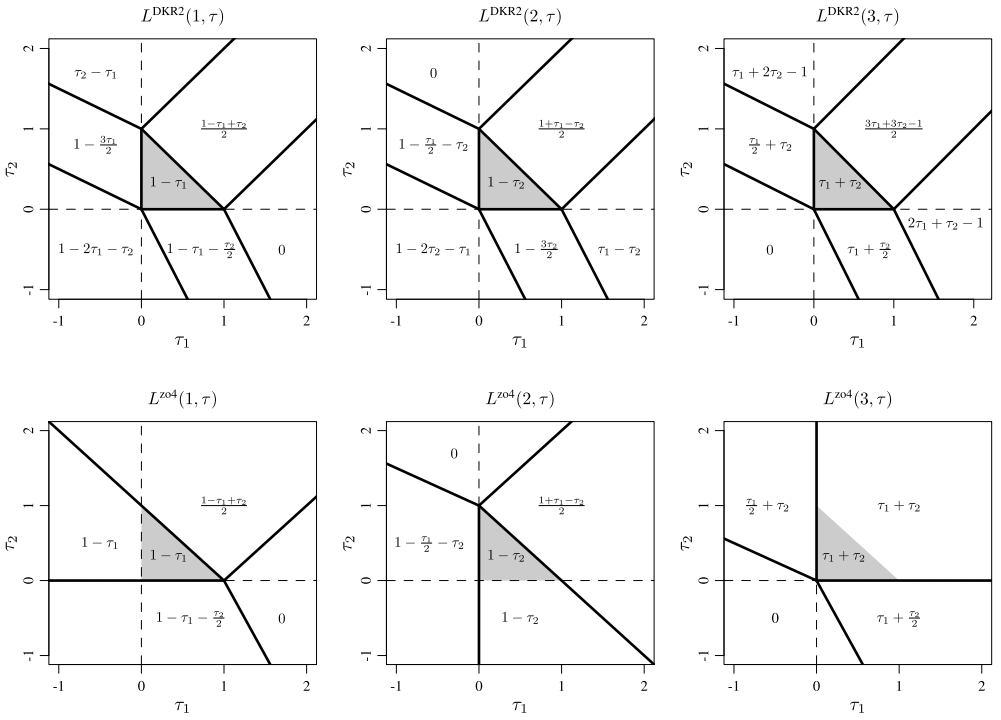


Fig. 4. Three-class hinge-like losses L^{DKR2} (top) and L^{zo4} (bottom). Regions separated by solid lines are associated with the function values indicated.

where (X_0, Y_0) is a test observation as in (2) and $\pi_j(x) = P(Y_0 = j|X_0 = x)$ as in (3). Let τ^{zo4} or τ^{DKR2} be the action function determined by minimizing the expected value of L^{zo4} or L^{DKR2} respectively. Then

$$E\big\{L^{\mathrm{zo4}}(Y_0, \tau^{\mathrm{zo4}}(X_0)) - H^{\mathrm{zo}}(\pi(X_0))\big\}$$
$$\leq E\big\{L^{\mathrm{zo4}}(Y_0, \tau^{\mathrm{DKR2}}(X_0)) - H^{\mathrm{zo}}(\pi(X_0))\big\} \quad (27)$$
$$\leq E\big\{L^{\mathrm{DKR2}}(Y_0, \tau^{\mathrm{DKR2}}(X_0)) - H^{\mathrm{zo}}(\pi(X_0))\big\}. \quad (28)$$

The first line follows by definition of τ^{zo4}. The second line follows because the fact that L^{zo4} is a tighter extension than L^{DKR2} implies that the right-hand side of (26) with $L = L^{\mathrm{zo4}}$

is no greater than that with $L = L^{\mathrm{DKR2}}$ for the same function $\tau(x)$. Hence the right-hand side of (26) with $L = L^{\mathrm{zo4}}$ and $\tau = \tau^{\mathrm{zo4}}$ is never greater and may be much smaller than that with $L = L^{\mathrm{DKR2}}$ and $\tau = \tau^{\mathrm{DKR2}}$. This suggests that the excess zero-one risk on the left-hand side of (26) with $\tau = \tau^{\mathrm{zo4}}$ may be smaller than that with $\tau = \tau^{\mathrm{DKR2}}$. The second reason is based on generalization bounds, in the form of upper bounds on the following quantity

$$\sup_{\tau \in \mathcal{T}} \left| \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \tau(X_i)) - E\big\{ L(Y_0, \tau(X_0)) \big\} \right|,$$

where $\mathcal{T}$ is a certain function class. As seen from [29], such generalization bounds depend linearly on the Lipschitz constant of the loss function $L(j, \tau)$ in τ. The fact that L^{zo4} is a tighter extension than L^{DKR2} implies a smaller Lipschitz constant for L^{zo4}, which leads to a smaller generalization bound for L^{zo4}. While the preceding reasoning may indicate potential advantages of the new hinge-like losses over existing ones, our discussion is heuristic and further comparison of these losses can be studied in future work.

B. Regret Bounds for Hinge-Like Losses

The preceding section mainly focuses on constructing multiclass hinge-like losses which induce the generalized entropy L^{zo} or H^{cw} as does the zero-one or cost-weighted classification loss, while achieving certain desirable properties geometrically compared with hinge-like losses in [11] and [12]. Here we derive classification regret bounds, which compare the regrets of our hinge-like losses with those of the zero-one and cost-weighted losses, where the actions are take from those of the hinge-like losses by a prediction mapping. Such bounds provide a quantitative guarantee on classification calibration, a qualitative property which leads to infinite-sample classification consistency under suitable technical conditions [15], [16].

Proposition 6: The following regret bounds hold for the hinge-like losses L^{cw3} and L^{zo4}.

(i) For $\eta \in \Delta_m$ and $\tau \in \mathbb{R}^{m-1}$, $m^{-1} B_{L^{\mathrm{cw}}}(\eta, \tau^\dagger) \le B_{L^{\mathrm{cw3}}}(\eta, \tau)$, that is,

$$\frac{1}{m} \left\{ \sum_{j=1}^{m} \eta_j L^{\mathrm{cw}}(j, \tau^\dagger) - H^{\mathrm{cw}}(\eta) \right\} \le \sum_{j=1}^{m} \eta_j L^{\mathrm{cw3}}(j, \tau) - H_{L^{\mathrm{cw3}}}(\eta), \tag{29}$$

where $\tau^\dagger = (\tau_1, \ldots, \tau_{m-1}, 1 - \sum_{k=1}^{m-1} \tau_{k+})^{\mathrm{T}}$.

(ii) For $\eta \in \Delta_m$ and $\tau \in \mathbb{R}^{m-1}$, $m^{-1} B_{L^{\mathrm{zo}}}(\eta, \tilde{\tau}) \le B_{L^{\mathrm{zo4}}}(\eta, \tau)$, that is,

$$\frac{1}{m} \left\{ \sum_{j=1}^{m} \eta_j L^{\mathrm{zo}}(j, \tilde{\tau}) - H^{\mathrm{zo}}(\eta) \right\} \le \sum_{j=1}^{m} \eta_j L^{\mathrm{zo4}}(j, \tau) - H_{L^{\mathrm{zo4}}}(\eta), \tag{30}$$

where $\tilde{\tau} = (\tau_1, \ldots, \tau_{m-1}, 1 - \sum_{k=1}^{m-1} \tau_k)^{\mathrm{T}}$.

The regret bounds (29) and (30) directly lead to classification calibration, which can be defined as follows, allowing a prediction mapping [15], [16]. For a loss $L(j, \gamma)$ with action space $\mathcal{A}$, let $\sigma = (\sigma_1, \ldots, \sigma_m)^{\mathrm{T}} : \mathcal{A} \to \mathbb{R}^m$ be a prediction mapping which carries an action in $\mathcal{A}$ to a vector in $\mathbb{R}^m$, to be used as the corresponding action in the zero-one or cost-weighted classification loss. The prediction mapping can be defined directly as the identity mapping, $\sigma(\gamma) = \gamma$, in the case of $\mathcal{A} \subset \mathbb{R}^m$, but needs to convert an action γ to a vector in $\mathbb{R}^m$ in the case of $\mathcal{A} \subset \mathbb{R}^{m-1}$. A loss $L(j, \gamma)$ with action space $\mathcal{A}$ and prediction mapping $\sigma(\cdot)$ is said to be classification calibrated for the zero-one loss if for any $\eta \in \Delta_m$ and $k \in [m]$ with $\eta_k < \max_{j \in [m]} \eta_j$,

$$\inf_{\gamma \in \mathcal{A}} \left\{ \sum_{j=1}^{m} \eta_j L(j, \gamma) - H_L(\eta) : \sigma_k(\gamma) = \max_{j \in [m]} \sigma_j(\gamma) \right\} > 0. \tag{31}$$

For $L = L^{\mathrm{zo4}}$ and $\sigma(\tau) = \tilde{\tau}$, inequality (30) implies that the left-hand side of (31) is no smaller than $(-\eta_k + \max_{j \in [m]} \eta_j)/m > 0$. Hence L^{zo4} is classification calibrated for the zero-one loss. Similarly, a loss $L(j, \gamma)$ with action space $\mathcal{A}$ and prediction mapping $\sigma(\cdot)$ is said to be classification calibrated for cost-weighted classification with cost matrix C if for any $\eta \in \Delta_m$ and $k \in [m]$ with $\eta^{\mathrm{T}} C_k > \max_{j \in [m]} \eta^{\mathrm{T}} C_j$,

$$\inf_{\gamma \in \mathcal{A}} \left\{ \sum_{j=1}^{m} \eta_j L(j, \gamma) - H_L(\eta) : \sigma_k(\gamma) = \max_{j \in [m]} \sigma_j(\gamma) \right\} > 0. \tag{32}$$

For $L = L^{\mathrm{cw3}}$ and $\sigma(\tau) = \tau^\dagger$, inequality (29) implies that the left-hand side of (32) is no smaller than $(\eta^{\mathrm{T}} C_k - \min_{j \in [m]} \eta^{\mathrm{T}} C_j)/m > 0$. Hence the loss L^{cw3} is classification calibrated with $\sigma(\tau) = \tau^\dagger$ for cost-weighted classification.

There is an interesting feature in the regret bound (29) for $L^{\mathrm{cw3}}(j, \tau)$, compared with the regret bound (30) for $L^{\mathrm{zo4}}(j, \tau)$. The prediction mapping associated with the loss $L^{\mathrm{cw3}}(j, \tau)$ with $\tau \in \mathbb{R}^{m-1}$ is $\tau^\dagger = (\tau_1, \ldots, \tau_{m-1}, 1 - \sum_{k=1}^{m-1} \tau_{k+})^{\mathrm{T}}$, whose components may sum to less than one, instead of $\tilde{\tau} = (\tau_1, \ldots, \tau_{m-1}, 1 - \sum_{k=1}^{m-1} \tau_k)^{\mathrm{T}}$, whose components necessarily sum to one. Although this difference warrants further study, using $\tau^\dagger$ instead of $\tilde{\tau}$ for classification ensures that the predicted value for mth class, $1 - \sum_{k=1}^{m-1} \tau_{k+}$, is not affected by any negative components among $(\tau_1, \ldots, \tau_{m-1})$. For example, if $m = 3$ and $(\tau_1, \tau_2) = (.6, -.3)$, then $1 - \sum_{k=1}^{2} \tau_{k+} = .4$ but $1 - \sum_{k=1}^{2} \tau_k = .7$. Using $\tau^\dagger$ means that class 1 is predicted, whereas using $\tilde{\tau}$ means that class 3 is predicted, which seems to be artificially caused by the negative value of τ_2. See Figure 5 for an illustration and Proposition 8 for an explanation.

The regret bounds (29) and (30) for the losses L^{cw3} and L^{zo4} are similar to those for the losses L^{LLW2} and L^{DKR2} in [12]. In fact, the regret bound for L^{LLW2} in Duchi et al. can be seen as (29) with $L^{\mathrm{cw3}}(j, \tau)$ and $L^{\mathrm{cw}}(j, \tau^\dagger)$ replaced by $L^{\mathrm{LLW2}}(j, \tau)$ and $L^{\mathrm{cw}}(j, \tilde{\tau})$ respectively, because L^{LLW2} is L^{LLW} multiplied by m after a reparametrization noted earlier.

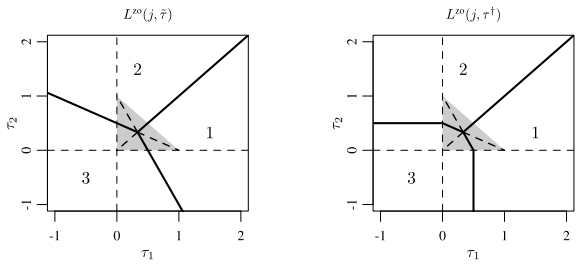


Fig. 5. Classification using the prediction mapping $\tilde{\tau}$ (left) or $\tau^\dagger$ (right), defined in Proposition 6, with $\tau = (\tau_1, \tau_2) \in \mathbb{R}^2$ for $m = 3$. Each region separated by solid lines from others is classified by the index of a maximum component of $\tilde{\tau}$ or $\tau^\dagger$.

The regret bound for L^{DKR2} in Duchi et al. can be seen as (30) with $L^{\mathrm{zo4}}(j,\tau)$ replaced by $L^{\mathrm{DKR2}}(j,\tau)$ and with $L^{\mathrm{zo}}(j,\tilde{\tau})$ unchanged. As mentioned in Section IV-A, further research is desired to compare these hinge-like losses in theory and empirical evaluation.

C. General Characterization and Regret Bounds

Our new hinge-like losses are explicitly derived to induce the same generalized entropy as the zero-one or cost-weighted classification loss, and shown to achieve comparable regret bounds to those for existing hinge-like losses. In this section, we provide a general result indicating that all losses with the same generalized entropy as the zero-one loss achieve a classification regret bound similarly as in Proposition 6. This result relies on a general characterization of such losses in terms of the value manifold defined below.

For a loss $L(j,\gamma)$ with action space $\mathcal{A}$, the value manifold is defined as $\mathcal{S}_L = \overline{\mathrm{conv}}(\mathcal{R}_L)$, where $\overline{\mathrm{conv}}$ denotes the closure of the convex hull and

$$\mathcal{R}_L = \{(L(1,\gamma),\ldots,L(m,\gamma))^{\mathrm{T}} : \gamma \in \mathcal{A}\}.$$

The concept of the set $\mathcal{R}_L$ and its convex hull, $\mathrm{conv}(\mathcal{R}_L)$, also plays an important role in [16], where the admissibility of $\mathrm{conv}(\mathcal{R}_L)$ can be equivalently defined as that of $\mathcal{S}_L$ because $\mathrm{conv}(\mathcal{R}_L)$ and $\mathcal{S}_L$ share the same boundary, denoted as $\partial\mathcal{S}_L$. Then the generalized entropy of L can be expressed such that for any $\eta \in \Delta_m$,

$$H_L(\eta) = \inf_{\gamma\in\mathcal{A}} \left\{ \sum_{j=1}^m \eta_j L(j,\gamma) \right\} = \inf_{z\in\mathcal{R}_L} \eta^{\mathrm{T}} z = \inf_{z\in\mathcal{S}_L} \eta^{\mathrm{T}} z, \tag{33}$$

similarly as in [16], Eq. (7). For the zero-one loss L^{zo}, the value manifold is denoted as

$$\mathcal{S}^{\mathrm{zo}} = \Big\{ (z_1,\ldots,z_m)^{\mathrm{T}} : \textstyle\sum_{j=1}^m z_j = m-1 \\ \text{and } 0 \le z_1,\ldots,z_m \le 1 \Big\}.$$

The set $\mathcal{S}^{\mathrm{zo}}$ is an $(m-1)$-dimensional polytope in $\mathbb{R}^m$, where each vertex is a m-dimensional vector with one component 0 and the remaining 1.

Proposition 7: A loss $L(j,\gamma)$ induces the same generalized entropy as the zero-one loss, i.e., $H_L(\eta) = H^{\mathrm{zo}}(\eta) = 1 - \max_{k\in[m]} \eta_k$ for $\eta \in \Delta_m$ if and only if

$$\mathcal{S}^{\mathrm{zo}} \subset \mathcal{S}_L \subset \mathcal{S}^{\mathrm{zo}*}, \tag{34}$$

where $S^{\mathrm{zo}*} = \{z + b : z \in \mathcal{S}^{\mathrm{zo}}, b \in \mathbb{R}^m_+\}$, also denoted as $\mathcal{S}^{\mathrm{zo}} + \mathbb{R}^m_+$.

Figure 6 shows, in the three-class setting, the value manifolds for the zero-one and several hinge-like losses with the same generalized entropy as the zero-one loss. These value manifolds all satisfy the inclusion property as stated in Proposition 7.

The following result establishes a general link from the generalized entropy of the zero-one loss to classification regret bounds. The link involves a particular prediction mapping $\sigma_L(\gamma)$, defined from the negative values of a given loss L. Such a prediction mapping is also exploited to study classification calibration in [16], with an additional assumption that the value manifold $\mathcal{S}_L$ is symmetric.

Proposition 8: Suppose that a loss $L(j,\gamma)$ with action space $\mathcal{A}$ induces the same generalized entropy as the zero-one loss (i.e., $H_L = H^{\mathrm{zo}}$). Then for $\eta \in \Delta_m$ and $\gamma \in \mathcal{A}$, $m^{-1} B_{L^{\mathrm{zo}}}(\eta, \sigma_L(\gamma)) \le B_L(\eta,\gamma)$, that is,

$$\frac{1}{m}\left\{ \sum_{j=1}^m \eta_j L^{\mathrm{zo}}(j,\sigma_L(\gamma)) - H^{\mathrm{zo}}(\eta) \right\} \le \sum_{j=1}^m \eta_j L(j,\gamma) - H_L(\eta), \tag{35}$$

where $\sigma_L(\gamma) = (-L(1,\gamma),\ldots,-L(m,\gamma))^{\mathrm{T}}$. Moreover, the loss $L(j,\gamma)$, with the prediction mapping σ_L, is classification calibrated for the zero-one loss.

Proposition 8 provides a theoretical support for our approach in constructing hinge-like losses with the same generalized entropy as the zero-one loss in Section IV-A. The regret bound (35) implies classification calibration similarly as discussed in Section IV-B. Compared with [16, Section 4], our result gives a more concrete sufficient condition for achieving classification calibration, in addition to a quantitative guarantee. Because the loss is unspecified except for its generalized entropy being H^{zo} in Proposition 8, our result is of a different nature from [12, Proposition 4], where classification calibration is shown to be achieved by the specific loss L_H in (13) given a concave function H (as its generalized entropy). Classification regret bounds then need to be proved on a case-by-case basis, for example, for the loss L^{DKR} given H^{zo} in [12].

By the nature of the zero-one loss, the regret bound (35) remains valid with $\sigma_L(\gamma)$ replaced by another prediction mapping $\sigma(\gamma)$ subject to the monotonicity property that the components in $\sigma(\gamma)$ are in the same order as in $\sigma_L(\gamma)$ for any $\gamma \in \mathcal{A}$. Then the regret bounds for L^{zo3}, as a special case of L^{cw3}, and L^{zo4} in Proposition 6 can be deduced from (35), because the monotonicity property is satisfied by the prediction mappings $\tau^\dagger$ and $\tilde{\tau}$ used in (29) and (30). See the Supplement for details. Possible extensions of Proposition 8 to cost-weighted classification can be studied in future work.

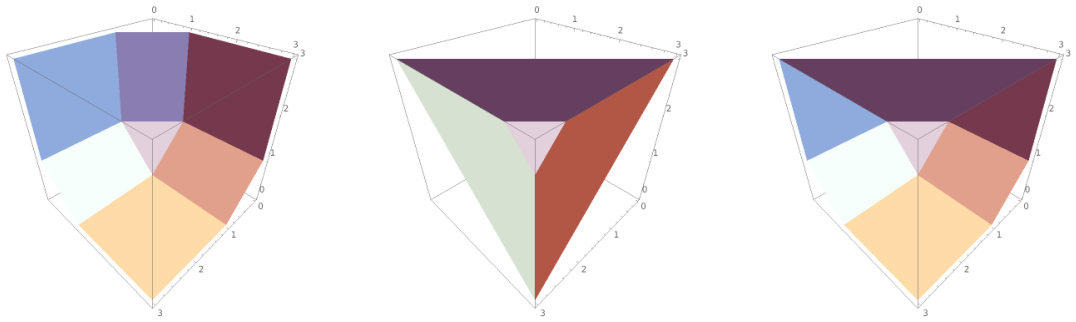


Fig. 6. The boundaries of value manifolds for three-class hinge-like losses $L^{\mathrm{DKR2}}/L^{\mathrm{zo4}}$ (left), L^{LLW2} (middle) and L^{zo3} (right). The triangle polytope in the center of each plot is the value manifold $\mathcal{S}^{\mathrm{zo}}$ for the zero-one loss. The boundary of $\mathcal{S}^{\mathrm{zo}*}$, defined as $\mathcal{S}^{\mathrm{zo}} + \mathbb{R}^m_+$, is the same as in the left plot.

D. Numerical Illustration

We provide a simple numerical experiment to illustrate potential advantages of the proposed hinge-like losses. As motivated by [30] in the binary setting, we consider a three-class setting to study misclassification rates committed by linear classifiers obtained through minimizing hinge-like (surrogate) losses when a small, optimal error rate can be achieved by a linear classifier based on minimizing the zero-one loss.

We investigate the proposed L^{zo4} loss, and compare it with L^{DKR}, L^{LLW}, and the hinge-like losses in [13] and [14], denoted as WW and CS. In the comparison, we also include the multinomial logistic regression, as well as the two common strategies based on binary classification, one-versus-one (OVO) and one-versus-all (OVA). For all methods, the action function $\tau(x)$ is linear in x with an intercept. The multinomial logistic regression is implemented through the `glmnet` R package [31], and all other methods are implemented through the `CVXR` R package [32] based on the `mosek` solver [33].

We consider a three-class data generating process as follows: 98% of (X, Y) are drawn such that $X|Y = j \sim N(\mu_j, 0.1^2 I)$ and $P(Y = j) = 1/3$ for $j = 1, 2, 3$, and the remaining 2% of (X, Y) are outliers with $Y = 1$ and X drawn from $N(\mu_4, 0.1^2 I)$, where $(\mu_1, \ldots, \mu_4)$ are $(-1, 0)$, $(0, 0.25)$, $(1, 0)$, and $(1 + \rho, 0)$ respectively. The parameter $\rho > 0$ controls the separation degree (how far away) between outliers and the main data. We train all methods on a training set of size 3,000 and calculate misclassification error rates on a test set of size 50,000. For each method, we do not apply regularization to isolate the effect of the loss function given the large training data size. We vary ρ from 0 to 2.5, i.e., moving outliers away while keeping main data fixed, to investigate how each method's performance changes.

For example, Figure 7 shows the training data and classification regions and test error rates from different methods for $\rho = 2$. The optimal error rate for linear classifiers is virtually 2%, achieved by the Bayes classifier on the main data (hence correctly classifying 98% of the entire data). In principle, this optimal linear classifier can be derived by minimizing the zero-one loss. However, minimizing hinge-like (surrogate) losses lead to error rates greater than the optimal rate in various degrees. For $\rho = 2$ as shown in Figure 7, the proposed L^{zo4} method achieves the smallest error rate, 4.27%, among all hinge-like methods under study.

In Figure 8a, we plot the test error rates over ρ from all methods. Only the proposed method, CS, and OVO achieve near-optimal error rates (around 2.4%) when $\rho < 0.5$. CS is slightly better than our method when $\rho < 1.5$, but its error rate quickly rises and becomes higher than ours, as ρ further increases. Compared with the other three classification calibrated methods, DKR, LLW, and multinomial logistic, our method achieves smaller error rates, sometimes to a large extent, over the range of ρ. The error rate from LLW decreases as ρ increases, which is paradoxical and may be further studied.

Figure 8b shows the loss values as well as test error rates over ρ from our method, DKR, and LLW, where the loss functions can be perfectly aligned on the probability simplex. In accordance with (27)–(28), depending on the proposed L^{zo4} loss being a tighter extension outside the probability simplex, the loss values from our method are always smaller than those from DKR and LLW. Then as suggested by the regret bound (26), our method also achieves smaller error rates than the other two methods, even though the regret bound only gives an upper bound on the error rates for all three methods.

V. Proper Scoring Rules

We investigate proper scoring rules in two similar directions as in Section IV: first deriving new proper scoring rules (while recovering existing ones) and second establishing classification regret bounds with respect to the zero-one or cost-weighted classification loss.

A. Examples of Proper Scoring Rules

We examine various examples of multi-class proper scoring rules, obtained from Proposition 3. In particular, it is of interest to study how commonly used two-class losses can be extended to multi-class ones. These examples lead to new multi-class proper scoring rules and shed new light on existing ones. See Section IV for a discussion of multi-class hinge-like losses related to zero-one classification losses, derived using Proposition 1.

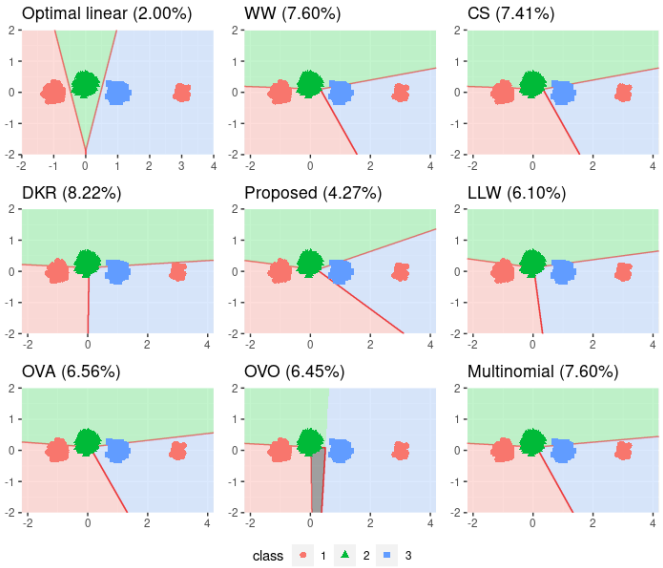
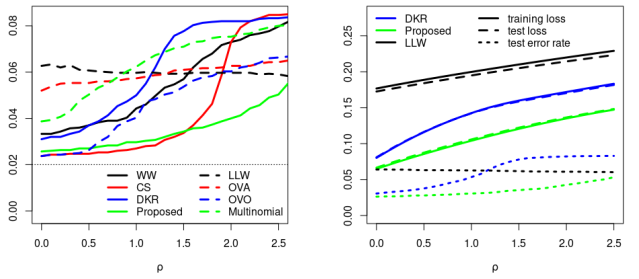


Fig. 7. Classification regions and test error rates when the separation degree $\rho = 2$. The gray area in the OVO's plot indicates where pairwise votes are tied and thus prediction can not be uniquely determined.



(a) Test error rates. (b) Loss values and test error rates.

Fig. 8. Trend plots over separation degree ρ. Loss values on the training and test sets are reported for three methods where the loss functions can be aligned.

1) Two-Class Losses: For two-class classification ($m = 2$) and a univariate convex function f_0 on $\overline{\mathbb{R}}_+$, the proper scoring rule (18) in Proposition 3 reduces to

$$L_{f_0}(j,q) = -\mathbb{1}_1(j)\partial f_0(u^q) + \mathbb{1}_2(j)\{u^q\partial f_0(u^q) - f(u^q)\}, \quad j = 1, 2, \quad (36)$$

where $u^q = q_1/q_2$ for $q = (q_1, q_2)^{\mathrm{T}} \in \Delta_2$, ∂f_0 denotes a sub-gradient of f_0, and $\mathbb{1}_k(j)$ is an indicator defined as 1 if $j = k$ or 0 otherwise. For a twice-differentiable function f_0, the gradient of the loss (36) can be directly calculated as

$$\frac{\mathrm{d}}{\mathrm{d}q_1} L_{f_0}(j,q) = -\{\mathbb{1}_1(j) - q_1\}w(q_1), \quad (37)$$

where $\mathrm{d}/\mathrm{d}q_1$ denotes a derivative taken with respect to q_1 with $q_2 = 1 - q_1$, and the weight function $w(q_1) = f_0''(u^q)/q_2^3$ with f_0'' the second derivative of f_0. From (37), $L_{f_0}(j,q)$ can be put into an integral representation in terms of $w(\cdot)$ and the cost-weighted binary classification loss [4], [21], [34]. The formula (36) in terms of f_0 differs from the integral representation or the canonical representation (7), even though they can be transformed into each other.

For concreteness, consider the following examples of two-class losses:

- *Likelihood Loss*: $L_\ell(j,q) = -\log q_j$ with $f_0 = t\log t - (1+t)\log(1+t)$,
- *Exponential Loss*: $L_e(j,q) = \mathbb{1}_1(j)\sqrt{q_2/q_1} + \mathbb{1}_2(j)\sqrt{q_1/q_2}$ with $f_0 = (\sqrt{t} - 1)^2$,
- *Calibration Loss*: $L_c(j,q) = \{\mathbb{1}_1(j)(q_2/q_1) + \mathbb{1}_2(j)\log(q_1/q_2)\}/2$ with $f_0 = -(\log t)/2$,

where all the expressions for $L(j,q)$ are up to additive constants in q. See Supplement Table S1 for further information. While the likelihood loss is tied to maximum likelihood estimation, the exponential loss is associated with boosting algorithms [35], [36]. The calibration loss is studied in [37] for logistic regression, where the fitted probabilities are used for inverse probability weighting. See the Supplement for a discussion on convexity of these losses with a logistic link.

Remark 2: The loss (36) was also derived in [38] for training a discriminator in generative adversarial learning [39], [40]. In that context, the loss for training a generator is, in a nonparametric limit, the negative Bayes risk from discrimination or the f_0-divergence by relationship (12) with $\pi^0 = 1_m/m$,

$$D_{f_0}(P_1\|P_2) = -\inf_{q:\mathcal{X}\to\Delta_2} E\{L_{f_0}(Y, q(X))\},$$

where P_1 is the data distribution represented by training data and P_2 is the model distribution represented by simulated data from the generator. Hence the generator can be trained to minimize various f_0-divergences, including forward and reverse Kullback–Liebler and Hellinger divergences. See Supplement Table S1 in [38].

2) Multi-Class Pairwise Losses: There can be numerous choices for extending a two-class loss (36) to multi-class ones, just as a univariate convex function f_0 can be extended in multiple ways to multivariate ones. A simple approach is to use an additive extension, $f(u_1, \ldots, u_{m-1}) = \sum_{k=1}^{m-1} f_0(u_k)$. The corresponding loss (18) is then

$$L_{f_0}^{\mathrm{pw,a}}(j,q) = \sum_{k=1}^{m-1}\Big[-\mathbb{1}_k(j)\partial f_0\Big(\frac{q_k}{q_m}\Big) + \mathbb{1}_m(j)\Big\{\frac{q_k}{q_m}\partial f_0\Big(\frac{q_k}{q_m}\Big) - f_0\Big(\frac{q_k}{q_m}\Big)\Big\}\Big], \quad (38)$$

Equivalently, the loss (38) can be obtained by applying the two-class loss (36) to a pair of classes, k and m, and summing up such pairwise losses for $k \in [m-1]$. In this sense, the loss (38) can be interpreted as performing multi-class classification via pairwise comparison of each class $k \in [m-1]$ with class m.

The preceding loss (38) is asymmetric with class m compared with the remaining classes $k \in [m-1]$. A symmetrized version can be obtained by varying the choice of a base class

and summing up the resulting losses as

$$\begin{aligned} & L_{f_0}^{\mathrm{pw,s}}(j,q) \\ &= \sum_{l,k\in[m],k\neq l} \Bigg[-\mathbb{1}_k(j)\partial f_0(\frac{q_k}{q_l}) \\ &\qquad + \mathbb{1}_l(j)\left\{ \frac{q_k}{q_l}\partial f_0(\frac{q_k}{q_l}) - f_0(\frac{q_k}{q_l}) \right\} \Bigg] \\ &= \sum_{k\in[m],k\neq j} \left\{ -\partial f_0(\frac{q_j}{q_k}) + \frac{q_k}{q_j}\partial f_0(\frac{q_k}{q_j}) - f_0(\frac{q_k}{q_j}) \right\}. \end{aligned} \tag{39}$$

See the Supplement for a proof. The symmetrized loss (39) can also be deduced from (18) with the choice $f(u_1,\ldots,u_{m-1}) = \sum_{l,k\in[m],k\neq l} u_l f_0(\frac{u_k}{u_l})$, where $u_m \equiv 1$. In spite of the interpretation via pairwise comparison, our approach involves optimizing the loss (38) or (39) *jointly* over $q \in \Delta_m$ using all m labels, and hence differs from the usual one-against-all or all-pairs approach, which performs binary classification with 2 reduced labels *separately* for multiple times. Further comparison of these approaches can be studied in future work.

Consider a multinomial logistic link $q^h = (q_1^h,\ldots,q_m^h)^{\mathrm{T}}$, where $h = (h_1,\ldots,h_m)^{\mathrm{T}}$ and

$$q_j^h = \frac{\exp(h_j)}{\sum_{k=1}^m \exp(h_k)}, \quad j\in[m]. \tag{40}$$

The link is a natural extension of the logistic link, because log ratios between $(q_1,\ldots,q_m)$ are related to contrasts between $(h_1,\ldots,h_m)$. To remove over-parametrization, a restriction is often imposed such as $h_m \equiv 0$ or $\sum_{k=1}^m h_k \equiv 0$. By the additive construction, the composite losses obtained from (38) and (39) can be easily shown to be convex in h whenever the two-class loss (36) with a logistic link $q_1^{h_0}/q_2^{h_0} = \exp(h_0)$ is convex in h_0.

For the two-class likelihood, exponential, and calibration losses above, the pairwise extensions (39) can be calculated as follows:

- *Pairwise likelihood loss*: $L_\ell^{\mathrm{pw,s}}(j,q) = 2\sum_{k\in[m],k\neq j}\log(1+\frac{q_k}{q_j})$,
- *Pairwise exponential loss*: $L_e^{\mathrm{pw,s}}(j,q) = 2\sum_{k\in[m],k\neq j}\sqrt{\frac{q_k}{q_j}}$,
- *Multi-class calibration loss*: $L_c^{\mathrm{pw,s}}(j,q) = \sum_{k\in[m],k\neq j}\{\log(\frac{q_k}{q_j}) + \frac{q_k}{q_j}\}/2$,

where additive constants in q are dropped for simplicity. See Supplement Table S1 for the expressions of the corresponding f, H, and gradients. By convexity of the associated two-class composite losses [4], we see that with the multinomial logistic link (40), the three composite losses, $L_\ell^{\mathrm{pw,s}}(j,q^h)$, $L_e^{\mathrm{pw,s}}(j,q^h)$, and $L_c^{\mathrm{pw,s}}(j,q^h)$, are all convex in h. In particular, the pairwise exponential composite loss is

$$L_e^{\mathrm{pw,s}}(j,q^h) = 2\sum_{k\in[m],k\neq j} \mathrm{e}^{(h_k-h_j)/2},$$

which is associated with multi-class boosting algorithms AdaBoost.M2 [41] or AdaBoost.MR [42]. See [7] for further study. The pairwise likelihood and calibration losses appear to be new. The pairwise likelihood loss, with $m \geq 3$, differs from the standard likelihood loss based on multinomial data, which will be discussed later. The multi-class calibration loss has recently been re-derived and studied for propensity score estimation with multi-valued treatments [43].

3) Multi-Class Simultaneous Losses: Apparently, there exist various multi-class proper scoring rules, which cannot be expressed as pairwise losses (38) or (39) and hence will be referred to as simultaneous losses. A notable example as mentioned above is the standard likelihood loss (or the logarithmic scoring rule) for multinomial data, $L(j,q) = -\log q_j$. In fact, a large class of multi-class simultaneous losses can be defined with the generalized entropy in the form

$$H_\beta(q) = \begin{cases} \|q\|_\beta, & \text{if } \beta\in(0,1), \\ -\|q\|_\beta, & \text{if } \beta\in(1,\infty), \end{cases}$$

where $\|q\|_\beta = \{\sum_{j=1}^m q_j^\beta\}^{1/\beta}$ is the L_β norm. The corresponding dissimilarity function is $f_\beta(t) = -\|\tilde{t}\|_\beta$ if $\beta\in(0,1)$ or $\|\tilde{t}\|_\beta$ if $\beta\in(1,\infty)$, where $\tilde{t} = (t_1,\ldots,t_{m-1},1)^{\mathrm{T}}$. The resulting scoring rule can be calculated by (18) as

$$L_\beta(j,q) = \begin{cases} (q_j/\|q\|_\beta)^{\beta-1}, & \text{if } \beta\in(0,1), \\ -(q_j/\|q\|_\beta)^{\beta-1}, & \text{if } \beta\in(1,\infty). \end{cases} \tag{41}$$

The case $\beta > 1$ is called a pseudo-spherical score [20], [44]. The limiting case $\beta\to 1$ is also known to yield the logarithmic score, $L(j,q) = -\log q_j$, after suitable rescaling. The case $\beta\in(0,1)$ seems previously unstudied. There are also two additional limiting cases as $\beta\to 0+$ or ∞. See Supplement Table S1 for further details.

Proposition 9: Define a rescaled version of H_β as

$$H_\beta^{\mathrm{r}}(q) = \frac{\|q\|_\beta - 1}{m^{1/\beta-1}-1}, \tag{42}$$

if $\beta\in(0,1)\cup(1,\infty)$, and $H_\beta^{\mathrm{r}}(q) = \lim_{\beta'\to\beta} H_{\beta'}^{\mathrm{r}}(q)$, if $\beta = 0,1,\infty$. Then the following proper scoring rules are obtained.

(i) Simultaneous exponential loss ($\beta = 0$): $L_0^{\mathrm{r}}(j,q) = (\prod_{k=1}^m \frac{q_k}{q_j})^{1/m}$ corresponding to $H_0^{\mathrm{r}}(q) = m(\prod_{j=1}^m q_j)^{1/m}$.

(ii) Pairwise exponential loss ($\beta = 1/2$): $L_{1/2}^{\mathrm{r}}(j,q) = (m-1)^{-1}\sum_{k\in[m],k\neq j}\sqrt{\frac{q_k}{q_j}}$ corresponding to $H_{\frac{1}{2}}^{\mathrm{r}}(q) = (m-1)^{-1}(\|q\|_{\frac{1}{2}} - 1)$.

(iii) Multinomial likelihood loss ($\beta = 1$): $L_1^{\mathrm{r}}(j,q) = -(\log m)^{-1}\log q_j$ corresponding to $H_1^{\mathrm{r}}(q) = -(\log m)^{-1}\sum_{j=1}^m q_j\log q_j$.

(iv) Multi-class zero-one loss ($\beta = \infty$): $L_\infty^{\mathrm{r}}(j,q) = (1-m^{-1})^{-1}1\{j\neq \mathrm{argmax}_{k\in[m]}q_k\}$ corresponding to $H_\infty^{\mathrm{r}}(q) = (1-m^{-1})^{-1}(1-\max_{j\in[m]}q_j)$.

Moreover, with a multinomial logistic link (40), the composite loss $L_\beta^{\mathrm{r}}(j,q^h)$ is convex in h if $\beta\in[0,1]$, but non-convex in h if $\beta > 1$.

There are several interesting features. First, with a multinomial logistic link (40), the scoring rule $L_0^{\mathrm{r}}(j,q)$ leads to a composite loss

$$L_0^{\mathrm{r}}(j,q^h) = \mathrm{e}^{\frac{1}{m}\sum_{k=1}^m (h_k-h_j)},$$

which coincides with the exponential loss in [6]. For this reason, $L_0^{\mathrm{r}}(j,q)$ is called the simultaneous exponential loss. Moreover, the scoring rule $L_{1/2}^{\mathrm{r}}(j,q)$ yields, up to a multiplicative factor, the pairwise exponential loss $L_e^{\mathrm{pw,s}}(j,q)$, which is connected with the boosting algorithms in [41] and [42] as mentioned earlier. The logarithmic rule $L_1^{\mathrm{r}}(j,q)$ corresponds to the standard likelihood loss based on multinomial data. Finally, the loss $L_\infty^{\mathrm{r}}(j,q)$ obtained as $\beta\to\infty$ recovers the

zero-one loss, which is a proper scoring rule (although not strictly proper). Further research is desired to study relative merits of these losses.

B. Regret Bounds for Proper Scoring Rules

We derive classification regret bounds for proper scoring rules, which compare the regrets of the proper scoring rules (as losses) with those of the corresponding zero-one and cost-weighted classification losses, similarly as in Proposition 6 for hinge-like losses. All such bounds are also called surrogate regret bounds, in the sense that the a proper scoring rule or a hinge-like loss can be considered a surrogate criterion for the zero-one or cost-weighted classification loss. Similarly as discussed in Section IV-B, these results provide a quantitative guarantee on classification calibration [15], [16].

Compared with hinge-like losses, a potential gain in using proper scoring rules is that classification regret bounds can be obtained with respect to a range of cost-weighted classification losses with different cost matrices C for a proper scoring rule, defined independently of C. The cost matrix is involved only to convert an action (in the form of a probability vector) from the scoring rule to a prediction for the cost-weighted classification loss. See Corollary 3. In contrast, for the regret bound (29), the hinge-like loss L^{cw3} depends on the cost-matrix C used in the classification loss L^{cw}. A similar observation is made by [21, Corollary 28] in two-class settings.

A general basis for deriving regret bounds, applicable to not just scoring rules but arbitrary losses $L(j,\gamma)$ with an action space $\mathcal{A} \subset \mathbb{R}^m$, can be cast as

$$\psi(B^{\mathrm{zo}}(\eta,\gamma)) \le B_L(\eta,\gamma), \tag{43}$$

where $B^{\mathrm{zo}}(\eta,\gamma) = B_{L^{\mathrm{zo}}}(\eta,\gamma)$, the regret of the zero-one loss $L^{\mathrm{zo}}(\eta,\gamma)$, and

$$\psi(t) = \inf_{\eta'\in\Delta_m,\gamma'\in\mathcal{A}:\, B^{\mathrm{zo}}(\eta',\gamma')=t} B_L(\eta',\gamma'), \quad t \ge 0.$$

In fact, (43) is a tautology from the definition of ψ. Various regret bounds can be obtained by identifying convenient lower bounds of ψ. In the two-class setting, the regret for the zero-one loss is $B^{\mathrm{zo}}(\eta,\gamma) = |2\eta_1 - 1| 1\{(2\eta_1-1)(\gamma_1-\gamma_2) \le 0\}$ for $\eta = (\eta_1,\eta_2)^{\mathrm{T}} \in \Delta_2$ and $\gamma = (\gamma_1,\gamma_2)^{\mathrm{T}}$. For $t > 0$, $B^{\mathrm{zo}}(\eta,\gamma) = t$ means $\eta_1 = (1 \pm t)/2$, and hence $\psi(t)$ can be simplified as

$$\psi^{\mathrm{BJM}}(t) = \min\left\{ \inf_{\gamma': t(\gamma_1'-\gamma_2') \le 0} B_L^1\left(\frac{1+t}{2}, \gamma'\right), \inf_{\gamma': t(\gamma_1'-\gamma_2') \ge 0} B_L^1\left(\frac{1-t}{2}, \gamma'\right) \right\},$$

where $B_L^1(\eta_1,\gamma)$ denotes $B_L(\eta,\gamma)$ as a function of (η_1,γ). Moreover, $\psi^{\mathrm{BJM}}(t)$ at $t=0$ also satisfies $\psi^{\mathrm{BJM}}(0) = 0 \le \psi(0)$. Therefore, (43) holds with ψ replaced by ψ^{BJM}:

$$\psi^{\mathrm{BJM}}(B^{\mathrm{zo}}(\eta,\gamma)) \le B_L(\eta,\gamma). \tag{44}$$

In the multi-class setting, the regret $B^{\mathrm{zo}}(\eta,\gamma)$ does not admit a direct simplification. Nevertheless, our results below for proper scoring rules can be seen as further manipulation of (43) by exploiting the fact that the regret (8) is a Bregman divergence due to the canonical representation (7) for proper scoring rules.

Remark 3: Replacing ψ^{BJM} in (44) by the greatest convex lower bound on ψ^{BJM} (or the Fenchel biconjugate of ψ^{BJM}) recovers the regret bound in [10, Theorem 1] in the symmetric case where $L(1,\gamma) = L(2,-\gamma)$. In general, there is a benefit from such a modification in the setting where covariates are restored, instead of being lifted out in most of our discussion. For a regret bound in the form $\phi(B^{\mathrm{zo}}(\eta,\gamma)) \le B_L(\eta,\gamma)$, if ϕ is convex, then application of Jensen's inequality gives

$$\phi\big[E\{B^{\mathrm{zo}}(\eta(X),\gamma(X))\}\big] \le E\big[\phi\{B^{\mathrm{zo}}(\eta(X),\gamma(X))\}\big] \le E\{B_L(\eta(X),\gamma(X))\},$$

where $E\{B^{\mathrm{zo}}(\eta(X),\gamma(X))$ and $E\{B_L(\eta(X),\gamma(X))\}$ are the average regret over X.

1) Zero-One Classification: Before presenting our general regret bounds for proper scoring rules with respect to cost-weighted classification in Sections V-B.2–V-B.3, we demonstrate novel implications of our general results in the simple but important setting of zero-one classification.

For a proper scoring rule $L(j,q)$, an application of our regret bound (56) or (60) with respect to the zero-one loss with $C_0 = 1_m$ shows that for any $\eta, q \in \Delta_m$,

$$\underline{\psi}\left(B^{\mathrm{zo}}(\eta,q)\right) \le B_L(\eta,q), \tag{45}$$

where $\underline{\psi}(\cdot)$ is defined as

$$\underline{\psi}(t) = \inf_{\substack{\eta',q'\in\Delta_m:\, \|\eta'-q'\|_{\infty 2}=t,\\ \max_{j\in[m]} q_j' \le 1/2}} B_L(\eta',q'), \quad t \ge 0. \tag{46}$$

For a vector $b = (b_1,\ldots,b_m)^{\mathrm{T}}$, $\|b\|_{\infty 2}$ denotes $\max_{j\ne k\in[m]} (|b_j|+|b_k|)$. Inequality (45) can be seen to extend the two-class regret bound (44) in Bartlett et al. (2006) to multi-class settings for proper scoring rules. Unlike the two-class setting, additional effort is needed to find a simple meaningful lower bound of $\underline{\psi}$ in the multi-class setting. Our current approach involves deriving a lower bound on the regret (or Bregman divergence) B_L by the L_1 norm, in the form such that for any $\eta, q \in \Delta_m$,

$$B_L(\eta,q) = H_L(q) - H_L(\eta) - (q-\eta)^{\mathrm{T}}\partial H_L(q) \ge \frac{\kappa_L}{2}\|\eta - q\|_1^2, \tag{47}$$

where $\kappa_L > 0$ is a constant depending on L, and $\|b\|_1 = \sum_{j=1}^m |b_j|$ is the L_1 norm for any vector $b = (b_1,\ldots,b_m)^{\mathrm{T}}$. Hence (47) can be interpreted as saying that $-H_L$ is strongly convex with respect to the L_1 norm with modulus κ_L. Because $\|\eta-q\|_1 \ge \|\eta-q\|_{\infty 2}$, the regret bound (45) together with (47) implies that for any $\eta, q \in \Delta_m$,

$$\frac{\kappa_L}{2}\left(B^{\mathrm{zo}}(\eta,q)\right)^2 \le B_L(\eta,q). \tag{48}$$

In general, the preceding discussion shows that a potentially improved lower bound on the Bregman divergence $B_L(\eta,q)$ by a non-quadratic function of $\|\eta-q\|_1$ can also be translated into a classification regret bound.

Our current approach does not exploit the restriction that $\max_{j\in[m]} q_j' \le 1/2$ in the definition of $\underline{\psi}$. Hence it is interesting to study how our results here can be improved. On the other hand, such an improvement, even if achieved, may be limited. See the later discussion on regret bounds for the pairwise exponential loss.

Our approach leads to the following result for two classes of proper scoring rules discussed in Section V-A: a class of

pairwise losses (39) with f_0 associated with a Beta family of weight functions as studied in [4], and a class of simultaneous losses (41). In all these cases, inequalities (47) can be of independent interest.

Proposition 10: Inequalities (47) and (48) hold for the following proper scoring rules.

(i) Consider a pairwise loss $L = L_{f_0}^{\mathrm{pw,s}}$ in (39), with a univariate function f_0 defined such that (37) holds with a weight function $w(q_1) = 2^{2\nu} q_1^{\nu-1} q_2^{\nu-1}$ for $(q_1, q_2)^{\mathrm{T}} \in \Delta_2$. If $\nu \le 0$, then (47) and (48) are valid with $\kappa_L = 2$. In general, the constant κ_L cannot be improved to be greater for $m = 2$ or for $m \ge 3$ and $\nu \in (-1, 0]$.

(ii) Consider a simultaneous loss $L = L_\beta$ in (41). Then (47) and (48) are valid with

$$\kappa_L = \begin{cases} (1-\beta) m^{(1-1/\beta)(2\beta-1)} 2^{2-2\beta}, & \text{if } \beta \in [1/2, 1), \\ (1-\beta) 2^{1/\beta - 1}, & \text{if } \beta \in (0, 1/2]. \end{cases}$$

The bounds from the two segments both give $\kappa_L = 1$ at $\beta = 1/2$. In general, the constant κ_L cannot be improved to be greater for $m = 2$ or for $m \ge 3$ and $\beta \in (0, 1/2]$.

We discuss several specific examples. The standard likelihood loss $L(j, q) = -\log q_j$ is equivalent to the simultaneous loss L_β in the limit of $\beta \to 1$ after properly rescaled. In this case, Pinsker's inequality states that (47) holds with $\kappa_L = 1$ [45, Lemma 12.6.1]:

$$\sum_{j=1}^m \eta_j \log(\eta_j / q_j) \ge \frac{1}{2} \left(\sum_{j=1}^m |\eta_j - q_j| \right)^2. \tag{49}$$

The resulting regret bound (48) for the standard likelihood loss L then gives

$$\frac{1}{2} \left(B^{\mathrm{zo}}(\eta, q) \right)^2 \le B_L(\eta, q). \tag{50}$$

This surrogate regret bound for the multinomial likelihood loss appears new, even though the Bregman divergence bound (49) is known. In the Supplement, we verify that (49) can be recovered from (47), using Proposition 10(ii) as $\beta \to 1$. The constant κ_L in Proposition 10(ii) may be improvable for fixed $\beta \in (1/2, 1)$, but is not improvable in the limit $\beta \to 1$, i.e., the constant $1/2$ in (49) is not improvable. On the other hand, an improved lower bound of the KL divergence than (49) can be found in terms of a non-quadratic function of $\|\eta - q\|_1$ (e.g., [21]). Such bounds can also be translated into classification regret bounds for the likelihood loss by the discussion from (47) to (48).

The pairwise exponential loss associated with multi-class boosting is defined equivalently as $L_e^{\mathrm{pw,s}}(j, q) = 2(L_{1/2} - 1) = 2 \sum_{k \in [m], k \ne j} \sqrt{q_k / q_j}$ in Section II-B. The two inequalities (47) obtained from Proposition 10, part (i) with $\nu = -1/2$ and part (ii) with $\beta = 1/2$, are equivalent to each other and both lead to

$$H_{L_{1/2}}(q) - H_{L_{1/2}}(\eta) - (q - \eta)^{\mathrm{T}} \partial H_{L_{1/2}}(q) \ge \frac{1}{2} \|\eta - q\|_1^2, \tag{51}$$

where $H_{L_{1/2}}(q) = \|q\|_{1/2}$ and $L_{1/2}(j, q) = (\|q\|_{1/2} / q_j)^{1/2} = \sum_{k=1}^m \sqrt{q_k / q_j}$. The resulting regret bound (48) for the rescaled pairwise exponential loss $L_{1/2}$ gives

$$\frac{1}{2} \left(B^{\mathrm{zo}}(\eta, q) \right)^2 \le B_{L_{1/2}}(\eta, q). \tag{52}$$

The two bounds (50) and (52) for the likelihood and rescaled pairwise exponential losses happen to be of the same form, due to the scaling used. For the two-class exponential loss defined as $L_e = L_{1/2} - 1$, the existing regret bound (44), corresponding to an exact calculation of $\underline{\psi}$ by the proof of (61) later, is

$$1 - \sqrt{1 - (B^{\mathrm{zo}}(\eta, q))^2} \le B_{L_{1/2}}(\eta, q),$$

which is slightly stronger than (52) because $1 - \sqrt{1 - \delta^2} \ge \delta^2/2$ for $\delta \in [0, 1]$, but $(1 - \sqrt{1 - \delta^2})/(\delta^2/2) \to 1$ as $\delta \to 0$. Therefore, our result (52) provides a reasonable extension of existing regret bounds to multi-class pairwise exponential losses.

A notable proper scoring rule which is not informed by Proposition 10 for $m \ge 3$ is the simultaneous exponential loss L_0^{r} as used in [6], even though the loss L_0^{r} is equivalent to the exponential loss for $m = 2$. See the Supplement for details.

Remark 4: Inequality (47) on the Bregman divergence in general differs from generalized Pinsker inequalities relating (two-distribution) f-divergences to the total variation studied in [21], Section 7.2, for binary experiments. For the pairwise exponential loss, the Bregman divergence on the left-hand side of (51) can be calculated as $(\sum_{j \in [m]} \sqrt{q_j})(\sum_{j \in [m]} \eta_j / \sqrt{q_j}) - (\sum_{j \in [m]} \sqrt{\eta_j})^2$, which is apparently not any f-divergence between probability vectors η and q. An exception is the classical Pinsker inequality (49): the Kullback–Liebler divergence on the left-hand side of (49) is both an f-divergence with $f(t) = t \log t$ and a Bregman divergence with $H_L(\eta) = -\sum_{j \in [m]} \eta_j \log \eta_j$.

Remark 5: In the two-class setting, a scoring rule satisfying inequality (47) is called a strongly proper loss, and surrogate regret bounds are obtained for strongly proper losses with respect to the area under the curve (AUC) in [46]. It is interesting to investigate possible extensions of such results to the multi-class setting.

2) Cost-Transformed Losses: We study two types of classification regret bounds with respect to a general cost-weighted classification loss as defined in Section II-C. This subsection deals with the first type where a classification regret bound is derived for a loss, allowed to depend on a pre-specified cost matrix C, similarly as the hinge-like loss L^{cw3} in (23). An action of the loss is directly taken as a prediction for the cost-weighted classification loss. See the next subsection on the second type of classification regret bounds.

For a general loss $L(j, \gamma)$ (not just scoring rules), define a cost-transformed loss, depending on a cost matrix C, as

$$\tilde{L}(j, \gamma) = c_{jM} L(j, \gamma) + \sum_{k \in [m], k \ne j} (c_{jM} - c_{jk}) \{ L(k, \gamma) - 1 \}, \tag{53}$$

where $c_{jM} = \max_{k \in [m]} c_{jk}$. In the special case where $C = 1_m 1_m^{\mathrm{T}} - I_m$ for the zero-one loss, the transformed loss $\tilde{L}(j, \gamma)$ reduces to the original loss $L(j, \gamma)$. A motivation for this construction is that the cost-weighted classification loss

can also be obtained in this way from the zero-one loss: $L^{\mathrm{cw}}(j,\gamma) = \tilde{L}^{\mathrm{zo}}(j,\gamma)$. In general, the risk and regret of the transformed loss can be related to those of the original loss as follows.

Lemma 4: The risks of the losses $\tilde{L}(j,\gamma)$ and $L(j,\gamma)$ satisfy

$$R_{\tilde{L}}(\eta,\gamma) = (1_m^{\mathrm{T}}\tilde{\eta})\, R_L(\tilde{\tilde{\eta}},\gamma) - D(\eta),$$

where $D(\eta) = \sum_{j\in[m]}\sum_{k\in[m],k\neq j}\eta_j(c_{jM} - c_{jk})$, $\tilde{\tilde{\eta}} = \tilde{\eta}/(1_m^{\mathrm{T}}\tilde{\eta}) \in \Delta_m$, and $\tilde{\eta} = (\tilde{\eta}_1,\ldots,\tilde{\eta}_m)^{\mathrm{T}} \in \mathbb{R}_+^m$ with

$$\tilde{\eta}_j = c_{jM}\eta_j + \sum_{k\in[m],k\neq j}(c_{kM} - c_{kj})\eta_k.$$

Moreover, the regrets of $\tilde{L}$ and L satisfy $B_{\tilde{L}}(\eta,\gamma) = (1_m^{\mathrm{T}}\tilde{\eta})\, B_L(\tilde{\tilde{\eta}},\gamma)$.

For a scoring rule $L(j,q)$ with actions defined as probability vectors $q \in \Delta_m$, there is a simple upper bound on the regret of the associated zero-one loss $L^{\mathrm{zo}}(j,q)$, which is instrumental to our derivation of classification regret bounds.

Lemma 5: For any $\eta, q \in \Delta_m$, it holds that

$$B^{\mathrm{zo}}(\eta,q) \le \|\eta - q\|_{\infty 2},$$

where $\|b\|_{\infty 2} = \max_{j\neq k\in[m]}(|b_j| + |b_k|)$ for any vector $b = (b_1,\ldots,b_m)^{\mathrm{T}}$. The bound is tight for any $m \ge 2$ in that there exist $\eta, q \in \Delta_m$ for which the bound becomes exact.

Combining the preceding two lemmas and invoking a similar argument as indicated by (43) leads to the following regret bound, depending on the action q.

Proposition 11: For a scoring rule $L(j,q)$, define a nondecreasing function ψ_q:

$$\psi_q(t) = \inf_{\eta'\in\Delta_m:\|\eta'-q\|_{\infty 2}\ge t} B_L(\eta',q), \quad t \ge 0.$$

Then the regrets of the cost-weighted classification loss $L^{\mathrm{cw}}(j,q)$ and the cost-transformed scoring rule $\tilde{L}(j,q)$ satisfy

$$\psi_q\left(\frac{B^{\mathrm{cw}}(\eta,q)}{1_m^{\mathrm{T}}\tilde{\eta}}\right) \le \frac{B_{\tilde{L}}(\eta,q)}{1_m^{\mathrm{T}}\tilde{\eta}}, \tag{54}$$

where $B^{\mathrm{cw}} = B_{L^{\mathrm{cw}}}$, and $\tilde{\eta}$ is defined, depending on η and C, as in Lemma 4.

A cost-transformed loss (53) from a proper scoring rule can be easily shown to remain a proper scoring rule. In this case, a uniform regret bound can be obtained from (54), by taking an infimum over q and incorporating simplification due to the representation of the regret (8) as a Bregman divergence for a proper scoring rule.

Corollary 2: For a proper scoring rule $L(j,q)$, the regrets of $L^{\mathrm{cw}}(j,q)$ and $\tilde{L}(j,q)$ satisfy

$$\underline{\psi}\left(\frac{B^{\mathrm{cw}}(\eta,q)}{1_m^{\mathrm{T}}\tilde{\eta}}\right) \le \frac{B_{\tilde{L}}(\eta,q)}{1_m^{\mathrm{T}}\tilde{\eta}}, \tag{55}$$

where $\underline{\psi}$ is defined in (46), and $\tilde{\eta}$ is defined, depending on η and C, in Lemma 4.

It is instructive to examine the regret bound (55) in the special case of class-weighted costs, where $C = C_0 1_m^{\mathrm{T}} - \mathrm{diag}(C_0)$ with $C_0 = (c_{10},\ldots,c_{m0})^{\mathrm{T}}$. The cost-transformed loss $\tilde{L}$ reduces to $\tilde{L}(j,q) = c_{j0}L(j,q)$. The regret bound (55) becomes

$$\underline{\psi}\left(\frac{B^{\mathrm{cw0}}(\eta,q)}{C_0^{\mathrm{T}}\eta}\right) \le \frac{B_{\tilde{L}}(\eta,q)}{C_0^{\mathrm{T}}\eta}, \tag{56}$$

where $B^{\mathrm{cw0}} = B_{L^{\mathrm{cw0}}}$. We defer a discussion of these results until after Corollary 3.

3) Cost-Independent Losses: We derive a different type of classification regret bounds than in the preceding subsection. Here a loss used for training is defined independently of any cost matrix, but an action of the loss can be converted after training to a prediction, depending on the cost matrix C, for the cost-weighted classification loss. For scoring rules, our derivation relies on the following extension of Lemma 5 on the regret of the cost-weighted classification loss, where a prediction is linearly converted from a probability vector.

Lemma 6: For any $\eta, q \in \Delta_m$, it holds that

$$B^{\mathrm{cw}}(\eta, \overline{C}^{\mathrm{T}}q) \le \|\overline{C}^{\mathrm{T}}(\eta - q)\|_{\infty 2},$$

where $\overline{C} = C_M 1_m^{\mathrm{T}} - C$ and $C_M = (c_{1M},\ldots,c_{mM})^{\mathrm{T}}$ with $c_{jM} = \max_{k\in[m]} c_{jk}$ for $j \in [m]$ as defined in the transformed loss (53).

By a similar argument as indicated by (43), we obtain a regret bound which compares the regret of a scoring rule $L(j,q)$ with that of the cost-weighted classification loss with a prediction depending on both q and C as in Lemma 6.

Proposition 12: For a scoring rule $L(j,q)$, define a nondecreasing function ψ_q^C:

$$\psi_q^C(t) = \inf_{\eta'\in\Delta_m:\|\overline{C}^{\mathrm{T}}(\eta'-q)\|_{\infty 2}\ge t} B_L(\eta',q), \quad t \ge 0.$$

Then the regrets of the cost-weighted classification loss $L^{\mathrm{cw}}(j,\overline{C}^{\mathrm{T}}q)$ and the scoring rule $L(j,q)$ satisfy

$$\psi_q^C\left(B^{\mathrm{cw}}(\eta,\overline{C}^{\mathrm{T}}q)\right) \le B_L(\eta,q). \tag{57}$$

For a *proper* scoring rule $L(j,q)$, the regret bound (57) can be strengthened (see the Supplement for a proof) such that for each $w \in \mathcal{W}_{\eta,q}$,

$$\psi_{q^w}^C\left(B^{\mathrm{cw}}(\eta,\overline{C}^{\mathrm{T}}q)\right) \le B_L(\eta,q), \tag{58}$$

where $q^w = (1-w)\eta + wq$ and

$$\mathcal{W}_{\eta,q} = \left\{ w \in [0,1] : \overline{C}_k^{\mathrm{T}}q^w = \max_j(\overline{C}_j^{\mathrm{T}}q^w) \right.$$
$$\left. \text{for } k = \mathrm{argmax}_j(\overline{C}_j^{\mathrm{T}}q) \right\} \ni 1.$$

By definition, $w \in \mathcal{W}_{\eta,q}$ means that using q^w yields the same classification as using q. Moreover, a uniform regret bound can be obtained from (58) by minimizing over q^w with $w \in \mathcal{W}_{\eta,q}$ such that $\max_{j\in[m]}\overline{C}_j^{\mathrm{T}}q^w \le 1_m^{\mathrm{T}}\overline{C}^{\mathrm{T}}q^w/2$.

Corollary 3: For a proper scoring rule $L(j,q)$, define

$$\underline{\psi}^C(t) = \inf_{\substack{\eta',q'\in\Delta_m:\|\overline{C}^{\mathrm{T}}(\eta'-q')\|_{\infty 2}=t,\\ \max_{j\in[m]}(\overline{C}_j^{\mathrm{T}}q')\le 1_m^{\mathrm{T}}\overline{C}^{\mathrm{T}}q'/2}} B_L(\eta',q'), \quad t \ge 0.$$

Then the regrets of $L^{\mathrm{cw}}(j,\overline{C}^{\mathrm{T}}q)$ and $L(j,q)$ satisfy

$$\underline{\psi}^C\left(B^{\mathrm{cw}}(\eta,\overline{C}^{\mathrm{T}}q)\right) \le B_L(\eta,q). \tag{59}$$

In the special case of class-weighted costs, corresponding to $C = C_0 1_m^{\mathrm{T}} - \mathrm{diag}(C_0)$ with $C_0 = (c_{10},\ldots,c_{m0})^{\mathrm{T}}$, define

$$\underline{\psi}^{C_0}(t) = \inf_{\substack{\eta',q'\in\Delta_m:\|C_0\circ(\eta'-q')\|_{\infty 2}=t,\\ \max_{j\in[m]}(c_{j0}q'_j)\le C_0^{\mathrm{T}}q'/2}} B_L(\eta',q'), \quad t \ge 0,$$

where $\circ$ denotes the component-wise product between two vectors. The regret bound (59) for proper scoring rules reduces to

$$\underline{\psi}^{C_0}\left(B^{\mathrm{cw0}}(\eta, C_0 \circ q)\right) \le B_L(\eta, q). \tag{60}$$

It is interesting to compare the two regret bounds (56) and (60). On one hand, for the zero-one loss with $C_0 = 1_m$, both of these bounds lead to the regret bound (45) discussed in Section V-B.1. On the other hand, the two bounds (56) and (60) in general serve different purposes. The bound (56) compares the regrets of the transformed scoring rule $\tilde{L}$ depending on C_0 and the classification loss L^{cw0} with the prediction always set to q. To use $\tilde{L}$, a different round of training is required for a different choice of C_0. The bound (60) relates the regrets of the original scoring rule L, independent of C_0, and the classification loss L^{cw0} with the prediction defined as $C_0 \circ q$. Only one round of training is needed to determine q when using L, and then the prediction can be adjusted from q according to the choice of C_0. Hence the bound (60) can be potentially more useful than (56).

For binary classification with $m = 2$, the regret bound (60) for proper scoring rules can be shown to recover Theorem 25 in [21]. For a *proper* scoring rule $L(j, q)$ and any $\eta, q \in \Delta_2$, it holds that

$$\min\left\{\psi^{\mathrm{RW}}(\delta), \psi^{\mathrm{RW}}(-\delta)\right\} \le B_L(\eta, q), \tag{61}$$

where $\delta = B^{\mathrm{cw0}}(\eta, C_0 \circ q)$, $\psi^{\mathrm{RW}}(\delta) = B_L^1((c_{20} + \delta)/(c_{10} + c_{20}), c_{20}/(c_{10} + c_{20}))$, and $B_L^1(\eta_1, q_1) = B_L(\eta, q)$ with $\eta = (\eta_1, \eta_2)^{\mathrm{T}}$ and $q = (q_1, q_2)^{\mathrm{T}}$, that is, $B_L^1(\eta_1, q_1)$ is $B_L(\eta, q)$ treated as a function of (η_1, q_1) only. See the Supplement for details.

VI. Conclusion

In this article, we are mainly concerned with constructing losses and establishing corresponding regret bounds in multi-class settings. Various topics remain to be studied in further research. Large sample theory can be studied regarding estimation and approximation errors, similarly as in [9] and [10], by taking advantage of our multi-class regret bounds. It is also of interest to incorporate estimation of a data quantizer [12], [17]. Computational algorithms need to be developed for implementing our new hinge-like losses and, in connection with boosting algorithms, for implementing composite losses based on new proper scoring rules. Numerical experiments are also desired to evaluate empirical performance of new methods.

Acknowledgment

The authors thank the Associate Editor and two referees for constructive comments leading to various improvements in the article.

References

[1] M. H. DeGroot, "Uncertainty, information, and sequential experiments," *Ann. Math. Stat.*, vol. 33, no. 2, pp. 404–419, 1962.

[2] P. D. Grünwald and A. P. Dawid, "Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory," *Ann. Statist.*, vol. 32, no. 4, pp. 1367–1433, Aug. 2004.

[3] L. J. Savage, "Elicitation of personal probabilities and expectations," *J. Amer. Stat. Assoc.*, vol. 66, no. 336, pp. 783–801, Dec. 1971.

[4] A. Buja, W. Stuetzle, and Y. Shen, "Loss functions for binary class probability estimation and classification: Structure and applications," Univ. Pennsylvania, Philadelphia, PA, USA, Tech. Rep., 2005.

[5] R. C. Williamson, E. Vernet, and M. D. Reid, "Composite multiclass losses," *J. Mach. Learn. Res.*, vol. 17, no. 222, pp. 7860–7911, 2016.

[6] H. Zou, J. Zhu, and T. Hastie, "New multicategory boosting algorithms based on multicategory Fisher-consistent losses," *Ann. Appl. Statist.*, vol. 2, no. 4, pp. 1290–1306, 2008.

[7] I. Mukherjee and R. E. Schapire, "A theory of multiclass boosting," *J. Mach. Learn. Res.*, vol. 14, pp. 437–497, Feb. 2013.

[8] Y. Lin, "Support vector machines and the Bayes rule in classification," *Data Mining Knowl. Discovery*, vol. 6, pp. 259–275, Jul. 2002.

[9] T. Zhang, "Statistical behavior and consistency of classification methods based on convex risk minimization," *Ann. Statist.*, vol. 32, no. 1, pp. 56–85, 2004.

[10] P. L. Bartlett, M. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *J. Amer. Statist. Assoc.*, vol. 101, pp. 138–156, Apr. 2006.

[11] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data," *J. Amer. Statist. Assoc.*, vol. 99, no. 465, pp. 67–81, 2004.

[12] J. Duchi, K. Khosravi, and F. Ruan, "Multiclass classification, information, divergence and surrogate risk," *Ann. Statist.*, vol. 46, no. 6B, pp. 3246–3275, Dec. 2018.

[13] J. Weston and C. Watkins, "Multi-class support vector machines," Dept. Comput. Sci., Roy. Holloway College, Univ. London, London, U.K., Tech. Rep. CSD-TR-98-04, 1998.

[14] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *J. Mach. Learn. Res.*, vol. 2, pp. 265–292, Mar. 2001.

[15] T. Zhang, "Statistical analysis of some multi-category large margin classification methods," *J. Mach. Learn. Res.*, vol. 5, pp. 1225–1251, Oct. 2004.

[16] A. Tewari and P. L. Bartlett, "On the consistency of multiclass classification methods," *J. Mach. Learn. Res.*, vol. 8, pp. 1007–1025, May 2007.

[17] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "On surrogate loss functions and f-divergences," *Ann. Statist.*, vol. 37, pp. 876–904, Apr. 2009.

[18] L. Györfi and T. Nemetz, "F-dissimilarity: A generalization of the affinity of several distributions," *Ann. Inst. Stat. Math.*, vol. 30, no. 1, pp. 105–113, Dec. 1978.

[19] D. García-García and R. C. Williamson, "Divergences and risks for multiclass experiments," in *Proc. 25th Annu. Conf. Learn. Theory*, 2012, pp. 28.1–28.20.

[20] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Amer. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, Jan. 2012.

[21] M. D. Reid and R. C. Williamson, "Information, divergence and risk for binary experiments," *J. Mach. Learn. Res.*, vol. 12, no. 3, pp. 731–817, Mar. 2011.

[22] I. Steinwart, "How to compare different loss functions and their risks," *Constructive Approx.*, vol. 26, no. 2, pp. 225–287, 2007.

[23] C. Scott, "Calibrated asymmetric surrogate losses," *Electron. J. Statist.*, vol. 6, pp. 958–992, May 2012.

[24] I. Good, "Rational decisions," *J. Roy. Stat. Soc. B, Methodol.*, vol. 14, pp. 107–114, Jan. 1952.

[25] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Roy. Statist. Soc. B, Methodol.*, vol. 28, no. 1, pp. 131–142, 1966.

[26] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 299–318, 1967.

[27] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[28] R. I. Boţ and G. Wanka, "The conjugate of the pointwise maximum of two convex functions revisited," *J. Global Optim.*, vol. 41, no. 4, pp. 625–632, Aug. 2008.

[29] Y. Lei, U. Dogan, D.-X. Zhou, and M. Kloft, "Data-dependent generalization bounds for multi-class classification," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 2995–3021, May 2019.

[30] S. Ben-David, D. Loker, N. Srebro, and K. Sridharan, "Minimizing the misclassification error rate using a surrogate convex loss," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 83–90.

[31] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Statist. Softw.*, vol. 33, no. 1, pp. 1–22, 2010.
[32] A. Fu, B. Narasimhan, and S. Boyd, "CVXR: An R package for disciplined convex optimization," *J. Stat. Softw.*, vol. 94, no. 14, pp. 1–34, 2020.
[33] MOSEK Aps. (2020). *MOSEK Rmosek Package. Version 9.1*. [Online]. Available: https://docs.mosek.com/9.1/rmosek.pdf
[34] M. J. Schervish, "A general method for comparing probability assessors," *Ann. Statist.*, vol. 17, no. 4, pp. 1856–1879, Dec. 1989.
[35] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *Ann. Statist.*, vol. 28, no. 2, pp. 337–407, 2000.
[36] R. E. Schapire and Y. Freund, *Boosting: Foundations and Algorithms*. Cambridge, MA, USA: MIT Press, 2012.
[37] Z. Tan, "Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data," *Biometrika*, vol. 107, no. 1, pp. 137–158, Mar. 2020.
[38] Z. Tan, Y. Song, and Z. Ou, "Calibrated adversarial algorithms for generative modelling," *Stat*, vol. 8, no. 1, p. e224, Jan. 2019.
[39] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[40] S. Nowozin, B. Cseke, and R. Tomioka, "f-GAN: Training generative neural samplers using variational divergence minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 271–279.
[41] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.
[42] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Mach. Learn.*, vol. 37, no. 3, pp. 297–336, 1999.
[43] W. Xu and Z. Tan, "High-dimensional model-assisted inference for treatment effects with multi-valued treatments," 2022, *arXiv:2201.09192*.
[44] I. Good, "Comment on 'measuring information and uncertainty,"' in *Foundations of Statistical Inference*, V. P. Godambe and D. A. Sprott, and R. Holt, Eds. Toronto, ON, Canada: Holt, Rinehart and Winston, 1971, pp. 337–339.
[45] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1991.
[46] S. Agarwal, "Surrogate regret bounds for the area under the ROC curve via strongly proper losses," in *Proc. 26th Annu. Conf. Learn. Theory*, 2013, pp. 338–353.

Zhiqiang Tan received the B.S. degree in applied mathematics from Tsinghua University and the Ph.D. degree in statistics from the University of Chicago. He is currently a Professor with the Department of Statistics, Rutgers University. Prior to that, he was an Assistant Professor with Johns Hopkins University. He has published extensively on statistical theory, methods, and applications in leading statistical and interdisciplinary journals and conferences. His research interests include Monte Carlo methods, causal inference, statistical learning, and related areas. He received an NSF CAREER Award, and is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics and an Elected Member of the International Statistical Institute.

Xinwei Zhang received the B.S. degree in mathematics and applied mathematics and the B.E. degree in computer science from the Renmin University of China in 2015 and the A.M. degree in statistics from Washington University in St. Louis in 2017. He is currently pursuing the Ph.D. degree in statistics with Rutgers University. His research interests include statistics and machine learning.