Artificial Intelligence 215 (2014) 55–78

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

# Multicategory large margin classification methods: Hinge losses vs. coherence functions

Zhihua Zhang [a,*], Cheng Chen [a], Guang Dai [a], Wu-Jun Li [b], Dit-Yan Yeung [c]

[a] Key Laboratory of Shanghai Education Commission for Intelligence Interaction & Cognition Engineering, Department of Computer Science and Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai, 200240, China
[b] National Key Laboratory for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Nanjing, 210023, China
[c] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China

ARTICLE INFO

*Article history:*
Received 3 August 2013
Received in revised form 9 May 2014
Accepted 16 June 2014
Available online 20 June 2014

*Keywords:*
Multiclass margin classification
Fisher consistency
Multicategory hinge losses
Coherence losses
Multicategory boosting algorithm

ABSTRACT

Generalization of large margin classification methods from the binary classification setting to the more general multicategory setting is often found to be non-trivial. In this paper, we study large margin classification methods that can be seamlessly applied to both settings, with the binary setting simply as a special case. In particular, we explore the Fisher consistency properties of multicategory majorization losses and present a construction framework of majorization losses of the 0–1 loss. Under this framework, we conduct an in-depth analysis about three widely used multicategory hinge losses. Corresponding to the three hinge losses, we propose three multicategory majorization losses based on a *coherence function*. The limits of the three coherence losses as the temperature approaches zero are the corresponding hinge losses, and the limits of the minimizers of their expected errors are the minimizers of the expected errors of the corresponding hinge losses. Finally, we develop multicategory large margin classification methods by using a so-called multiclass $\mathcal{C}$-loss.

## 1. Introduction

Large margin classification methods have become increasingly popular since the advent of the support vector machine (SVM) [4] and boosting [7,8]. Recent developments include the large-margin unified machine of Liu et al. [16] and the flexible assortment machine of Qiao and Zhang [19]. Typically, large margin classification methods approximately solve an otherwise intractable optimization problem defined with the 0–1 loss. These algorithms were originally designed for binary classification problems. Unfortunately, generalization of them to the multicategory setting is often found to be non-trivial. The goal of this paper is to solve multicategory classification problems using the same margin principle as that for binary problems.

The conventional SVM based on the hinge loss function possesses support vector interpretation (or data sparsity) but does not have uncertainty (that is, the SVM does not directly estimate the conditional class probability). The non-differentiable hinge loss function also makes it non-trivial to extend the conventional SVM from binary classification

[*] Corresponding author.
*E-mail addresses:* zhzhang@gmail.com (Z. Zhang), jackchen1990@gmail.com (C. Chen), guang.gdai@gmail.com (G. Dai), liwujun@nju.edu.cn (W.-J. Li), dyyeung@cse.ust.hk (D.-Y. Yeung).

http://dx.doi.org/10.1016/j.artint.2014.06.002
0004-3702/© 2014 Elsevier B.V. All rights reserved.

problems to multiclass classification problems in the same margin principle [25,3,26,12,5,13]. Thus, one seemingly natural approach to constructing a classifier for the binary and multiclass problems is to consider a smooth loss function.

For example, regularized logistic regression models based on the negative multinomial log-likelihood function (also called the logit loss) [31,10] are competitive with SVMs. Moreover, it is natural to exploit the logit loss in the development of a multicategory boosting algorithm [9]. Recently, Zhang et al. [30] proposed a smooth loss function that called coherence function for developing binary large margin classification methods. The coherence function establishes a bridge between the hinge loss and the logit loss. In this paper, we study the application of the coherence function in the multiclass classification problem.

### 1.1. Multicategory margin classification methods

We are concerned with an $m$-class ($m > 2$) classification problem with a set of training data points $\{(\mathbf{x}_i, c_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^p$ is an input vector and $c_i \in \{1, 2, \dots, m\}$ is its corresponding class label. We assume that each $\mathbf{x}$ belongs to one and only one class. Our goal is to find a classifier $\phi(\mathbf{x}) : \mathbf{x} \to c \in \{1, \dots, m\}$.

Let $P_c(\mathbf{x}) = \Pr(C = c|X = \mathbf{x})$, $c = 1, \dots, m$, be the class conditional probabilities given $\mathbf{x}$. The expected error at $\mathbf{x}$ is then defined by

$$\sum_{c=1}^{m} \mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}} P_c(\mathbf{x}),$$

where $\mathbb{I}_{\{\#\}}$ is 1 if # is true and 0 otherwise. The empirical error on the training data is thus given by

$$\epsilon = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{\phi(\mathbf{x}_i) \neq c_i\}}.$$

Given that $\epsilon$ is equal to its minimum value zero when all training data points are correctly classified, we wish to use $\epsilon$ as a basis for devising classification methods.

Suppose the classifier is modeled using an $m$-vector $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \dots, g_m(\mathbf{x}))^T$, where the induced classifier is obtained via maximization in a manner akin to discriminant analysis: $\phi(\mathbf{x}) = \operatorname{argmax}_j\{g_j(\mathbf{x})\}$. For simplicity of our analysis, we assume that for a fixed $\mathbf{x}$, each $g_j$ itself lies in a compact set. We also assume that the maximizing argument of $\max_j g_j(\mathbf{x})$ is unique. Of course this excludes the trivial case that $g_j = 0$ for all $j \in \{1, \dots, m\}$. However, this assumption does not imply that the maximum value is unique; indeed, adding a constant to each component $g_j(\mathbf{x})$ does not change the maximizing argument. To remove this redundancy, it is convenient to impose a sum-to-zero constraint. Thus we define

$$\mathcal{G} = \left\{ \left(g_1(\mathbf{x}), \dots, g_m(\mathbf{x})\right)^T \,\middle|\, \sum_{j=1}^{m} g_j(\mathbf{x}) = 0 \right\}$$

and assume $\mathbf{g} \in \mathcal{G}$ in this paper unless otherwise specified. Zou et al. [33] referred to such a $\mathbf{g}$ as the *margin vector*. Liu and Shen [15] referred to $\max_j(g_j(\mathbf{x}) - g_c(\mathbf{x}))$ as the generalized margin of $(\mathbf{x}, c)$ with respect to (w.r.t.) $\mathbf{g}$.

Since a margin vector $\mathbf{g}$ induces a classifier, we explore the minimization of $\epsilon$ w.r.t. $\mathbf{g}$. However, this minimization problem is intractable because $\mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}}$ is the 0–1 function. A wide variety of margin-based classifiers can be understood as minimizers of a *surrogate loss* function $\psi_c(\mathbf{g}(\mathbf{x}))$, which upper bounds the 0–1 loss $\mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}}$. That is, various tractable *surrogate loss* functions $\psi_c(\mathbf{g}(\mathbf{x}))$ are thus used to upper approximate $\mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}}$. The corresponding empirical risk function is given by

$$\hat{\mathcal{R}}(\mathbf{g}) = \frac{1}{n} \sum_{i=1}^{n} \psi_{c_i}\big(\mathbf{g}(\mathbf{x}_i)\big).$$

If $\alpha$ is a positive constant that does not depend on $(\mathbf{x}, c)$, $\operatorname{argmin}_{\mathbf{g}(\mathbf{x}) \in \mathcal{G}} \frac{1}{\alpha} \hat{\mathcal{R}}(\mathbf{g})$ is equivalent to $\operatorname{argmin}_{\mathbf{g}(\mathbf{x}) \in \mathcal{G}} \hat{\mathcal{R}}(\mathbf{g})$. We thus present the following definition.

**Definition 1.** A surrogate loss $\psi_c(\mathbf{g}(\mathbf{x}))$ is said to be the majorization of $\mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}}$ w.r.t. $(\mathbf{x}, c)$ if $\psi_c(\mathbf{g}(\mathbf{x})) \geq \alpha \mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}}$ where $\alpha$ is a positive constant that does not depend on $(\mathbf{x}, c)$.

In practice, convex majorization functions play an important role in the development of classification algorithms. On one hand, the convexity makes the resulting optimization problems computationally tractable. On the other hand, the classification methods usually have better statistical properties.

Given a majorization function $\psi_c(\mathbf{g}(\mathbf{x}))$, the classifier resulted from the minimization of $\hat{\mathcal{R}}(\mathbf{g})$ w.r.t. the margin vector $\mathbf{g}$ is called a large margin classifier or a margin-based classification method. In the binary classification setting, a wide variety of classifiers can be understood as minimizers of a majorization loss function of the 0–1 loss. If such functions satisfy other technical conditions, the resulting classifiers can be shown to be Bayes consistent [1]. It seems reasonable to pursue a similar development in the case of multicategory classification, and indeed such a proposal has been made by Zou et al. [33] (also see [24,23]).

**Definition 2.** A surrogate function $\psi_c(\mathbf{g}(\mathbf{x}))$ is said to be *Fisher-consistent w.r.t.* a margin vector $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_m(\mathbf{x}))^T$ at $(\mathbf{x}, c)$ if (i) the following risk minimization problem

$$\hat{\mathbf{g}}(\mathbf{x}) = \underset{\mathbf{g}(\mathbf{x}) \in \mathcal{G}}{\operatorname{argmin}} \sum_{c=1}^{m} \psi_c\big(\mathbf{g}(\mathbf{x})\big) P_c(\mathbf{x}) \tag{1}$$

has a unique solution $\hat{\mathbf{g}}(\mathbf{x}) = (\hat{g}_1(\mathbf{x}), \ldots, \hat{g}_m(\mathbf{x}))^T$; and (ii)

$$\underset{c}{\operatorname{argmax}}\, \hat{g}_c(\mathbf{x}) = \underset{c}{\operatorname{argmax}}\, P_c(\mathbf{x}).$$

Zou et al. [33] assumed $\psi_c(\mathbf{g}(\mathbf{x}))$ as an independent and identical setting; that is, $\psi_c(\mathbf{g}(\mathbf{x})) \triangleq \eta(g_c(\mathbf{x}))$ where $\eta$ is some loss function. As we see, Definition 2 does not require that the function $\psi_c(\mathbf{g}(\mathbf{x}))$ depends only on $g_c(\mathbf{x})$. Thus, this definition refines the definition of Zou et al. [33]. The definition is related to the notion of *infinite-sample consistency* (ISC) of Zhang [28]. ISC says that an exact solution of Problem (1) leads to a Bayes rule. However, it does not require that the solution of Problem (1) be unique. Additionally, Zhang [28] especially discussed two other settings: pairwise comparison $\psi_c(\mathbf{g}(\mathbf{x})) \triangleq \sum_{j \neq c} \eta(g_c(\mathbf{x}) - g_j(\mathbf{x}))$ and constrained comparison $\psi_c(\mathbf{g}(\mathbf{x})) \triangleq \sum_{j \neq c} \eta(-g_j(\mathbf{x}))$.

In this paper, we are concerned with multicategory classification methods in which binary and multicategory problems are solved following the same principle. One of the principled approaches is due to Lee et al. [13]. The authors proposed a multicategory SVM (MSVM) which treats the $m$-class problem simultaneously. Moreover, Lee et al. [13] proved that their MSVM satisfies a Fisher consistency condition. Unfortunately, this desirable property does not hold for many other multiclass SVMs (see, e.g., [25,3,26,12]). The multiclass SVM of [5] possesses this property only if there is a dominating class (that is, $\max_j P_j(\mathbf{x}) > 1/2$).

Recently, Liu and Shen [15] proposed a so-called multicategory $\psi$-learning algorithm by using a multicategory $\psi$ loss, and Wu and Liu [27] devised robust truncated-hinge-loss SVMs. These two algorithms are parallel to the multiclass SVM of Crammer and Singer [5] and enjoy a generalized pairwise comparison setting.

Additionally, Zhu et al. [32] and Saberian and Vasconcelos [21] devised several multiclass boosting algorithms, which solve binary and multicategory problems under the same principle. Mukherjee and Schapire [17] created a general framework for studying multiclass boosting, which formalizes the interaction between the boosting algorithm and the weak learner. We note that Gao and Koller [11] applied the multiclass hinge loss of Crammer and Singer [5] to devise a multiclass boosting algorithm. However, this algorithm is cast under an output coding framework.

### 1.2. Contributions and outline

In this paper, we study the Fisher consistency properties of multicategory surrogate losses. First, assuming that losses are twice differentiable, we present a Fisher consistency property under a more general setting, including the independent and identical, constrained comparison and generalized pairwise comparison settings. We next propose a framework for constructing a majorization function of the 0–1 loss. This framework provides us with a natural and intuitive perspective for construction of three extant multicategory hinge losses. Under this framework, we conduct an in-depth analysis on the Fisher consistency properties of these three extant multicategory hinge losses. In particular, we give a sufficient condition that the multiclass hinge loss used by Vapnik [25], Bredensteiner and Bennett [3], Weston and Watkins [26], Guermeur [12] satisfies the Fisher consistency. Moreover, we constructively derive the minimizers of the expected errors of the multiclass hinge losses of Crammer and Singer [5].

The framework also inspires us to propose a class of multicategory majorization functions which are based on the coherence function [30]. The coherence function is a smooth and convex majorization of the hinge function. Especially, its limit as the temperature approaches zero gives the hinge loss. Moreover, its relationship with the logit loss is also shown. Zhang et al. [30] originally exploited the coherence function in binary classification problems. We investigate its application in the development of multicategory margin classification methods. Based on the coherence function, we in particular present three multicategory coherence losses which correspond to the three extant multicategory hinge losses. These multicategory coherence losses are infinitely smooth and convex and they satisfy the Fisher consistency condition.

The coherence losses have the advantage over the hinge losses that they provide an estimate of the conditional class probability, and over the multicategory logit loss that their limiting versions at zero temperature are just their corresponding multicategory hinge loss functions. Thus they are very appropriate for use in the development of multicategory large margin classification methods, especially boosting algorithms. We propose in this paper a multiclass $\mathcal{C}$ learning algorithm and a multiclass GentleBoost algorithm, both based on our multicategory coherence loss functions.

The remainder of this paper is organized as follows. Section 2 gives a general result on Fisher consistency. In Section 3, we discuss the methodology for the construction of multicategory majorization losses and present two majorization losses based on the coherence function. Section 4 develops another multicategory coherence loss that we call the multiclass $\mathcal{C}$-loss. Based on the multiclass $\mathcal{C}$-loss, a multiclass $\mathcal{C}$ learning algorithm and a multiclass GentleBoost algorithm are given in Section 5. We conduct empirical analysis for the multicategory large margin algorithms in Section 6, and conclude our work in Section 7. All proofs are deferred to the appendix.

## 2. A general result on Fisher consistency

Using the notion and notation given in Section 1.1, we now consider a more general setting than the pairwise comparison. Let $\mathbf{g}^c(\mathbf{x}) = (g_1(\mathbf{x}) - g_c(\mathbf{x}), \ldots, g_{c-1}(\mathbf{x}) - g_c(\mathbf{x}), g_{c+1}(\mathbf{x}) - g_c(\mathbf{x}), \ldots, g_m(\mathbf{x}) - g_c(\mathbf{x}))^T$. We define $\psi_c(\mathbf{g}(\mathbf{x}))$ as a function of $\mathbf{g}^c(\mathbf{x})$ (thereafter denoted $f(\mathbf{g}^c(\mathbf{x}))$). It is clear that the pairwise comparison $\psi_c(\mathbf{g}(\mathbf{x})) = \sum_{j \neq c} \eta(g_c(\mathbf{x}) - g_j(\mathbf{x}))$, the multiclass hinge loss of Crammer and Singer [5], the multicategory $\psi$-loss of Liu and Shen [15], and the truncated hinge loss of Wu and Liu [27] follow this generalized definition. Moreover, for these cases, we note that $f(\mathbf{g}^c)$ is symmetric.[1]

Furthermore, we present a unifying definition of $\boldsymbol{\psi}(\mathbf{g}) = (\psi_1(\mathbf{g}), \ldots, \psi_m(\mathbf{g}))^T : \mathbb{R}^m \to \mathbb{R}^m$ where we ignore the dependency of $\mathbf{g}$ on $\mathbf{x}$. Let $\Psi$ be a set of mappings $\boldsymbol{\psi}(\mathbf{g})$ satisfying the conditions: (i) when fixed $g_c$ $\psi_c(\mathbf{g})$ is symmetric w.r.t. the remaining arguments and (ii) $\psi_c(\mathbf{g}) = \psi_j(\mathbf{g}^{jc})$ where $\mathbf{g}^{jc}$ is obtained by only exchanging $g_c$ and $g_j$ of $\mathbf{g}$. Obviously, the mapping $\boldsymbol{\psi}(\mathbf{g})$ defined via the independent and identical setting, the constrained comparison, or the generalized pairwise comparison with symmetric $f$ belongs to $\Psi$. With this notion, we give an important theorem of this paper as follows.

**Theorem 3.** *Let $\boldsymbol{\psi}(\mathbf{g}) \in \Psi$ be a twice differentiable function from $\mathbb{R}^m$ to $\mathbb{R}^m$. Assume that the Hessian matrix of $\psi_c(\mathbf{g})$ w.r.t. $\mathbf{g}$ is conditionally positive definite for $c = 1, \ldots, m$. Then the minimizer $\hat{\mathbf{g}} = (\hat{g}_1, \ldots, \hat{g}_m)^T$ of $\sum_c \psi_c(\mathbf{g}(\mathbf{x}))P_c(\mathbf{x})$ in $\mathcal{G}$ exists and is unique. Furthermore, if $\frac{\partial \psi_c(\mathbf{g})}{\partial g_c} - \frac{\partial \psi_c(\mathbf{g})}{\partial g_j}$ where $j \neq c$ is negative for any $\mathbf{g} \in \mathcal{G}$, then $P_l > P_k$ implies $\hat{g}_l > \hat{g}_k$.*

The proof of the theorem is given in Appendix A.1. Note that an $m \times m$ real matrix $\mathbf{A}$ is said to be conditionally positive definite if $\mathbf{y}^T \mathbf{A} \mathbf{y} > 0$ for any nonzero real vector $\mathbf{y} = (y_1, \ldots, y_m)^T$ with $\sum_{j=1}^m y_j = 0$. The condition that $\frac{\partial \psi_c(\mathbf{g})}{\partial g_c} - \frac{\partial \psi_c(\mathbf{g})}{\partial g_j} < 0$ on $\mathcal{G}$ for $j \neq c$ is not necessary for Fisher consistency. For example, in the setting $\psi_c(\mathbf{g}(\mathbf{x})) = \eta(g_c(\mathbf{x}))$, Zou et al. [33] proved that if $\eta(z)$ is a twice differentiable function with $\eta'(0) < 0$ and $\eta''(z) > 0$ $\forall z$, then $\psi_c(\mathbf{g}(\mathbf{x}))$ is Fisher-consistent. In the setting $\psi_c(\mathbf{g}(\mathbf{x})) = \sum_{j \neq c} \eta(-g_j)$, we note that $\sum_c \sum_{j \neq c} \eta(-g_j)P_c = \sum_{c=1} \eta(-g_c)(1 - P_c)$. Based on the proof of Zou et al. [33], we have that if $\eta(z)$ is a twice differentiable function with $\eta'(0) < 0$ and $\eta''(z) > 0$ $\forall z$, then $\psi_c(\mathbf{g}(\mathbf{x})) = \sum_{j \neq c} \eta(-g_j(\mathbf{x}))$ is Fisher-consistent. That is, in these two cases, we can relax the condition that $\frac{\partial \psi_c(\mathbf{g})}{\partial g_c} - \frac{\partial \psi_c(\mathbf{g})}{\partial g_j} < 0$ for any $\mathbf{g} \in \mathcal{G}$ as $\frac{\partial \psi_c(\mathbf{0})}{\partial g_c} - \frac{\partial \psi_c(\mathbf{0})}{\partial g_j} < 0$, for $j \neq c$.

We have the following corollary, whose proof is given in Appendix A.2. We will see two concrete cases of this corollary (that is, Theorems 10 and 13).

**Corollary 4.** *Assume $\psi_c(\mathbf{g}) = f(\mathbf{g}^c)$ where $f(\mathbf{z})$ is a symmetric and twice differentiable function from $\mathbb{R}^{m-1}$ to $\mathbb{R}$. If the Hessian matrix of $f(\mathbf{z})$ w.r.t. $\mathbf{z}$ is positive definite, then the minimizer $\hat{\mathbf{g}} = (\hat{g}_1, \ldots, \hat{g}_m)^T$ of $\sum_c \psi_c(\mathbf{g}(\mathbf{x}))P_c(\mathbf{x})$ in $\mathcal{G}$ exists and is unique. Furthermore, if $\frac{\partial \psi_c(\mathbf{g})}{\partial g_c} - \frac{\partial \psi_c(\mathbf{g})}{\partial g_j}$ where $j \neq c$ is negative for any $\mathbf{g}$, then $P_l > P_k$ implies $\hat{g}_l > \hat{g}_k$.*

Theorem 3 or Corollary 4 shows that $\psi_c(\mathbf{g})$ admits the ISC of Zhang [28]. Thus, under the conditions in Theorem 3 or Corollary 4, we also have the relationship between the approximate minimization of the risk based on $\psi_c$ and the approximate minimization of the classification error. In particular, if

$$\mathbb{E}_X\left[\sum_{c=1}^m \psi_c\big(\hat{\mathbf{g}}(X)\big)P_c(X)\right] \le \inf_{\mathbf{g} \in \mathcal{G}} \mathbb{E}_X\left[\sum_{c=1}^m \psi_c\big(\mathbf{g}(X)\big)P_c(X)\right] + \epsilon_1$$

for some $\epsilon_1 > 0$, then there exists an $\epsilon_2 > 0$ such that

$$\mathbb{E}_X\left[\sum_{c=1, c \neq \hat{\phi}(X)}^m P_c(X)\right] \le \mathbb{E}_X\left[\sum_{c=1, c \neq \phi^*(X)}^m P_c(X)\right] + \epsilon_2$$

where $\hat{\phi}(X) = \operatorname{argmax}_j \{\hat{g}_j(X)\}$, $\phi^*(X) = \operatorname{argmax}_j \{P_j(X)\}$ and $\mathbb{E}_X[\sum_{c=1, c \neq \phi^*(X)}^m P_c(X)]$ is the optimal error. This result directly follows from Theorem 3 in Zhang [28].

## 3. Multicategory majorization losses

Given $\mathbf{x}$ and its label $c$, we let $\mathbf{g}(\mathbf{x})$ be a margin vector at $\mathbf{x}$ and the induced classifier be $\phi(\mathbf{x}) = \operatorname{argmax}_j g_j(\mathbf{x})$. In the binary case, it is clear that $\phi(\mathbf{x}) = c$ if and only if $g_c(\mathbf{x}) > 0$, and that $g_c(\mathbf{x}) \le 0$ is a necessary and sufficient condition of

[1] A symmetric function of $p$ variables is one whose value at any $p$-tuple of arguments is the same as its value at any permutation of that $p$-tuple.

$\phi(\mathbf{x}) \neq c$. Thus, we always have $\mathbb{I}_{\{g_c(\mathbf{x})\leq 0\}} = \mathbb{I}_{\{\phi(\mathbf{x})\neq c\}}$. Furthermore, let $g_1(\mathbf{x}) = -g_2(\mathbf{x}) = \frac{1}{2} f(\mathbf{x})$ and encode $y = 1$ if $c = 1$ and $y = -1$ if $c = 2$. Then the empirical error is

$$\epsilon = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{\phi(\mathbf{x}_i)\neq c_i\}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{g_{c_i}(\mathbf{x}_i)\leq 0\}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{y_i f(\mathbf{x}_i)\leq 0\}}.$$

In the multicategory case, $\phi(\mathbf{x}) = c$ implies $g_c(\mathbf{x}) > 0$ but $\phi(\mathbf{x}) \neq c$ does not imply $g_c(\mathbf{x}) \leq 0$. We shall see that $g_c(\mathbf{x}) \leq 0$ is a sufficient but not necessary condition of $\phi(\mathbf{x}) \neq c$. In general, we only have $\mathbb{I}_{\{g_c(\mathbf{x})\leq 0\}} \leq \mathbb{I}_{\{\phi(\mathbf{x})\neq c\}}$. Although $g_j(\mathbf{x}) - g_c(\mathbf{x}) > 0$ for some $j \neq c$ is a necessary and sufficient condition of $\phi(\mathbf{x}) \neq c$, it in most cases yields an optimization problem which is not easily solved. This is an important reason why it is not trivial to develop multicategory AdaBoost and SVMs with the same principle as binary AdaBoost and SVMs.

### 3.1. Methodology

Recall that $\sum_{j=1}^{m} g_j(\mathbf{x}) = 0$ and there is at least one $j \in \{1, \ldots, m\}$ such that $g_j(\mathbf{x}) \neq 0$. If $g_c(\mathbf{x}) \leq 0$, then there exists one $l \in \{1, \ldots, m\}$ such that $l \neq c$ and $g_l(\mathbf{x}) > 0$. As a result, we have $\phi(\mathbf{x}) \neq c$. Therefore, $g_c(\mathbf{x}) \leq 0$ implies $\phi(\mathbf{x}) \neq c$. Unfortunately, if $\phi(\mathbf{x}) \neq c$, $g_c(\mathbf{x}) \leq 0$ does not necessarily hold. For example, consider the case that $m = 3$ and $c = 1$. Assume that $\mathbf{g}(\mathbf{x}) = (2, 3, -5)$. Then we have $\phi(\mathbf{x}) = 2 \neq 1$ and $g_1(\mathbf{x}) = 2 > 0$. In addition, it is clear that $\phi(\mathbf{x}) = c$ implies $g_c(\mathbf{x}) > 0$. However, $g_c(\mathbf{x}) > 0$ does not imply $\phi(\mathbf{x}) = c$.

On the other hand, it is obvious that $\phi(\mathbf{x}) = c$ is equivalent to $\phi(\mathbf{x}) \neq j$ for all $j \neq c$. In terms of the above discussions, a condition of making $\phi(\mathbf{x}) = c$ is $g_j(\mathbf{x}) \leq 0$ for $j \neq c$. To summarize, we immediately have the following theorem.

**Proposition 5.** *For $(\mathbf{x}, c)$, let $\mathbf{g}(\mathbf{x})$ be a margin vector at $\mathbf{x}$ and the induced classifier be $\phi(\mathbf{x}) = \arg\max_j g_j(\mathbf{x})$. Then*

(a) $\mathbb{I}_{\{g_c(\mathbf{x})\leq 0\}} \leq \mathbb{I}_{\{\phi(\mathbf{x})\neq c\}} = \mathbb{I}_{\{\bigcup_{j\neq c} g_j(\mathbf{x})-g_c(\mathbf{x})>0\}} \leq \mathbb{I}_{\{\bigcup_{j\neq c} g_j(\mathbf{x})>0\}}$

(b) $\mathbb{I}_{\{\bigcap_{j\neq c} g_j(\mathbf{x})\leq 0\}} \leq \mathbb{I}_{\{\bigcap_{j\neq c} g_j(\mathbf{x})-g_c(\mathbf{x})\leq 0\}} = \mathbb{I}_{\{\phi(\mathbf{x})=c\}} \leq \mathbb{I}_{\{g_c(\mathbf{x})>0\}}$

(c) $\mathbb{I}_{\{\bigcup_{j\neq c} g_j(\mathbf{x})>0\}} \leq \sum_{j\neq c} \mathbb{I}_{\{g_j(\mathbf{x})>0\}}$

(d) $\mathbb{I}_{\{\bigcup_{j\neq c} g_j(\mathbf{x})-g_c(\mathbf{x})>0\}} \leq \sum_{j\neq c} \mathbb{I}_{\{g_j(\mathbf{x})-g_c(\mathbf{x})>0\}}$.

Proposition 5 shows that $g_c(\mathbf{x}) \leq 0$ is the sufficient condition of $\phi(\mathbf{x}) \neq c$, while $g_j(\mathbf{x}) > 0$ for some $j \neq c$ is its necessary condition. The following theorem shows that they become sufficient and necessary when $\mathbf{g}$ has one and only one positive element.

**Proposition 6.** *Under the conditions in Proposition 5. The relationship of*

$$\mathbb{I}_{\{g_c(\mathbf{x})\leq 0\}} = \mathbb{I}_{\{\phi(\mathbf{x})\neq c\}} = \mathbb{I}_{\{\bigcup_{j\neq c} g_j(\mathbf{x})-g_c(\mathbf{x})>0\}} = \mathbb{I}_{\{\bigcup_{j\neq c} g_j(\mathbf{x})>0\}} = \sum_{j\neq c} \mathbb{I}_{\{g_j(\mathbf{x})>0\}}$$

*holds if and only if the margin vector $\mathbf{g}(\mathbf{x})$ has only one positive element.*

In the binary case, this relationship always holds because $g_1(\mathbf{x}) = -g_2(\mathbf{x})$. Recently, Zou et al. [33] derived multicategory boosting algorithms using $\exp(-g_c(\mathbf{x}))$. In their discrete boosting algorithm, the margin vector $\mathbf{g}(\mathbf{x})$ is modeled as an $m$-vector function with one and only one positive element. In this case, $\mathbb{I}_{\{g_c(\mathbf{x})\leq 0\}}$ is equal to $\mathbb{I}_{\{\phi(\mathbf{x})\neq c\}}$. Consequently, $\exp(-g_c(\mathbf{x}))$ is a majorization of $\mathbb{I}_{\{\phi(\mathbf{x})\neq c\}}$ because $\exp(-g_c(\mathbf{x}))$ is an upper bound of $\mathbb{I}_{\{g_c(\mathbf{x})\leq 0\}}$. Therefore, this discrete AdaBoost algorithm still approximates the original empirical 0–1 loss function. In the general case, however, Proposition 6 implies that $\exp(-g_c(\mathbf{x}))$ is not the majorization of $\mathbb{I}_{\{\phi(\mathbf{x})\neq c\}}$.

### 3.2. Approaches

Proposition 5 provides us with approaches for constructing majorization functions of the 0–1 loss function $\mathbb{I}_{\{\phi(\mathbf{x})\neq c\}}$. Clearly, $\sum_{j\neq c} \mathbb{I}_{\{g_j(\mathbf{x})>0\}}$ and $\sum_{j\neq c} \mathbb{I}_{\{g_j(\mathbf{x})-g_c(\mathbf{x})>0\}}$ are separable, so they are more tractable respectively than $\mathbb{I}_{\{\bigcup_{j\neq c} g_j(\mathbf{x})>0\}}$ and $\mathbb{I}_{\{\bigcup_{j\neq c} g_j(\mathbf{x})-g_c(\mathbf{x})>0\}}$. Thus, $\sum_{j\neq c} \mathbb{I}_{\{g_j(\mathbf{x})>0\}}$ and $\sum_{j\neq c} \mathbb{I}_{\{g_j(\mathbf{x})-g_c(\mathbf{x})>0\}}$ are popularly employed in practical applications.

In particular, suppose $\eta(g_j(\mathbf{x}))$ upper bounds $\mathbb{I}_{\{g_j(\mathbf{x})\leq 0\}}$; that is, $\eta(g_j(\mathbf{x})) \geq \mathbb{I}_{\{g_j(\mathbf{x})\leq 0\}}$. Note that $\eta(g_j(\mathbf{x})) \geq \mathbb{I}_{\{g_j(\mathbf{x})\leq 0\}}$ if and only if $\eta(-g_j(\mathbf{x})) \geq \mathbb{I}_{\{g_j(\mathbf{x})\geq 0\}}$. Thus $\eta(-g_j(\mathbf{x}))$ upper bounds $\mathbb{I}_{\{g_j(\mathbf{x})\geq 0\}}$, and hence $\eta(g_j(\mathbf{x}) - g_l(\mathbf{x}))$ upper bounds $\mathbb{I}_{\{g_l(\mathbf{x})-g_j(\mathbf{x})>0\}}$. It then follows from Proposition 5 that $\sum_{j\neq c} \eta(-g_j(\mathbf{x}))$ and $\sum_{j\neq c} \eta(g_c(\mathbf{x}) - g_j(\mathbf{x}))$ are majorizations of $\mathbb{I}_{\{\phi(\mathbf{x})\neq c\}}$. Consequently, we can define two classes of majorizations for $\mathbb{I}_{\{\phi(\mathbf{x})\neq c\}}$. The first one is

$$\psi_c\big(\mathbf{g}(\mathbf{x})\big) = \sum_{l \neq c} \eta\big(g_c(\mathbf{x}) - g_l(\mathbf{x})\big), \tag{2}$$

while the second one is

$$\psi_c\big(\mathbf{g}(\mathbf{x})\big) = \sum_{l \neq c} \eta\big(-g_l(\mathbf{x})\big). \tag{3}$$

This leads us to two approaches for constructing majorization $\psi_c(\mathbf{g}(\mathbf{x}))$ of $\mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}}$. Zhang [28] referred to them as *pairwise comparison* and *constrained comparison*. A theoretical analysis of these two classes of majorization functions has also been presented by Zhang [28]. His analysis mainly focused on consistency of empirical risk minimization and the ISC property of surrogate losses. Our results in Section 3.1 show a direct and intuitive connection of these two approaches with the original 0–1 loss.

### 3.3. Multicategory hinge losses

Using distinct $\eta(g_j(\mathbf{x}))$ $(\geq \mathbb{I}_{\{g_j(\mathbf{x}) \leq 0\}})$ in the two approaches, we can construct different multicategory losses for large margin classifiers. For example, let $\eta(g_j(\mathbf{x})) = (1 - g_j(\mathbf{x}))_+$ which upper bounds $\mathbb{I}_{\{g_j(\mathbf{x}) \leq 0\}}$. Then $\sum_{j \neq c}(1 - g_c(\mathbf{x}) + g_j(\mathbf{x}))_+$ and $\sum_{j \neq c}(1 + g_j(\mathbf{x}))_+$ are candidate majorizations for $\mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}}$, which yield two multiclass SVM methods.

In the multicategory SVM (MSVM), Lee et al. [13] employed $\sum_{j \neq c}(1 + g_j(\mathbf{x}))_+$ as a multicategory hinge loss. Moreover, Lee et al. [13] proved that this multicategory hinge loss is Fisher-consistent. In particular, the minimizer of $\sum_{c=1}^{m} \sum_{j \neq c}(1 + g_j(\mathbf{x}))_+ P_c(\mathbf{x})$ w.r.t. $\mathbf{g} \in \mathcal{G}$ is $\hat{g}_l(\mathbf{x}) = m - 1$ if $l = \operatorname{argmax}_j(P_j(\mathbf{x}))$ and $\hat{g}_l(\mathbf{x}) = -1$ otherwise.

The pairwise comparison $\sum_{j \neq c}(1 - g_c(\mathbf{x}) + g_j(\mathbf{x}))_+$ was used by Vapnik [25], Weston and Watkins [26], Bredensteiner and Bennett [3], Guermeur [12]. Unfortunately, Lee et al. [13], Zhang [28], Liu [14] showed that solutions of the corresponding optimization problem do not always implement the Bayes decision rule. However, we find that it is still Fisher-consistent under certain conditions. In particular, we have the following theorem (the proof is given in Appendix B.1).

**Theorem 7.** *Let $P_j(\mathbf{x}) > 0$ for $j = 1, \ldots, m$, $P_l(\mathbf{x}) = \max_j P_j(\mathbf{x})$ and $P_k(\mathbf{x}) = \max_{j \neq l} P_j(\mathbf{x})$, and let*

$$\hat{\mathbf{g}}(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{g}(\mathbf{x}) \in \mathcal{G}} \sum_{c=1}^{m} P_c(\mathbf{x}) \sum_{j \neq c} \big(1 - g_c(\mathbf{x}) + g_j(\mathbf{x})\big)_+ .$$

*If $P_l(\mathbf{x}) > 1/2$ or $P_k(\mathbf{x}) < 1/m$, then $\hat{g}_l(\mathbf{x}) = 1 + \hat{g}_k(\mathbf{x}) \geq 1 + \hat{g}_j(\mathbf{x})$ for $j \neq l, k$.*

This theorem implies that $\hat{g}_l(\mathbf{x}) > \hat{g}_j(\mathbf{x})$, so the majorization function $\sum_{j \neq c}(1 - g_c(\mathbf{x}) + g_j(\mathbf{x}))_+$ is Fisher-consistent when $P_l(\mathbf{x}) > 1/2$ or $P_k(\mathbf{x}) < 1/m$. In the case that $m = 3$, Liu [14] showed that this majorization function yields the Fisher consistency when $P_k < \frac{1}{3}$, while the consistency is not always satisfied when $1/2 > P_l > P_k \geq 1/3$. Theorem 7 shows that for any $m \geq 3$ the consistency is also satisfied whenever $P_k < \frac{1}{m}$.

As we have seen, $\mathbb{I}_{\{\bigcup_{j \neq c} g_j(\mathbf{x}) - g_c(\mathbf{x}) > 0\}}$ can be also used as a starting point to construct a majorization of $\mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}}$. Since $\mathbb{I}_{\{\bigcup_{j \neq c} g_j(\mathbf{x}) - g_c(\mathbf{x}) > 0\}} = \mathbb{I}_{\{\max_{j \neq c} g_j(\mathbf{x}) - g_c(\mathbf{x}) > 0\}}$, we call this construction approach the *maximum pairwise comparison*. In fact, this approach was employed by Crammer and Singer [5], Liu and Shen [15] and Wu and Liu [27]. Especially, Crammer and Singer [5] used the surrogate:

$$\xi_c\big(\mathbf{g}(\mathbf{x})\big) = \max\big\{g_j(\mathbf{x}) + 1 - \mathbb{I}_{\{j=c\}}\big\} - g_c(\mathbf{x}). \tag{4}$$

It is easily seen that

$$\mathbb{I}_{\{\bigcup_{j \neq c} g_j(\mathbf{x}) - g_c(\mathbf{x}) > 0\}} \leq \max_j \big\{g_j(\mathbf{x}) + 1 - \mathbb{I}_{\{j=c\}}\big\} - g_c(\mathbf{x}) \leq \sum_{j \neq c} \big(1 + g_j(\mathbf{x}) - g_c(\mathbf{x})\big)_+ ,$$

which implies that $\xi_c(\mathbf{g}(\mathbf{x}))$ is a tighter upper bound of $\mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}}$ than $\sum_{j \neq c}(1 - g_c(\mathbf{x}) + g_j(\mathbf{x}))_+$. Note that Crammer and Singer [5] did not assume $\mathbf{g} \in \mathcal{G}$, but Liu and Shen [15] argued that this assumption is also necessary. Zhang [28] showed that $\xi_c(\mathbf{g}(\mathbf{x}))$ is Fisher-consistent only when $P_l(\mathbf{x}) > 1/2$. However, the author did not give an explicit expression of the minimizer of the expected error in question in the literature. Here we present the constructive solution of the corresponding minimization problem in the following theorem (the proof is given in Appendix B.2).

**Theorem 8.** *Consider the following optimization problem of*

$$\hat{\mathbf{g}}(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{g}(\mathbf{x}) \in \mathcal{G}} \sum_{c=1}^{m} \Big\{ \max_j \big(g_j(\mathbf{x}) + 1 - \mathbb{I}_{\{j=c\}}\big) - g_c(\mathbf{x}) \Big\} P_c(\mathbf{x}). \tag{5}$$

*Assume that $P_l(\mathbf{x}) = \max_j P_j(\mathbf{x})$.*

(1) *If $P_l(\mathbf{x}) > 1/2$, then $\hat{g}_j(\mathbf{x}) = \mathbb{I}_{\{j=l\}} - \frac{1}{m}$ for $j = 1, \dots, m$;*
(2) *If $P_l(\mathbf{x}) = 1/2$, then $0 \leq \hat{g}_l(\mathbf{x}) - \hat{g}_j(\mathbf{x}) \leq 1$ and $\hat{g}_j(\mathbf{x}) = \hat{g}_c(\mathbf{x})$ for $c, j \neq l$;*
(3) *If $P_l(\mathbf{x}) < 1/2$, then $\hat{g}_c(\mathbf{x}) = 0$ for $c = 1, \dots, m$.*

This theorem shows that the majorization function $\max_j\{g_j(\mathbf{x}) + 1 - \mathbb{I}_{\{j=c\}}\} - g_c(\mathbf{x})$ is Fisher-consistent when $P_l(\mathbf{x}) > 1/2$. Otherwise, the solution of (5) degenerates to the trivial point. As we have seen from Theorems 7 and 8, $P_l(\mathbf{x}) > 1/2$ is a sufficient condition for both $\max_j\{g_j(\mathbf{x}) + 1 - \mathbb{I}_{\{j=c\}}\} - g_c(\mathbf{x})$ and $\sum_{j\neq c}(1 - g_c(\mathbf{x}) + g_j(\mathbf{x}))_+$ to be Fisher-consistent. Moreover, they satisfy the condition $\hat{g}_l(\mathbf{x}) = 1 + \hat{g}_k(\mathbf{x})$ where $k = \operatorname{argmax}_{j\neq l} P_j(\mathbf{x})$. However, as shown in Theorem 7, $\sum_{j\neq c}(1 - g_c(\mathbf{x}) + g_j(\mathbf{x}))_+$ still yields the Fisher-consistent property when $P_k < \frac{1}{m}$. Thus, the consistency condition for the pairwise comparison hinge loss is weaker than that for the maximum pairwise comparison hinge loss.

### 3.4. Multicategory coherence losses

To construct a smooth majorization function of $\mathbb{I}_{\{\phi(\mathbf{x})\neq c\}}$, we define $\eta(g_c(\mathbf{x}))$ as the coherence function which was proposed by Zhang et al. [30]. The coherence function is

$$\eta_T(z) \triangleq T \log\left[1 + \exp\frac{1-z}{T}\right], \quad T > 0 \tag{6}$$

where $T$ is called the temperature parameter. Clearly, $\eta_T(z) \geq (1-z)_+ \geq \mathbb{I}_{\{z\leq 0\}}$. Moreover, $\lim_{T\to 0}\eta_T(z) = (1-z)_+$. Thus, we directly have two majorizations of $\mathbb{I}_{\{\phi(\mathbf{x})\neq c\}}$ based on the constrained comparison method and the pairwise comparison method.

Using the constrained comparison, we give a smooth approximation to $\sum_{j\neq c}(1 + g_j(\mathbf{x}))_+$ for the MSVM of Lee et al. [13]. That is,

$$L_T\big(\mathbf{g}(\mathbf{x}), c\big) \triangleq T \sum_{j\neq c} \log\left[1 + \exp\left(\frac{1 + g_j(\mathbf{x})}{T}\right)\right].$$

It is immediate that $L_T(\mathbf{g}(\mathbf{x}), c) \geq \sum_{j\neq c}(1 + g_j(\mathbf{x}))_+$ and $\lim_{T\to 0} L_T(\mathbf{g}(\mathbf{x}), c) = \sum_{j\neq c}(1 + g_j(\mathbf{x}))_+$. Furthermore, we have the following theorem (the proof is given in Appendix B.3).

**Theorem 9.** *Assume that $P_c(\mathbf{x}) > 0$ for $c = 1, \dots, m$. Consider the optimization problem*

$$\max_{\mathbf{g}(\mathbf{x})\in\mathcal{G}} \sum_{c=1}^m L_T\big(\mathbf{g}(\mathbf{x}), c\big) P_c(\mathbf{x}) \tag{7}$$

*for a fixed $T > 0$ and let $\hat{\mathbf{g}}(\mathbf{x}) = (\hat{g}_1(\mathbf{x}), \dots, \hat{g}_m(\mathbf{x}))^T$ be its solution. Then $\hat{\mathbf{g}}(\mathbf{x})$ is unique. Moreover, if $P_l(\mathbf{x}) < P_j(\mathbf{x})$, we have $\hat{g}_l(\mathbf{x}) < \hat{g}_j(\mathbf{x})$. Furthermore, we have*

$$\lim_{T\to 0} \hat{g}_c(\mathbf{x}) = \begin{cases} m-1 & \textit{if } c = \operatorname{argmax}_j P_j(\mathbf{x}), \\ -1 & \textit{otherwise}. \end{cases}$$

*Additionally, having obtained $\hat{\mathbf{g}}(\mathbf{x})$, $P_c(\mathbf{x})$ is given by*

$$P_c(\mathbf{x}) = 1 - \frac{(m-1)(1 + \exp(-\frac{1+\hat{g}_c(\mathbf{x})}{T}))}{m + \sum_{j=1}^m \exp(-\frac{1+\hat{g}_j(\mathbf{x})}{T})}. \tag{8}$$

Although there is no explicit expression for $\hat{\mathbf{g}}(\mathbf{x})$ in Problem (7), Theorem 9 shows that its limit at $T = 0$ is equal to the minimizer of $\sum_{c=1}^m \sum_{j\neq c}(1 + g_j(\mathbf{x}))_+ P_c(\mathbf{x})$, which was studied by Lee et al. [13].

Based on the pairwise comparison, we have a smooth alternative to multiclass hinge loss $\sum_{j\neq c}(1 + g_c(\mathbf{x}) - g_j(\mathbf{x}))_+$, which is

$$G_T\big(\mathbf{g}(\mathbf{x}), c\big) \triangleq T \sum_{j\neq c} \log\left[1 + \exp\left(\frac{1 + g_j(\mathbf{x}) - g_c(\mathbf{x})}{T}\right)\right]. \tag{9}$$

It is also immediate that $G_T(\mathbf{g}(\mathbf{x}), c) \geq \sum_{j\neq c}(1 + g_c(\mathbf{x}) - g_j(\mathbf{x}))_+$ and $\lim_{T\to 0} G_T(\mathbf{g}(\mathbf{x}), c) = \sum_{j\neq c}(1 + g_c(\mathbf{x}) - g_j(\mathbf{x}))_+$.

**Theorem 10.** *Assume that $P_c(\mathbf{x}) > 0$ for $c = 1, \dots, m$. Let $P_l = \max_j P_j(\mathbf{x})$ and $P_k(\mathbf{x}) = \max_{j\neq l} P_j(\mathbf{x})$. Consider the optimization problem*

$$\max_{\mathbf{g}(\mathbf{x})\in\mathcal{G}} \sum_{c=1}^m G_T\big(\mathbf{g}(\mathbf{x}), c\big) P_c(\mathbf{x})$$

*for a fixed $T > 0$ and let $\hat{\mathbf{g}}(\mathbf{x}) = (\hat{g}_1(\mathbf{x}), \dots, \hat{g}_m(\mathbf{x}))^T$ be its solution. Then $\hat{\mathbf{g}}(\mathbf{x})$ is unique. Moreover, if $P_i(\mathbf{x}) < P_j(\mathbf{x})$, we have $\hat{g}_i(\mathbf{x}) < \hat{g}_j(\mathbf{x})$. Additionally, if $P_l(\mathbf{x}) > 1/2$ or $P_k(\mathbf{x}) < 1/m$, then*

$$\lim_{T\to 0} \hat{g}_l(\mathbf{x}) = 1 + \lim_{T\to 0} \hat{g}_k(\mathbf{x}) \geq 1 + \lim_{T\to 0} \hat{g}_j(\mathbf{x}) \quad \textit{for } j \neq l, k,$$

*whenever the limits exist.*

The proof of Theorem 10 is given in Appendix B.4. We see that the limit of $\hat{g}_l(\mathbf{x})$ at $T = 0$ agrees with that shown in Theorem 7. Unfortunately, based on $G_T(\mathbf{g}(\mathbf{x}), c)$, it is hard to obtain an explicit expression of the class conditional probabilities $P_c(\mathbf{x})$ via the $\hat{g}_c(\mathbf{x})$.

## 4. Multiclass $\mathcal{C}$-losses

In this section, we present a smooth and Fisher-consistent majorization of the multiclass hinge loss $\xi_c(\mathbf{g}(\mathbf{x}))$ in (4) using the idea behind the coherence function. We call this new majorization *multiclass $\mathcal{C}$-loss*. We will see that this multiclass $\mathcal{C}$-loss bridges the multiclass hinge loss $\xi_c(\mathbf{g}(\mathbf{x}))$ and the negative multinomial log-likelihood (logit) of the form

$$\gamma_c\big(\mathbf{g}(\mathbf{x})\big) = \log\sum_{j=1}^{m}\exp\big(g_j(\mathbf{x}) - g_c(\mathbf{x})\big) = \log\left[1 + \sum_{j\neq c}\exp\big(g_j(\mathbf{x}) - g_c(\mathbf{x})\big)\right]. \tag{10}$$

In the 0–1 loss the misclassification costs are specified as 1. It is natural to set the misclassification costs as a positive constant $u > 0$. This setting will reveal an important connection between the hinge loss and the logit loss. The empirical error on the training data is then

$$\epsilon = \frac{u}{n}\sum_{i=1}^{n}\mathbb{I}_{\{\phi(\mathbf{x}_i)\neq c_i\}}.$$

In this setting, we can extend the multiclass hinge loss $\xi_c(\mathbf{g}(\mathbf{x}))$ as

$$H_u\big(\mathbf{g}(\mathbf{x}), c\big) = \max_j\big\{g_j(\mathbf{x}) + u - u\mathbb{I}_{\{j=c\}}\big\} - g_c(\mathbf{x}). \tag{11}$$

It is clear that $H_u(\mathbf{g}(\mathbf{x}), c) \geq u\mathbb{I}_{\{\phi(\mathbf{x})\neq c\}}$. To establish the connection among the multiclass $\mathcal{C}$-loss, the multiclass hinge loss and the logit loss, we employ this setting to present the definition of the multiclass $\mathcal{C}$-loss.

We now express $\max\{g_j(\mathbf{x}) + u - u\mathbb{I}_{\{j=c\}}\}$ as $\sum_{j=1}^{m}\omega_j^c(\mathbf{x})[g_j(\mathbf{x}) + u - u\mathbb{I}_{\{j=c\}}]$ where

$$\omega_j^c(\mathbf{x}) = \begin{cases} 1 & j = \operatorname{argmax}_l\{g_l(\mathbf{x}) + u - u\mathbb{I}_{\{l=c\}}\} \\ 0 & \text{otherwise.}\end{cases}$$

Motivated by the idea behind deterministic annealing [20], we relax this hard function $\omega_j^c(\mathbf{x})$, retaining only $\omega_j^c(\mathbf{x}) \geq 0$ and $\sum_{j=1}^{m}\omega_j^c(\mathbf{x}) = 1$. With such soft $\omega_j^c(\mathbf{x})$, we maximize $\sum_{j=1}^{m}\omega_j^c(\mathbf{x})[g_j(\mathbf{x}) + u - u\mathbb{I}_{\{j=c\}}] - g_c(\mathbf{x})$ under entropy penalization; namely,

$$\max_{\{\omega_j^c(\mathbf{x})\}}\left\{F \triangleq \sum_{j=1}^{m}\omega_j^c(\mathbf{x})\big[g_j(\mathbf{x}) + u - u\mathbb{I}_{\{j=c\}}\big] - g_c(\mathbf{x}) - T\sum_{j=1}^{m}\omega_j^c(\mathbf{x})\log\omega_j^c(\mathbf{x})\right\}, \tag{12}$$

where $T > 0$ is also referred to as the temperature. The maximization of $F$ w.r.t. $\omega_j^c(\mathbf{x})$ is straightforward, and it gives rise to the following distribution

$$\omega_j^c(\mathbf{x}) = \frac{\exp[\frac{g_j(\mathbf{x}) + u - u\mathbb{I}_{\{j=c\}}}{T}]}{\sum_l \exp[\frac{g_l(\mathbf{x}) + u - u\mathbb{I}_{\{l=c\}}}{T}]} \tag{13}$$

based on the Karush–Kuhn–Tucker condition. The corresponding maximum of $F$ is obtained by plugging (13) back into (12):

$$C_{T,u}\big(\mathbf{g}(\mathbf{x}), c\big) \triangleq T\log\left[1 + \sum_{j\neq c}\exp\frac{u + g_j(\mathbf{x}) - g_c(\mathbf{x})}{T}\right], \quad T > 0,\ u > 0. \tag{14}$$

Note that for $T > 0$ we have

$$\begin{aligned} T \log\left[1 + \sum_{j \neq c} \exp \frac{u + g_j(\mathbf{x}) - g_c(\mathbf{x})}{T}\right] &= T \log \sum_j \exp\left(\frac{g_j(\mathbf{x}) + u - u\mathbb{I}_{\{j=c\}}}{T}\right) - g_c(\mathbf{x}) \\ &\geq \max_j \left\{g_j(\mathbf{x}) + u - u\mathbb{I}_{\{j=c\}}\right\} - g_c(\mathbf{x}) \\ &\geq u\mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}}. \end{aligned}$$

We thus call $C_{T,u}(\mathbf{g}(\mathbf{x}), c)$ the multiclass $\mathcal{C}$-loss. Clearly, $C_{T,u}(\mathbf{g}(\mathbf{x}), c)$ is infinitely smooth and convex in $\mathbf{g}(\mathbf{x})$ (see Appendix C for the proof). Moreover, the Hessian matrix of $C_{T,u}(\mathbf{g}(\mathbf{x}), c)$ w.r.t. $\mathbf{g}(\mathbf{x})$ is conditionally positive definite.

### 4.1. Properties

We now investigate the relationships between the multiclass $\mathcal{C}$-loss $C_{T,u}(\mathbf{g}(\mathbf{x}), c)$ and the multiclass hinge loss $H_u(\mathbf{g}(\mathbf{x}), c)$, and between $C_{T,u}(\mathbf{g}(\mathbf{x}), c)$ and multiclass coherence loss $G_T(\mathbf{g}(\mathbf{x}), c)$. In particular, we have the following proposition.

**Proposition 11.** *Let $G_T(\mathbf{g}(\mathbf{x}), c)$, $H_u(\mathbf{g}(\mathbf{x}), c)$ and $C_{T,u}(\mathbf{g}(\mathbf{x}), c)$ be defined by (9), (11) and (14), respectively. Then,*

(i) $\mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}} \leq C_{T,1}(\mathbf{g}(\mathbf{x}), c) < G_T(\mathbf{g}(\mathbf{x}), c)$.
(ii) $H_u(\mathbf{g}(\mathbf{x}), c) \leq C_{T,u}(\mathbf{g}(\mathbf{x}), c) \leq H_u(\mathbf{g}(\mathbf{x}), c) + T \log m$.

The proof is given in Appendix C.1. We see from this proposition that $C_{T,1}(\mathbf{g}(\mathbf{x}), c)$ is a majorization of $\mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}}$ tighter than $G_T(\mathbf{g}(\mathbf{x}), c)$. When treating $\mathbf{g}(\mathbf{x})$ fixed and considering $\omega_j^c(\mathbf{x})$ and $C_{T,u}(\mathbf{g}(\mathbf{x}), c)$ as functions of $T$, we have the following proposition.

**Proposition 12.** *For fixed $\mathbf{g}(\mathbf{x}) \neq \mathbf{0}$ and $u > 0$, we have*

(i) $\lim_{T\to\infty} C_{T,u}(\mathbf{g}(\mathbf{x}), c) - T \log m = \frac{1}{m} \sum_{j \neq c}(u + g_j(\mathbf{x}) - g_c(\mathbf{x}))$ *and*

$$\lim_{T\to\infty} \omega_j^c(\mathbf{x}) = \frac{1}{m} \quad \textit{for } j = 1, \ldots, m.$$

(ii) $\lim_{T\to 0} C_{T,u}(\mathbf{g}(\mathbf{x}), c) = H_u(\mathbf{g}(\mathbf{x}), c)$ *and*

$$\lim_{T\to 0} \omega_j^c(\mathbf{x}) = \begin{cases} 1 & j = \operatorname{argmax}_l \{g_l(\mathbf{x}) + 1 - \mathbb{I}_{\{l=c\}}\} \\ 0 & \textit{otherwise.} \end{cases}$$

(iii) $C_{T,u}(\mathbf{g}(\mathbf{x}), c)$ *is increasing in $T$.*

The proof is given in Appendix C.2. It is worth noting that Proposition 12-(ii) shows that at $T = 0$, $C_{T,1}(\mathbf{g}(\mathbf{x}), c)$ reduces to the multiclass hinge loss $\xi_c(\mathbf{g}(\mathbf{x}))$ of Crammer and Singer [5]. Additionally, when $u = 0$, we have

$$C_{T,0}\big(\mathbf{g}(\mathbf{x}), c\big) = T \log\left[1 + \sum_{j \neq c} \exp \frac{g_j(\mathbf{x}) - g_c(\mathbf{x})}{T}\right],$$

which was proposed by Zhang et al. [29]. When $T = 1$, it is the logit loss $\gamma_c(\mathbf{g}(\mathbf{x}))$ in (10). Thus, $C_{1,1}(\mathbf{g}(\mathbf{x}), c)$ bridges the hinge loss $\xi_c(\mathbf{g}(\mathbf{x}))$ and the logit loss $\gamma_c(\mathbf{g}(\mathbf{x}))$.

Consider that

$$\lim_{T\longrightarrow 0} C_{T,0}\big(\mathbf{g}(\mathbf{x}), c\big) = \max_j \big(g_j(\mathbf{x}) - g_c(\mathbf{x})\big).$$

This shows that $C_{T,0}(\mathbf{g}(\mathbf{x}), c)$ no longer converges to the majorizations of $\mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}}$ as $T \to 0$. However, as a special case of $u = 1$, we have $\lim_{T\to 0} C_{T,1}(\mathbf{g}(\mathbf{x}), c) = \xi_c(\mathbf{g}(\mathbf{x})) \geq \mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}}$; that is, $C_{T,1}(\mathbf{g}(\mathbf{x}), c)$ converges to the majorization of $\mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}}$. In fact, Proposition 12-(ii) implies that for an arbitrary $u > 0$, the limit of $C_{T,u}(\mathbf{g}(\mathbf{x}), c)$ at $T = 0$ is still the majorization of $\mathbb{I}_{\{\phi(\mathbf{x}) \neq c\}}$. We thus see an essential difference between $C_{T,1}(\mathbf{g}(\mathbf{x}), c)$ and $C_{T,0}(\mathbf{g}(\mathbf{x}), c)$, which are respectively the generalizations of the $\mathcal{C}$-loss and the logit loss.

For notational simplicity, here and later we denote $C_{T,1}(\mathbf{g}(\mathbf{x}), c)$ by $C_T(\mathbf{g}(\mathbf{x}), c)$. Throughout our analysis in this section, we assume that the maximizing argument $l = \operatorname{argmax}_j g_j(\mathbf{x})$ is unique. This implies that $g_l(\mathbf{x}) > g_j(\mathbf{x})$ for $j \neq l$. The following theorem shows that the $\mathcal{C}$-loss is Fisher-consistent.

**Table 1**
Summary of the Multicategory Loss Functions w.r.t. $(\mathbf{x}, c)$. Here “I”, “II” and “III” represent `the constrained comparison`, `pairwise comparison` and `maximum pairwise comparison` settings, respectively.

| | | |
|---|---|---|
| Hinge | $\sum_{j\neq c}(1+g_j(\mathbf{x}))_+$ [13] | I |
| | $\sum_{j\neq c}(1-g_c(\mathbf{x})+g_j(\mathbf{x}))_+$ [25] | II |
| | $\xi_c(\mathbf{g}(\mathbf{x}))=\max\{g_j(\mathbf{x})+1-\mathbb{I}_{\{j=c\}}\}-g_c(\mathbf{x})$ [5] | III |
| Coherence | $L_T(\mathbf{g}(\mathbf{x}),c)=T\sum_{j\neq c}\log[1+\exp(\frac{1+g_j(\mathbf{x})}{T})]$ | I |
| | $G_T(\mathbf{g}(\mathbf{x}),c)=T\sum_{j\neq c}\log[1+\exp(\frac{1+g_j(\mathbf{x})-g_c(\mathbf{x})}{T})]$ [see Eq. (9)] | II |
| | $C_{T,u}(\mathbf{g}(\mathbf{x}),c)=T\log[1+\sum_{j\neq c}\exp\frac{u+g_j(\mathbf{x})-g_c(\mathbf{x})}{T}]$ [see Eq. (14)] | III |
| Logit | $\gamma_c(\mathbf{g}(\mathbf{x}))=\log[1+\sum_{j\neq c}\exp(g_j(\mathbf{x})-g_c(\mathbf{x}))]$ [see Eq. (10)] | |

**Theorem 13.** *Assume that $P_c(\mathbf{x})>0$ for $c=1,\dots,m$. Consider the optimization problem:*

$$\operatorname*{argmax}_{\mathbf{g}(\mathbf{x})\in\mathcal{G}}\sum_{c=1}^{m} C_{T,u}\big(\mathbf{g}(\mathbf{x}),c\big)P_c(\mathbf{x}) \tag{15}$$

*for fixed $T>0$ and $u\geq 0$. Let $\hat{\mathbf{g}}(\mathbf{x})=(\hat{g}_1(\mathbf{x}),\dots,\hat{g}_m(\mathbf{x}))^T$ be the solution. Then $\hat{\mathbf{g}}(\mathbf{x})$ is unique. Moreover, if $P_i(\mathbf{x})<P_j(\mathbf{x})$, we have $\hat{g}_i(\mathbf{x})<\hat{g}_j(\mathbf{x})$. Furthermore, after obtaining $\hat{\mathbf{g}}(\mathbf{x})$, $P_c(\mathbf{x})$ is given by*

$$P_c(\mathbf{x})=\frac{\sum_{l=1}^{m}\exp\frac{u+\hat{g}_l(\mathbf{x})+\hat{g}_c(\mathbf{x})-u\mathbb{I}_{\{l=c\}}}{T}}{\sum_{j=1}^{m}\sum_{l=1}^{m}\exp\frac{u+\hat{g}_l(\mathbf{x})+\hat{g}_j(\mathbf{x})-u\mathbb{I}_{\{l=j\}}}{T}}. \tag{16}$$

In the case of $u=0$ and $T=1$, it follows from Theorem 13 that $P_c(\mathbf{x})=\frac{\exp(\hat{g}_c(\mathbf{x}))}{\sum_{j=1}^{m}\exp(\hat{g}_j(\mathbf{x}))}$. This is identical to the solution for logistic regression.

**Theorem 14.** *Let $\hat{\mathbf{g}}(\mathbf{x})=(\hat{g}_1(\mathbf{x}),\dots,\hat{g}_m(\mathbf{x}))^T$ be the solution of optimization problem (15) where $P_c(\mathbf{x})>0$ for $c=1,\dots,m$, and let $P_l(\mathbf{x})=\max_c P_c(\mathbf{x})$.*

(1) *If $P_l(\mathbf{x})>1/2$, then*

$$\lim_{T\to 0}\hat{g}_c(\mathbf{x})=\begin{cases} u(m-1)/m & \text{if } c=l,\\ -u/m & \text{otherwise}.\end{cases}$$

(2) *If $P_l(\mathbf{x})<1/2$, then*

$$\lim_{T\to 0}\hat{g}_c(\mathbf{x})=0 \quad \text{for } c=1,\dots,m.$$

The proofs of Theorems 13 and 14 are given in Appendix B.5. Theorem 14 shows a very important asymptotic property of the solution $\hat{g}_c(\mathbf{x})$. Especially when $u=1$, $\hat{g}_c(\mathbf{x})$ as $T\to 0$ converges to the solution of Problem (5) which is based on the multiclass hinge loss $\xi_c(\mathbf{g}(\mathbf{x}))$ of Crammer and Singer [5] (see Theorem 8).

**Remark 1.** We present three multicategory coherence functions $L_T(\mathbf{g}(\mathbf{x}),c)$, $G_T(\mathbf{g}(\mathbf{x}),c)$ and $C_T(\mathbf{g}(\mathbf{x}),c)$. They are respectively upper bounds of three multicategory hinge losses studied in Section 3.3, so they are majorizations of the 0–1 loss $\mathbb{I}_{\{\phi(\mathbf{x})\neq c\}}$. When $m=2$, these three losses become identical. Our theoretical analysis shows that their limits as the temperature approaches zero become the corresponding hinge losses, and the limits of the minimizers of their expected errors are the minimizers of the expected errors of the corresponding hinge losses (see Theorems 9, 10 and 14). We summarize the multicategory loss functions discussed in the paper in Table 1.

**Remark 2.** The coherence losses $L_T(\mathbf{g}(\mathbf{x}),c)$ and $C_T(\mathbf{g}(\mathbf{x}),c)$ can result in explicit expressions for the class conditional probabilities (see (8) and (16)). Thus, this can provide us with an approach for conditional class probability estimation in the multicategory SVMs of Lee et al. [13] and of Crammer and Singer [5]. Roughly speaking, one replaces the solutions of classification models based on the multicategory coherence losses with those of the corresponding multiclass SVMs in (8) and (16), respectively. Based on $G_T(\mathbf{g}(\mathbf{x}),c)$, however, there does not exist an explicit expression for the class probability similar to (8) or (16). In this case, the above approach for class probability estimation does not apply to the multiclass SVM model of Vapnik [25], Bredensteiner and Bennett [3], Weston and Watkins [26], Guermeur [12].

**Remark 3.** An advantage of $C_T(\mathbf{g}(\mathbf{x}),c)$ over $L_T(\mathbf{g}(\mathbf{x}),c)$ is in that it can make condition $\mathbf{g}(\mathbf{x})\in\mathcal{G}$ automatically satisfy in developing a classification method. Moreover, we see that the multiclass $\mathcal{C}$-loss $C_T(\mathbf{g}(\mathbf{x}),c)$ bridges the hinge loss and the logit loss. Thus, it is applicable to the construction of multiclass large margin classification methods. This motivates us to devise multiclass large margin classification methods based on $C_T(\mathbf{g}(\mathbf{x}), c)$.

## 5. Applications of the multiclass $\mathcal{C}$-loss in classification problems

In this section, we develop a multiclass large margin classifier and a multiclass boosting algorithm. Recall that we let $C_T(\mathbf{g}(\mathbf{x}), c)$ denote $C_{T,1}(\mathbf{g}(\mathbf{x}), c)$; that is, we always set $u = 1$ here and later.

### *5.1. The multiclass $\mathcal{C}$ learning*

Using the multiclass $\mathcal{C}$-loss $C_T(\mathbf{g}(\mathbf{x}), c)$, we now construct margin-based classifiers that we refer to as multiclass $\mathcal{C}$ learning (MCL). We first consider the linear case and then turn to the kernelized case. In the linear case, where $g_j(\mathbf{x}) = a_j + \mathbf{x}^T\mathbf{b}_j$, we pose the following optimization problem:

$$\min_{a,\mathbf{b}} \frac{1}{2}\sum_{j=1}^m \|\mathbf{b}_j\|^2 + \frac{\gamma}{n}\sum_{i=1}^n C_T\big(\mathbf{g}(\mathbf{x}_i), c_i\big)$$

$$\text{s.t.} \sum_{j=1}^m a_j\mathbf{1}_n + \mathbf{X}\sum_{j=1}^m \mathbf{b}_j = 0, \tag{17}$$

where $\gamma > 0$ is the regularization parameter, $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ is the $n\times p$ input data matrix, and $\mathbf{1}_n$ represents the $n\times 1$ of ones. Note that here we use the result from Liu and Shen [15] that the infinite constraint $\sum_{j=1}^m g_j(\mathbf{x})$ $\forall \mathbf{x} \in \mathcal{X}$ can be reduced to $\sum_{j=1}^m a_j\mathbf{1}_n + \mathbf{X}\sum_{j=1}^m \mathbf{b}_j = 0$, which is a function solely of the training data.

Given a reproducing kernel $K(\cdot,\cdot)$ from $\mathcal{X}\times\mathcal{X} \to \mathbb{R}$, we attempt to find a margin vector $(g_1(\mathbf{x}), \dots, g_m(\mathbf{x})) = (a_1 + h_1(\mathbf{x}), \dots, a_m + h_m(\mathbf{x})) \in \prod_{j=1}^m(\{1\} + \mathcal{H}_K)$, where $\mathcal{H}_K$ is a reproducing kernel Hilbert space. The solution of the following problem

$$\min_{\mathbf{g}(\mathbf{x})} \frac{1}{2}\sum_{j=1}^m \big\|h_j(\mathbf{x})\big\|^2_{\mathcal{H}_K} + \frac{\gamma}{n}\sum_{i=1}^n C_T\big(\mathbf{g}(\mathbf{x}_i), c_i\big) \tag{18}$$

under the constraints $\sum_{j=1}^m g_j(\mathbf{x}) = 0$ $\forall \mathbf{x} \in \mathcal{X}$ is

$$g_j(\mathbf{x}) = a_j + \sum_{i=1}^n \beta_{ji}K(\mathbf{x}_i, \mathbf{x}), \quad j = 1, \dots, m$$

with constraints $\sum_{j=1}^m g_j(\mathbf{x}_i) = 0$ for $i = 1, \dots, n$. This result follows readily from that of Lee et al. [13]. We see that kernel-based MCL solves the following optimization problem:

$$\min_{a,\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}_j^T\mathbf{K}\boldsymbol{\beta}_j + \frac{\gamma}{n}\sum_{i=1}^n C_T\big(\mathbf{g}(\mathbf{x}_i), c_i\big)$$

$$\text{s.t.} \sum_{j=1}^m a_j\mathbf{1}_n + \mathbf{K}\sum_{j=1}^m \boldsymbol{\beta}_j = 0, \tag{19}$$

where $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{jn})^T$ and $\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]$ is the $n\times n$ kernel matrix.

The minimization problem in (19) or (17) is a convex minimization problem and the objective function is differentiable; thus, the problem is readily solved. In particular, we make use of Newton-type methods to solve this problem. We further alternatively update $a_j$'s and $\boldsymbol{\beta}_j$'s. The details are given in Appendix D.

To end this subsection, we establish a connection of multiclass $\mathcal{C}$ learning with the multiclass SVM of Crammer and Singer [5], which is defined by

$$\min_{\mathbf{g}(\mathbf{x})} \frac{1}{2}\sum_{j=1}^m \big\|h_j(\mathbf{x})\big\|^2_{\mathcal{H}_K} + \frac{\gamma}{n}\sum_{i=1}^n \xi_{c_i}\big(\mathbf{g}(\mathbf{x}_i)\big) \tag{20}$$

under the constraints $\sum_{j=1}^m g_j(\mathbf{x}) = 0$ $\forall \mathbf{x} \in \mathcal{X}$. From Proposition 12, MCL reduces to the multiclass SVM of Crammer and Singer [5] as $T \to 0$. In fact, we have the following theorem (the proof is given in Appendix E).

**Theorem 15.** *Assume that $\gamma$ in Problems (20) and (18) are same. The minimizer of (18) approaches the minimizer of (20) as $T \to 0$.*

**Algorithm 1** GentleBoost.C($\{(\mathbf{x}_i, c_i)\}_{i=1}^{n} \subset \mathbb{R}^p \times \{1, \dots, m\}, T, H$).

```
1: Start with uniform weights $w_{ij} = 1/n$ for $i = 1, \dots, n$ and $j = 1, \dots, m$, and $\beta_j(\mathbf{x}) = 1/m$ and $g_j(\mathbf{x}) = 0$ for $j = 1, \dots, m$.
2: Repeat for $h = 1$ to $H$:
   (a) Repeat for $j = 1, \dots, m$:
       (i) Compute working responses and weights in the $j$th class,
           $$z_{ij} = T\frac{\mathbb{I}_{\{j=c_i\}} - \beta_j(\mathbf{x}_i)}{\beta_j(\mathbf{x}_i)(1 - \beta_j(\mathbf{x}_i))},$$
           $$w_{ij} = \beta_j(\mathbf{x}_i)\big(1 - \beta_j(\mathbf{x}_i)\big).$$
       (ii) Fit the regression function $g_j^{(h)}(\mathbf{x})$ by weighted least-squares of the working response $z_{ij}$ to $\mathbf{x}_i$ with weights $w_{ij}$ on the training data.
       (iii) Set $g_j(\mathbf{x}) \leftarrow g_j(\mathbf{x}) + g_j^{(h)}(\mathbf{x})$.
   (b) Set $g_j(\mathbf{x}) \leftarrow \frac{m-1}{m}[g_j(\mathbf{x}) - \frac{1}{m}\sum_{l=1}^{m} g_l(\mathbf{x})]$ for $j = 1, \dots, m$.
   (c) Compute $\beta_j(\mathbf{x}_i)$ for $j = 1, \dots, m$ as
       $$\beta_j(\mathbf{x}_i) = \begin{cases} \frac{\exp\frac{1+g_j(\mathbf{x}_i)-g_{c_i}(\mathbf{x}_i)}{T}}{1+\sum_{j\neq c_i}\exp\frac{1+g_j(\mathbf{x}_i)-g_{c_i}(\mathbf{x}_i)}{T}} & \text{if } j \neq c_i, \\ \frac{1}{1+\sum_{j\neq c_i}\exp\frac{1+g_j(\mathbf{x}_i)-g_{c_i}(\mathbf{x}_i)}{T}} & \text{if } j = c_i. \end{cases}$$
3: Output $\phi(\mathbf{x}) = \operatorname{argmax}_j g_j(\mathbf{x})$.
```

### 5.2. *The multiclass GentleBoost algorithm*

Like the negative multinomial log-likelihood function, when the multiclass $\mathcal{C}$-loss is used to devise multicategory discrete boosting algorithms, a closed-form solution no longer exists. We instead use the multiclass $\mathcal{C}$-loss to devise a genuine multicategory margin-based boosting algorithm. With a derivation similar (also see Appendix F for a brief derivation) to that in Friedman et al. [9], Zou et al. [33], Zhu et al. [32], our GentleBoost algorithm is shown in Algorithm 1.

## 6. Experimental evaluation

Our primary goal in this paper has been to provide statistical analysis of multicategory large margin classification methods based on hinge losses and coherence losses. However, we have also developed a multiclass $\mathcal{C}$ learning algorithm and a multiclass gentleBoost algorithm using the multiclass $\mathcal{C}$-loss. In this section, we conduct empirical analysis of these algorithms.

### 6.1. *Results of multiclass $\mathcal{C}$ learning*

We present the results of experiments evaluating multiclass $\mathcal{C}$ learning (MCL) and comparing it with the multiclass SVM [5], multiclass $\psi$-learning [15] and penalized logistic regression (PLR) [31]. All the algorithms were implemented in the linear setting.

Our first two experiments used the setup presented by Liu and Shen [15]. The first two datasets were generated from three bivariate $t$-distributions: $t((\sqrt{3}, 1)^T, \mathbf{I}_2)$, $t((-\sqrt{3}, 1)^T, \mathbf{I}_2)$ and $t((0, -2)^T, \mathbf{I}_2)$. Here $\mathbf{I}_2$ is the $2\times2$ identity matrix. In the first dataset, the degree of freedom (df) is equal to 1, while it is equal to 3 in the second dataset. All algorithms were trained using 150 samples and tested using an additional $10^6$ samples.

These authors found the multiclass SVM and multiclass $\psi$-learning to work best on these datasets and we have reported their results for these approaches in the first two columns of Table 2. This table displays the test errors, which were averaged over 100 randomly repeated simulations.

We implemented both MCL and PLR, using the same Newton-type method in both cases. The Newton iteration stops when the maximum iteration number (200) is reached or when the difference of successive loss values is less than 0.001. The initial values of $a_j$ and $\mathbf{b}_j$ are set to 0.

Adopting the procedure of Liu and Shen [15], our reported results were based on choosing the optimal value of the regularization parameter $\tau = \frac{2\gamma}{n}$ via a simple grid search on $[10^{-3}, 10^3]$. As shown in Table 2, $\psi$-learning has the lowest testing error for the first dataset, slightly outperforming MCL. MCL is best on the second dataset.

The third dataset can be obtained from Statlog (http://www.liacc.up.pt/ML/) and it consists of images of the letters “D,” “O” and “Q,” with 805, 753 and 783 cases respectively. 200 of the 2341 letters were randomly selected for training and the rest were retained for testing. The results are summarized in the third column of Table 2, where the test errors were averaged over 10 randomly repeated simulations. We see that MCL has the smallest test error, followed by $\psi$-learning and PLR.

Finally, we also performed experiments on text categorization using the WebKB dataset [6]. This dataset contains web pages gathered from computer science departments in several universities. The pages can be divided into seven categories. In the experiments, we used the four most populous categories, namely, student, faculty, course, and project, resulting in a total of 4192 pages. Based on information gain, 300 features were selected. We then randomly selected 70% of the data for training while the remaining 30% were used for testing. We repeated this procedure 30 times, and reported the final

**Table 2**
Test error rates and their standard deviations in parentheses (%). The results with MCL are based on $T = 1$ and $\tau = 2$, $T = 0.5$ and $\tau = 10$, $T = 0.2$ and $\tau = 10$, and $T = 1$ and $\tau = 10$ for the four datasets, respectively.

| data | SVM | $\psi$-L | MCL | PLR |
|---|---|---|---|---|
| t-data ($df = 1$) | 43.05 (±14.05) | **34.94** (±12.09) | 35.07 (±11.12) | 36.57 (±12.06) |
| t-data ($df = 3$) | 15.05 (±0.45) | 14.95 (±0.33) | **14.80** (±0.20) | 14.94 (±0.32) |
| letters | 8.16 (±0.75) | 7.72 (±0.82) | **7.42** (±0.29) | 7.54 (±0.40) |
| WebKB | N/A | N/A | **9.61** (±0.73) | 10.04 (± 0.66) |

**Table 3**
Summary of benchmark datasets.

| Dataset | # Train | # Test | # Features | # Classes |
|---|---|---|---|---|
| Vowel | 528 | 462 | 10 | 11 |
| Waveform | 300 | 4700 | 21 | 3 |
| Segmentation | 210 | 2100 | 19 | 7 |
| Optdigits | 3823 | 1797 | 64 | 10 |
| Pendigits | 7494 | 3498 | 16 | 10 |
| Satimage | 4435 | 2000 | 36 | 6 |

**Table 4**
Test error rates of our method and related methods (in %), and the results with GBoost.C are based on $T = 1$. The best result for each dataset is shown in bold.

| Dataset | CART | AdaBoost.MH | GD-MCBoost | MBoost.L | GBoost.E | GBoost.C |
|---|---|---|---|---|---|---|
| Vowel | 54.10 | 50.87 | 50.43 | 49.13 | 50.43 | **47.62** |
| Waveform | 31.60 | 18.22 | 17.45 | 17.23 | 17.62 | **16.53** |
| Segmentation | 9.80 | 5.29 | 4.43 | 4.10 | 4.52 | **4.05** |
| Optdigits | 16.60 | 5.18 | 3.78 | 3.28 | 5.12 | **3.17** |
| Pendigits | 8.32 | 5.86 | 3.60 | **3.12** | 3.95 | 3.14 |
| Satimage | 14.80 | 10.00 | 10.75 | 9.25 | 12.00 | **8.75** |

errors as an average over the 30 replicates. The results are shown in the final row of Table 2, where we have restricted the comparison to MCL and PLR. We see that MCL yields an improvement over PLR.

We also conducted a systematic study of the effect of the hyperparameters $\tau$ and $T$ on the letter dataset. We found that the results were relatively insensitive to particular values of these hyperparameters over an order of magnitude for $T$ and three orders of magnitude for $\tau$. There was a tradeoff; larger $\tau$ favors a smaller value of $T$.

### *6.2. Results of multiclass GentleBoost algorithm*

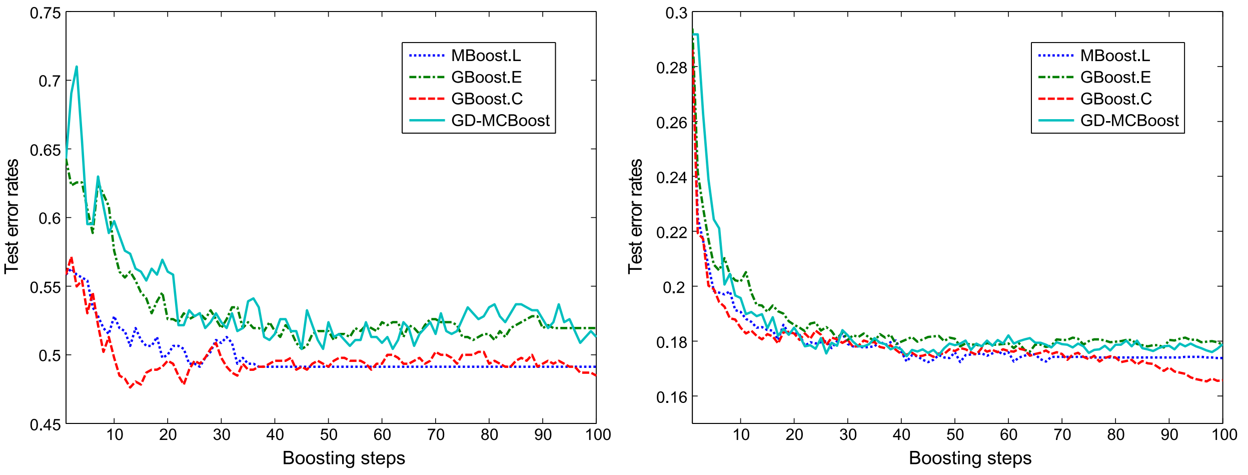
We also compare our multiclass gentleBoost algorithm (called GBoost.C) with some representative multicategory boosting algorithms, including AdaBoost.MH [22], multicategory LogitBoost (MBoost.L) [9], multicategory GentleBoost (GBoost.E) [33] and GD-MCBoost [21], on six publicly available datasets (Vowel, Waveform, Image Segmentation, Optdigits, Pendigits and Satimage) from the UCI Machine Learning Repository. Following the settings in Friedman et al. [9], Zou et al. [33], we use predefined training samples and test samples for these six datasets. Summary information for the datasets is given in Table 3. We use the code released by Saberian and Vasconcelos [21] to implement their GD-MCBoost algorithm.

Based on the experimental strategy in Zou et al. [33], eight-node regression trees are used as weak learners for all the boosting algorithms with the exception of AdaBoost.MH, which is based on eight-node classification trees. From the experiments, we observe that the performance of all the methods becomes stable after about 50 boosting steps. Hence, the number of boosting steps for all the methods is set to 100 ($H = 100$) in all the experiments. The test error rates (in %) of all the boosting algorithms are shown in Table 4, from which we can see that all the boosting methods achieve much better results than CART, and our method slightly outperforms the other boosting algorithms.

Among all the datasets tested, Vowel and Waveform are the most difficult for classification. The notably better performance of our method for these two datasets reveals its promising properties. Fig. 1 depicts the test error curves of MBoost.L, GBoost.E, GBoost.C and GD-MCBoost on these two datasets.

As we established in Section 3.1, GBoost.E does not implement a margin-based decision because the loss function used in this algorithm is not the majorization function of the 0–1 loss. Our experiments show that GD-MCBoost, MBoost.L and GBoost.C are comparable, and outperform GBoost.E. The results reported in Table 4 and Fig. 1 are based on the setting of $T = 1$. Recall that $\gamma_c(\mathbf{g}(\mathbf{x}))$ (see Eq. (10)) is the special case of $C_{T,0}(\mathbf{g}(\mathbf{x}), c)$ with $T = 1$, so the comparison of GBoost.C with MBoost.L is fair based on $T = 1$.

Proposition 12 shows that $C_T(\mathbf{g}(\mathbf{x}), c)$ $(= C_{T,1}(\mathbf{g}(\mathbf{x}), c))$ approaches $\max_j\{g_j(\mathbf{x}) + 1 - \mathbb{I}_{\{j=c\}}\} - g_c(\mathbf{x})$ as $T \to 0$. This encourages us to try to decrease $T$ gradually over the boosting steps. However, when $T$ gets very small, it can lead to numerical



(a) Vowel

(b) Waveform

**Fig. 1.** Test error rates versus boosting steps.

**Table 5**
Test error rates of our method (GBoost.C) with different values of $T$ (in %). The best and worst results for each dataset are shown in bold, and the average (ave) over these different values of $T$ and the corresponding standard deviation (stad) are also given for each dataset.

| Dataset | $T = 0.1$ | 0.2 | 0.4 | 0.8 | 1.6 | 3.2 | 6.4 | 12.8 | 25.6 | ave (±std) |
|---|---|---|---|---|---|---|---|---|---|---|
| Vowel | **50.43** | 49.57 | 49.57 | **45.45** | 48.48 | 47.62 | 49.13 | 48.70 | 46.75 | 48.41 (±1.56) |
| Waveform | 16.68 | **16.60** | 17.17 | 17.45 | 17.13 | 17.17 | **17.45** | 17.32 | 16.94 | 17.10 (±0.31) |
| Segmentation | 4.05 | 4.29 | **4.38** | 4.19 | 4.05 | **3.52** | 3.95 | 4.00 | 4.24 | 4.07 (±0.25) |
| Optdigits | **3.00** | 3.00 | 3.00 | 3.12 | **3.23** | 3.23 | 3.23 | 3.23 | 3.23 | 3.14 (±0.11) |
| Pendigits | 3.06 | 3.34 | **3.40** | 3.32 | **3.00** | 3.12 | 3.09 | 3.14 | 3.20 | 3.19 (±0.14) |
| Satimage | 7.90 | **7.85** | 8.70 | 8.85 | **9.10** | 9.10 | 8.85 | 8.75 | 8.95 | 8.67 (±0.47) |

problems and often makes the algorithm unstable. To observe the effect of $T$, we use the different values of $T$ to implement our boosting algorithm. The results are shown in Table 5. The experiments show that when $T$ takes a value in $[0.1, 20]$, our algorithm (GBoost.C) is always able to obtain promising performance. In other words, the performance of our algorithm is less sensitive to the value of $T$.

## 7. Conclusion

In this paper, we have studied a class of multicategory coherence loss functions as well as the relationship between the multicategory coherence and hinge losses. As majorization functions of the 0–1 loss, the multicategory coherence loss functions are Fisher-consistent, infinitely smooth and convex. Thus, it is appropriate for the design of margin-based boosting algorithms. In particular, we have devised a multiclass $C$ learning algorithm and a multiclass GentleBoost algorithm. While our main focus has been theoretical, we have also shown experimentally that our algorithms are effective.

## Acknowledgements

The authors would like to thank the three anonymous referees for their insightful comments on the original version of this paper. Wu-Jun Li is supported by the NSFC (No. 61100125), the 863 Program of China (No. 2012AA011003), and the Program for Changjiang Scholars and Innovative Research Team in University of China (IRT1158, PCSIRT). Z. Zhang is supported in part by the Natural Science Foundation of China (No. 61070239).

## Appendix A. The proof of Theorem 3 and Corollary 4

In our derivation, we just write $P_c$ and $g_c$ for $P_c(\mathbf{x})$ and $g_c(\mathbf{x})$ for notational simplicity. In order to prove the theorem, we first present the following definition and lemma, which can be found in Ortega and Rheinboldt [18].

**Definition 16.** A mapping $\mathbf{f}: D \subset \mathbb{R}^p \to \mathbb{R}^p$ is monotone on $D_0 \subset D$ if

$$\big(\mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v})\big)^T (\mathbf{u} - \mathbf{v}) \geq 0, \quad \forall \mathbf{u}, \mathbf{v} \in D_0;$$

and $\mathbf{f}$ is strictly monotone on $D_0$ if the above strict inequality holds whenever $\mathbf{u} \neq \mathbf{v}$.

**Lemma 17.** *Let $\mathbf{f} : D \subset \mathbb{R}^p \to \mathbb{R}^p$ be continuously differentiable on an open convex set $D_0 \subset D$. Then*

(a) *$\mathbf{f}(\mathbf{u})$ is monotone on $D_0$ if and only if $\mathbf{f}'(\mathbf{u})$ is positive semidefinite for all $\mathbf{u} \in D_0$.*
(b) *If $\mathbf{f}'(\mathbf{u})$ is positive definite for all $\mathbf{u} \in D_0$, then $f$ is strictly monotone on $D_0$.*
(c) *If $\mathbf{f}'(\mathbf{u})$ is conditionally positive definite for all $\mathbf{u} \in D_0$, then $f$ is strictly monotone on $\{\mathbf{u} | \sum_{j=1}^{p} u_j = 0\} \subset D_0$.*

### *A.1. The proof of Theorem 3*

To solve the constrained minimization problem in (1), we define the Lagrangian as follows.

$$L = \sum_{c=1}^{m} \psi_c(\mathbf{g}) P_c + \lambda \sum_{c=1}^{m} g_c.$$

Since the Hessian matrix of $L$ w.r.t. $\mathbf{g}$ is $\sum_c \mathbf{H}_c P_c$ where $\mathbf{H}_c = [\frac{\partial^2 \psi_c(\mathbf{g})}{\partial g_j \partial g_k}]$ is conditionally positive definite, the solution $\hat{g}_c$ exists and is unique. Moreover, we have $\nabla \psi_c(\mathbf{g}) = (\frac{\partial \psi_c(\mathbf{g})}{\partial g_1}, \ldots, \frac{\partial \psi_c(\mathbf{g})}{\partial g_m})^T$ is strictly monotone for $\mathbf{g} \in \mathcal{G}$.

The first partial derivative of $L$ w.r.t. $g_k$ is

$$\frac{\partial L}{\partial g_k} = \frac{\partial \psi_k}{\partial g_k} P_k + \sum_{c \neq k} \frac{\partial \psi_c}{\partial g_k} P_c + \lambda.$$

Based on the Karush–Kuhn–Tucker (KKT) conditions, we have

$$\frac{\partial \psi_k(\hat{\mathbf{g}})}{\partial g_k} P_k = -\sum_{c \neq k} \frac{\partial \psi_c(\hat{\mathbf{g}})}{\partial g_k} P_c - \lambda, \quad k = 1, \ldots, m.$$

Without loss of generality, we assume $P_1 > P_2$. Hence,

$$\left[\frac{\partial \psi_1(\hat{\mathbf{g}})}{\partial g_1} - \frac{\partial \psi_1(\hat{\mathbf{g}})}{\partial g_2}\right] P_1 - \left[\frac{\partial \psi_2(\hat{\mathbf{g}})}{\partial g_2} - \frac{\partial \psi_2(\hat{\mathbf{g}})}{\partial g_1}\right] P_2 = \sum_{c \neq 1,2} \left[\frac{\partial \psi_c(\hat{\mathbf{g}})}{\partial g_2} - \frac{\partial \psi_c(\hat{\mathbf{g}})}{\partial g_1}\right] P_c. \tag{21}$$

We now prove $\hat{g}_1 > \hat{g}_2$ by contradiction. Let us assume $\hat{g}_1 \leq \hat{g}_2$. On one hand, using the strict monotony of $\nabla \psi_c$ and the assumption of $\boldsymbol{\psi} \in \Psi$ yields

$$\begin{aligned}
&(\hat{g}_1 - \hat{g}_2)\left[\left(\frac{\partial \psi_1(\hat{\mathbf{g}})}{\partial g_1} - \frac{\partial \psi_1(\hat{\mathbf{g}})}{\partial g_2}\right) - \left(\frac{\partial \psi_2(\hat{\mathbf{g}})}{\partial g_2} - \frac{\partial \psi_2(\hat{\mathbf{g}})}{\partial g_1}\right)\right] \\
&\quad = (\hat{g}_1 - \hat{g}_2)\left[\left(\frac{\partial \psi_1(\hat{\mathbf{g}})}{\partial g_1} - \frac{\partial \psi_1(\hat{\mathbf{g}})}{\partial g_2}\right) - \left(\frac{\partial \psi_1(\hat{\mathbf{g}}^{12})}{\partial g_2} - \frac{\partial \psi_1(\hat{\mathbf{g}}^{12})}{\partial g_1}\right)\right] \\
&\quad = \left(\hat{\mathbf{g}} - \hat{\mathbf{g}}^{12}\right)^T \left(\nabla \psi_1(\hat{\mathbf{g}}) - \nabla \psi_1\left(\hat{\mathbf{g}}^{12}\right)\right) > 0
\end{aligned}$$

whenever $g_2 \neq g_1$. Here $\hat{\mathbf{g}}^{12} = (\hat{g}_2, \hat{g}_1, \hat{g}_3, \ldots, \hat{g}_m)^T$ and $\hat{\mathbf{g}} - \hat{\mathbf{g}}^{12} = (\hat{g}_1 - \hat{g}_2)(1, -1, 0, \ldots, 0)^T$. We thus have

$$0 > \frac{\partial \psi_2(\hat{\mathbf{g}})}{\partial g_2} - \frac{\partial \psi_2(\hat{\mathbf{g}})}{\partial g_1} \geq \frac{\partial \psi_1(\hat{\mathbf{g}})}{\partial g_1} - \frac{\partial \psi_1(\hat{\mathbf{g}})}{\partial g_2},$$

which implies that the right-hand side of Eq. (21) is negative. The above first inequality is based on the assumption of the theorem.

On the other hand, for $c \neq 1, 2$, using the strict monotony of $\nabla \psi_c$, we have

$$(\hat{g}_2 - \hat{g}_1)\left[\frac{\partial \psi_c(\hat{\mathbf{g}})}{\partial g_2} - \frac{\partial \psi_c(\hat{\mathbf{g}})}{\partial g_1} - \frac{\partial \psi_c(\hat{\mathbf{g}}^{12})}{\partial g_2} + \frac{\partial \psi_c(\hat{\mathbf{g}}^{12})}{\partial g_1}\right] = \left(\hat{\mathbf{g}} - \hat{\mathbf{g}}^{12}\right)^T \left(\nabla \psi_c(\hat{\mathbf{g}}) - \nabla \psi_c\left(\hat{\mathbf{g}}^{12}\right)\right) > 0$$

whenever $\hat{g}_1 \neq \hat{g}_2$. Furthermore, the symmetry of $\psi_c(\mathbf{g})$ when fixed $g_c$ implies that $\frac{\partial \psi_c(\hat{\mathbf{g}})}{\partial g_2} = \frac{\partial \psi_c(\hat{\mathbf{g}}^{12})}{\partial g_1}$ and $\frac{\partial \psi_c(\hat{\mathbf{g}})}{\partial g_1} = \frac{\partial \psi_c(\hat{\mathbf{g}}^{12})}{\partial g_2}$. Hence,

$$(\hat{g}_2 - \hat{g}_1)\left[\frac{\partial \psi_c(\hat{\mathbf{g}})}{\partial g_2} - \frac{\partial \psi_c(\hat{\mathbf{g}})}{\partial g_1}\right] > 0 \quad \text{whenever } \hat{g}_1 \neq \hat{g}_2,$$

which implies that the left-hand side of Eq. (21) is nonnegative. Thus, the assumption that $\hat{g}_1 \leq \hat{g}_2$ is impossible.

### *A.2. The proof of Corollary 4*

In order to prove Corollary 4, it suffices to prove the following lemma.

**Lemma 18.** *Let $\mathbf{u} = (u_1, \dots, u_m)^T \in \mathbb{R}^m$ and $\mathbf{u}^c = (u_1^c, \dots, u_m^c)^T$ with $u_j^c = u_j - u_c$. Then $\mathbf{B}_c = [\frac{\partial^2 f(\mathbf{u}^c)}{\partial u_j^c \partial u_k^k}]_{j,k \neq c}$ is an $(m-1)\times(m-1)$ positive definite matrix, if and only if $\mathbf{H}_c = [\frac{\partial \psi_c(\mathbf{u})}{\partial u_j u_k}]_{j,k=1}^m$ is an $m\times m$ conditionally positive definite matrix.*

**Proof.** Without loss of generality, we only consider the case of $c = m$. Clearly,

$$\frac{\partial \psi_m(\mathbf{u})}{\partial u_l} = \begin{cases} f_l'(\mathbf{u}^m) & l \neq m, \\ -\sum_{j=1}^{m-1} f_j'(\mathbf{u}^m) & l = m. \end{cases}$$

Subsequently,

$$\frac{\partial^2 \psi_m(\mathbf{u})}{\partial u_l^2} = \begin{cases} f_{ll}''(\mathbf{u}^m) & l \neq m, \\ \sum_{j=1}^{m-1} \sum_{i=1}^{m-1} f_{ji}''(\mathbf{u}^m) & l = m, \end{cases}$$

$\frac{\partial^2 \psi_m(\mathbf{u})}{\partial u_l \partial u_k} = f_{lk}''(\mathbf{u}^m)$ for $l \neq m$ and $k \neq m$, $\frac{\partial^2 \psi_m(\mathbf{u})}{\partial u_m \partial u_k} = -\sum_{j=1}^{m-1} f_{jk}''(\mathbf{u}^m)$ for $k \neq m$, and $\frac{\partial^2 \psi_m(\mathbf{u})}{\partial u_l \partial u_m} = -\sum_{j=1}^{m-1} f_{lj}''(\mathbf{u}^m)$ for $l \neq m$. We thus can express $\mathbf{H}_m$ as

$$\mathbf{H}_m = \begin{bmatrix} \mathbf{B}_m & -\mathbf{B}_m \mathbf{1}_{m-1} \\ -\mathbf{1}_{m-1}^T \mathbf{B}_m & \mathbf{1}_{m-1}^T \mathbf{B}_m \mathbf{1}_{m-1} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{m-1} \\ -\mathbf{1}_{m-1}^T \end{bmatrix} \mathbf{B}_m [\mathbf{I}_{m-1}, -\mathbf{1}_{m-1}].$$

Given any nonzero $\mathbf{z} \in \mathbb{R}^{m-1}$, we have

$$\begin{aligned} &\left[\mathbf{z}^T, -\mathbf{z}^T \mathbf{1}_{m-1}\right] \begin{bmatrix} \mathbf{B}_m & -\mathbf{B}_m \mathbf{1}_{m-1} \\ -\mathbf{1}_{m-1}^T \mathbf{B}_m & \mathbf{1}_{m-1}^T \mathbf{B}_m \mathbf{1}_{m-1} \end{bmatrix} [\mathbf{z} \quad -\mathbf{z}^T \mathbf{1}_{m-1}] \\ &= \mathbf{z}^T [\mathbf{I}_{m-1}, -\mathbf{1}_{m-1}] \begin{bmatrix} \mathbf{I}_{m-1} \\ -\mathbf{1}_{m-1}^T \end{bmatrix} \mathbf{B}_m [\mathbf{I}_{m-1}, -\mathbf{1}_{m-1}] \begin{bmatrix} \mathbf{I}_{m-1} \\ -\mathbf{1}_{m-1}^T \end{bmatrix} \mathbf{z} \\ &= \mathbf{z}^T \left(\mathbf{I}_{m-1} + \mathbf{1}_{m-1} \mathbf{1}_{m-1}^T\right) \mathbf{B}_m \left(\mathbf{I}_{m-1} + \mathbf{1}_{m-1} \mathbf{1}_{m-1}^T\right) \mathbf{z}, \end{aligned}$$

where $\mathbf{I}_m$ is the $m\times m$ identity matrix and $\mathbf{1}_m$ is the $m\times 1$ vector of ones. Consider that $\mathbf{I}_{m-1} + \mathbf{1}_{m-1}\mathbf{1}_{m-1}^T$ is positive definite. Thus, we obtain that $\mathbf{H}_m$ is conditionally positive definite if and only if $\mathbf{B}_m$ is positive definite. □

## Appendix B. Fisher consistency

### *B.1. The proof of Theorem 7*

Without loss of generality, we assume that $P_1 > 1/2 > P_2 \geq \cdots P_m > 0$.

Suppose that $P_1 > 1/2$. First, it is immediate to obtain $\hat{g}_1 \geq \hat{g}_2 \geq \cdots \geq \hat{g}_m$ from Theorem 5 of Zhang [28]. Now note that

$$\begin{aligned} L(\mathbf{g}) &= \sum_{c=1}^m \sum_{j \neq c} (1 - g_c + g_j)_+ P_c \\ &= (1 - g_1 + g_2)_+ P_1 + \sum_{j \neq 1,2} (1 - g_1 + g_j)_+ P_1 + \sum_{c \neq 1} (1 - g_c + g_1)_+ P_c + \sum_{c \neq 1} \sum_{j \neq 1,c} (1 - g_c + g_j)_+ P_c. \end{aligned}$$

We are given $\mathbf{u} = (u_1, \dots, u_m)^T \in \mathcal{G}$ such that $u_1 \geq u_2 \geq u_m$. Let $\rho \triangleq u_1 - u_2$. We define $f_1 = u_1 + \frac{m-1}{m}(1-\rho)$ and $f_j = u_j - \frac{1-\rho}{m}$ for $j = 2, \dots, m$. Clearly, $\sum_{j=1}^m f_j = 0$ and $f_1 = f_2 + 1$. We consider two cases.

In the first case where $0 \leq \rho \leq 1$, we have

$$\begin{aligned} L(\mathbf{u}) - L(\mathbf{f}) &= (1-\rho)P_1 + \sum_{j \neq 1,2} (1 - u_1 + u_j)_+ P_1 + \sum_{c \neq 1} (1 - u_c + u_1) P_c - \sum_{c \neq 1} (2 - \rho - u_c + u_1) P_c \\ &= (1-\rho)(2P_1 - 1) + \sum_{j \neq 1,2} (1 - u_1 + u_j)_+ P_1, \end{aligned}$$

which implies that $L(\mathbf{u}) - L(\mathbf{f}) > 0$ whenever $\rho \neq 1$.

In the second case where $\rho > 1$, we then have

$$L(\mathbf{u}) - L(\mathbf{f}) = \sum_{c \neq 1}(1 - u_c + u_1)P_c - \sum_{c \neq 1}(2 - \rho - u_c + u_1)P_c = (\rho - 1)(1 - P_1) > 0.$$

In summary, the minimizer $\hat{g}_j$ should satisfy $\hat{g}_1 = 1 + \hat{g}_2$ whenever $P_1 > \frac{1}{2}$.

Now suppose that $P_2 < \frac{1}{m}$. Assume that $\mathbf{u} = (u_1, \ldots, u_m)^T$ is a minimizer of $L$. We first show that $u_i - u_{i+1} \leq 1$. If there were an integer $k$ such that $1 \leq k \leq m - 1$ and $u_k - u_{k+1} > 1$, we would be able to give a new minimizer $\mathbf{v} = (v_1, \ldots, v_m)^T$ by letting $v_i = u_i + [1 - (u_k - u_{k+1})]$ for $i = 1, \ldots, k$ and $v_j = u_j$ for $j = k + 1, \ldots, m$. Then, for any pair $(i, j)$ where $i \in \{1, \ldots, k\}$ and $j \in \{k+1, \ldots, m\}$, we have the following four inequalities: $1 + v_j - v_i \leq 0$, $1 + u_j - u_i \leq 0$, $1 + v_i - v_j > 0$, and $1 + u_i - u_j > 0$. Therefore, we can get

$$\begin{aligned}
L(\mathbf{v}) - L(\mathbf{u}) &= \sum_{i=1}^{k} P_i \sum_{j=k+1}^{m} \big[(1 + v_j - v_i)_+ - (1 + u_j - u_i)_+\big] \\
&\quad + \sum_{j=k+1}^{m} P_j \sum_{i=1}^{k} \big[(1 + v_i - v_j)_+ - (1 + u_i - u_j)_+\big] \\
&= \sum_{j=k+1}^{m} P_j \sum_{i=1}^{k} (v_i - u_i) \\
&= \sum_{j=k+1}^{m} P_j \sum_{i=1}^{k} \big[1 - (u_k - u_{k+1})\big] < 0.
\end{aligned}$$

Second, we consider two cases. In the first case we assume $u_2 = u_3 = \cdots = u_m$. Letting $a \triangleq u_1 - u_2$, we can obtain $a \leq 1$ from the previous discussion. We thus have

$$\begin{aligned}
L(\mathbf{u}) &= P_1 \sum_{i=2}^{m}(1 + u_i - u_1)_+ + \sum_{i=2}^{m} P_i(1 + u_1 - u_i)_+ + \sum_{i=2}^{m} P_i \sum_{j \neq 1, i}(1 + u_j - u_i)_+ \\
&= (m - 1)(1 - a)P_1 + (1 + a)(1 - P_1) + (m - 2)(1 - P_1) \\
&= (1 - mP_1)a + (m - 1)P_1 + (1 - P_1) + (m - 2)(1 - P_1).
\end{aligned}$$

Noting that $P_1 > \frac{1}{m}$ (due to $P_m \leq \cdots \leq P_2 < \frac{1}{m}$), we obtain that $\mathbf{u}$ is a minimizer of $L$ if and only if $a = 1$. Consequently, we have $u_1 = 1 + u_2$.

In the second case we assume there exists a $k \in \{2, \ldots, m - 1\}$ such that $u_2 = \cdots = u_k$ and $u_k - u_{k+1} > 0$. In this case we let $a \triangleq u_1 - u_2$ and $b \triangleq u_k - u_{k+1}$. If $a$ were smaller than 1, we would be able to find a new minimizer $\mathbf{v}$ of $L$. Since $0 \leq a < 1$ and $0 < b \leq 1$, we have $\rho \triangleq \min\{b, 1 - a\} > 0$. Let $v_i = u_i - \rho$ for $i = 2, \cdots, k$ and $v_j = u_j$ for $j = 1, k + 1, \cdots, m$. Then for any pair $(i, j)$ where $i \in \{2, \cdots, k\}$ and $j \in \{1, k+1, \cdots, m\}$ we have the following three inequalities: $1 + v_i - v_j \geq 0$, $1 + u_i - u_j \geq 0$ and $(1 + v_j - v_i)_+ - (1 + u_j - u_i)_+ \leq (u_i - v_i)$. The third inequality follows from the convexity of the hinge function. Then, we have

$$\begin{aligned}
L(\mathbf{v}) - L(\mathbf{u}) &= \sum_{i=2}^{k} P_i \sum_{j \in \{1, k+1, \cdots, m\}} \big[(1 + v_j - v_i)_+ - (1 + u_j - u_i)_+\big] \\
&\quad + \sum_{j \in \{1, k+1, \cdots, m\}} P_j \sum_{i=2}^{k} \big[(1 + v_i - v_j)_+ - (1 + u_i - u_j)_+\big] \\
&\leq \sum_{i=2}^{k} P_i \sum_{j \in \{1, k+1, \cdots, m\}} (u_i - v_i) + \sum_{j \in \{1, k+1, \cdots, m\}} P_j \sum_{i=2}^{k} (v_i - u_i) \\
&= \rho \left[ (m - k + 1) \sum_{i=2}^{k} P_i - (k - 1)\left(1 - \sum_{i=2}^{k} P_i\right) \right] \\
&= \rho \left( m \sum_{i=2}^{k} P_i - k + 1 \right).
\end{aligned}$$

Since $P_2 < \frac{1}{m}$ and $P_2 \geq \cdots \geq P_m$, we obtain that $\sum_{i=2}^{k} P_i < \frac{k-1}{m}$. Hence, $L(\mathbf{v}) - L(\mathbf{u}) < 0$.

In summary, the minimizer $\mathbf{u}$ should satisfy $u_1 = u_2 + 1$. Recall that in the previous proof, we ignore the assumption that $\sum_{i=1}^m u_i = 0$. In fact, if $L$ attains its minimum at $\mathbf{u}$ with $\sum_{i=1}^m u_i = C$, we can obtain a new vector $\mathbf{u}'$ by letting $u_i' = u_i - \frac{C}{m}$. Then, we have $L(\mathbf{u}) = L(\mathbf{u}')$ and $\sum_{i=1}^m u_i' = 0$.

### B.2. The proof of Theorem 8

Without loss of generality, we assume that $P_l = \max_j(P_j(\mathbf{x}))$. This implies $g_l = \max_j(g_j(\mathbf{x}))$. Thus,

$$L = \sum_{c=1}^m \Big\{\max_j [g_j + 1 - \mathbb{I}_{\{j=c\}}] - g_c\Big\} P_c = \Big\{\max_j [g_j + 1 - \mathbb{I}_{\{j=l\}}] - g_l\Big\} P_l + \sum_{c \neq l} (1 + g_l - g_c) P_c.$$

**Case 1.** $g_l \geq \max_{j \neq l}\{1 + g_j\}$. In this case, we have

$$L = \sum_{c \neq l} (1 + g_l - g_c) P_c = 1 - P_l + \sum_{c \neq l} (g_l - g_c) P_c \geq 1 - P_l + \sum_{c \neq l} P_c$$

due to $g_l - g_c \geq 1$ for $c \neq l$. It is obvious that $L$ attains its minimum value

$$L_{\min} = 2(1 - P_l)$$

when $g_l - g_c = 1$ for $c \neq l$. Combining $\sum_{j=1}^m g_j = 0$, we have $g_l = (m-1)/m$ and $g_c = -1/m$ for $c \neq l$. Furthermore, we have $L_{\min} \geq 1$ if $P_l \leq 1/2$ and $L_{\min} < 1$ otherwise.

**Case 2.** $\max_j [g_j + 1 - \mathbb{I}_{\{j=l\}}] = 1 + g_k$ for $k \neq l$. In this case, we have $0 \leq g_l - g_k \leq 1$ and $g_k - g_j \geq 0$ for $j \neq l$. Note that

$$\begin{aligned} L &= (1 + g_k - g_l) P_l + \sum_{c \neq l} (1 + g_l - g_c) P_c \\ &= (g_l - g_k)\Big(\sum_{c \neq l} P_c - P_l\Big) + \sum_{c \neq l} (g_k - g_c) P_c + 1 \\ &= (g_l - g_k)(1 - 2P_l) + \sum_{c \neq l} (g_k - g_c) P_c + 1. \end{aligned}$$

If $P_l < 1/2$, then $L \geq 1$. Especially, $L$ attains the minimum value 1 when $g_l - g_k = 0$ and $g_k - g_c = 0$ for $c \neq l$. That is, $g_c = 0$ for $c = 1, \ldots, m$.

If $P_l = 1/2$, $L$ attains the minimum value 1 whenever the $g_c$ satisfy that $g_l - g_c \leq 1$ and $g_k - g_c = 0$ for $c \neq l$.

If $P_l > 1/2$, then $L \geq 2(1 - P_l)$. Further, $L$ attains the minimum value $2(1 - P_l)$ when $g_l - g_k = 1$ and $g_k - g_c = 0$ for $c \neq l$; that is, $g_l = (m-1)/m$ and $g_c = -1/m$ for $c \neq l$.

### B.3. The proof of Theorem 9

Consider the following Lagrangian function:

$$\begin{aligned} L &= \sum_{c=1}^m \sum_{j \neq c} T \log\big[1 + \exp\big((1 + g_j)/T\big)\big] P_c - \lambda \sum_{c=1}^m g_c \\ &= \sum_{c=1}^m T \log\big[1 + \exp\big((1 + g_c)/T\big)\big](1 - P_c) - \lambda \sum_{c=1}^m g_c, \end{aligned}$$

where $\lambda$ is the Lagrange multiplier. The first-order derivatives of $L$ w.r.t. the $g_c$ are

$$\frac{\partial L}{\partial g_c} = \frac{\exp((1 + g_c)/T)}{1 + \exp((1 + g_c)/T)} (1 - P_c) - \lambda.$$

The Hessian matrix $[\frac{\partial^2 L}{\partial g_l \partial g_j}] = \mathrm{diag}(\frac{\partial^2 L}{\partial g_1^2}, \ldots, \frac{\partial^2 L}{\partial g_m^2})$ where

$$\frac{\partial^2 L}{\partial g_c^2} = \frac{T(1 - P_c)\exp((1 + g_c)/T)}{[1 + \exp((1 + g_c)/T)]^2}$$

is positive definite and the minimizer $\hat{g}_c$ of the optimization problem (7) exists and is unique. This minimizer is obtained as the solution of $\frac{\partial L}{\partial g_c} = 0$ for $c = 1, \ldots, m$:

$$\hat{g}_c = T \log \frac{\lambda}{1 - P_c - \lambda} - 1.$$

Since $\frac{\lambda}{1-P_c-\lambda} > 0$, we have $0 < \lambda < 1 - P_c$ for $c = 1, \ldots, m$. We thus have $\hat{g}_l > \hat{g}_j$ if and only if $P_l > P_j$. Moreover, we obtain (8).

Let $l = \operatorname{argmax}_c P_c$. It then follows from $\sum_{c=1}^m \hat{g}_c = 0$ that $\hat{g}_l > 0$, and hence,

$$\frac{1 - P_l}{1 + \exp(-1/T)} < \lambda < 1 - P_l.$$

This implies $\lim_{T\to 0} \lambda = 1 - P_l$. As a result, we have $\lim_{T\to 0} \hat{g}_c = -1$ for $c \neq l$ and $\lim_{T\to 0} \hat{g}_l = m - 1$ due to $\sum_{c=1}^m \hat{g}_c = 0$.

### B.4. The proof of Theorem 10

We prove the theorem according to Corollary 4 where $\psi_c(\mathbf{g}) = f(\mathbf{g}^c) = G_T(\mathbf{g}(\mathbf{x}), c)$. It is directly calculated that for $j \neq c$,

$$f'_j(\mathbf{g}^c) = \frac{\exp \frac{1+g_j-g_c}{T}}{1 + \exp \frac{1+g_j-g_c}{T}} \triangleq \beta_j^c,$$

$f''_{jj}(\mathbf{g}^c) = \beta_j^c(1 - \beta_j^c)$ and $f''_{jk}(\mathbf{g}^c) = 0$ for $k \neq j, c$. Thus, the Hessian matrix $\mathbf{B}_c = [f''_{jk}(\mathbf{g}^c)]_{j,k\neq c}$ is positive definite. As a result, Corollary 4 shows that the minimizer $\hat{\mathbf{g}}$ exists and is unique.

Additionally, it is always satisfied that $\frac{\partial \psi_c(\mathbf{g})}{\partial g_c} = -\sum_{j\neq c} f'_j(\mathbf{g}^c) < 0$ for $j \neq c$. Thus, we obtain that $P_l > P_k$ implies $\hat{g}_l > \hat{g}_k$ from Corollary 4.

Recall that the minimizer $\hat{g}_c$ satisfies the condition of

$$P_c \sum_{j\neq c} \beta_j^c = \sum_{j\neq c} \beta_c^j P_j$$

where $\beta_c^j$ is defined via the $\hat{g}_c$ and we still denote them by the $\beta_c^j$ for simplicity. By the implicit function theorem, obviously, $\hat{g}_c$ is a continuous function of $T$ on $(0, \infty)$. Thus, its limit at $T = 0$ is bounded due to $\sum_{c=1}^m \hat{g}_c = 0$ and the boundedness of the $\hat{g}_c$. Without loss of generality, we assume that $P_1 > P_2 \geq \cdots \geq P_m$. In this case, we always have $\lim_{T\to 0} \hat{g}_1 \geq \lim_{T\to 0} \hat{g}_2 \geq \cdots \geq \lim_{T\to 0} \hat{g}_m$,

$$\lim_{T\to 0} \beta_1^j = \lim_{T\to 0} \frac{\exp \frac{1+\hat{g}_1-\hat{g}_j}{T}}{1 + \exp \frac{1+\hat{g}_1-\hat{g}_j}{T}} = 1, \quad \text{and} \quad \lim_{T\to 0} \beta_2^j = \lim_{T\to 0} \frac{\exp \frac{1+\hat{g}_2-\hat{g}_j}{T}}{1 + \exp \frac{1+\hat{g}_2-\hat{g}_j}{T}} = 1 \quad \text{for } j \neq 1.$$

If $\lim_{T\to 0} \hat{g}_1 > \lim_{T\to 0} \hat{g}_2 + 1$ had been satisfied, we would obtain

$$0 = \lim_{T\to 0} P_1 \sum_{j\neq 1} \beta_j^1 = \lim_{T\to 0} \sum_{j\neq 1} \beta_1^j P_j = 1 - P_1.$$

On the other hand, if $\lim_{T\to 0} \hat{g}_1 < \lim_{T\to 0} \hat{g}_2 + 1$ had been satisfied, we would obtain

$$P_1 = \lim_{T\to 0} \sum_{j\neq 1} \beta_1^j P_j \Big/ \sum_{j\neq 1} \beta_j^1 = (1 - P_1) \Big/ \left(1 + \sum_{j\neq 1,2} \lim_{T\to 0} \beta_j^1\right) \leq 1 - P_1 \quad \text{or}$$

$$P_2 = \lim_{T\to 0} \frac{P_2 + P_1\beta_2^1 + \sum_{j\neq 1,2} \beta_2^j P_j}{1 + \sum_{j\neq 2} \beta_j^2} = \frac{1}{1 + \lim_{T\to 0} \sum_{j\neq 2} \beta_j^2} \geq \frac{1}{m}.$$

Therefore, we obtain that $\lim_{T\to 0} \hat{g}_1 = \lim_{T\to 0} \hat{g}_2 + 1$ whenever $P_1 > \frac{1}{2}$ or $P_2 < \frac{1}{m}$.

### B.5. The proof of Theorems 13 and 14

We also prove the theorem in terms of Corollary 4. In the current case we have $\psi_c(\mathbf{g}) = f(\mathbf{g}^c) = C_T(\mathbf{g}(\mathbf{x}), c)$. We take computations for $j \neq c$ as

$$f'_j(\mathbf{g}^c) = \frac{\exp \frac{u+g_j-g_c}{T}}{1 + \sum_{l\neq c} \exp \frac{u+g_l-g_c}{T}} \triangleq \beta_j^c,$$

$f''_{jj}(\mathbf{g}^c) = \beta_j^c(1 - \beta_j^c)/T$ and $f''_{jk}(\mathbf{g}^c) = \beta_j^c \beta_j^k / T$ for $k \neq l, c$.

We denote $\mathbf{\Delta}_m = \operatorname{diag}(\beta_1^m, \ldots, \beta_{m-1}^m)$ and $\boldsymbol{\beta}_m = (\beta_1^c, \ldots, \beta_{m-1}^m)^T$. Then the Hessian matrix $\mathbf{B}_m$ is

$$\mathbf{B}_m = \frac{1}{T}\big(\boldsymbol{\Delta}_m - \boldsymbol{\beta}_m \boldsymbol{\beta}_m^T\big).$$

Since $\sum_{k \neq j,m} \beta_j^k \beta_k^m < \beta_j^m (1 - \beta_j^m)$, the matrix $\mathbf{B}_m$ is strictly diagonally dominant. Thus, $\mathbf{B}_m$ is positive definite. It then follows from Corollary 4 that the minimizer $\hat{\mathbf{g}}$ exists and is unique. Noting that $\frac{\partial \psi_c(\mathbf{g})}{\partial g_c} = -\sum_{j \neq c} \beta_j^c < 0$, we further obtain that $P_l > P_k$ implies $\hat{g}_l > \hat{g}_k$.

Since $\hat{\mathbf{g}}$ is the solution of the optimization problem in question, it should satisfy the first-order condition:

$$\frac{P_c \sum_{l=1}^m \exp \frac{u + \hat{g}_l - \hat{g}_c}{T}}{1 + \sum_{l \neq c} \exp \frac{u + \hat{g}_l - \hat{g}_c}{T}} = \sum_{j=1}^m \frac{P_j \times \exp \frac{u + \hat{g}_c - \hat{g}_j}{T}}{1 + \sum_{l \neq j} \exp \frac{u + \hat{g}_l - \hat{g}_j}{T}},$$

from which we get

$$\frac{P_c}{P_k} = \frac{\exp(\hat{g}_c/T)}{\exp(\hat{g}_k/T)} \frac{\exp(\hat{g}_c/T) + \sum_{l \neq c} \exp((u + \hat{g}_l)/T)}{\exp(\hat{g}_k/T) + \sum_{l \neq k} \exp((u + \hat{g}_l)/T)}. \tag{22}$$

From (22) and using the fact that $\sum_{c=1}^m P_c = 1$, we have (16).

We now consider the proof of Theorem 14. First, it is clear that $\hat{g}_c$ is continuous in $T$ on $(0, \infty)$, so its limit at $T = 0$ exists ($\infty$ allowed). For notational simplicity, we just use the $g_c$ instead of the $\hat{g}_c(\mathbf{x})$. Second, the above proof shows that $\lambda = 0$. And $\frac{\partial L}{\partial g_c} = 0$ yields $P_c = \sum_{j=1}^m \beta_c^j P_j$. Namely, for $c = 1, \dots, m$,

$$P_c = \frac{1}{1 + \sum_{i \neq c} \exp((u + g_i - g_c)/T)} P_c + \sum_{j \neq c} \frac{\exp((u + g_c - g_j)/T)}{1 + \sum_{i \neq j} \exp((u + g_i - g_j)/T)} P_j.$$

Let $l = \operatorname{argmax}_c P_c$ and $k = \operatorname{argmin}_c P_c$. We thus have that $\lim_{T \to 0} g_k \leq \lim_{T \to 0} g_j \leq \lim_{T \to 0} g_l$ for $j \neq k, l$. Note that

$$\begin{aligned} P_k = {} & \frac{P_k}{1 + \sum_{i \neq k} \exp(\frac{u + g_i - g_k}{T})} + \frac{P_l}{\exp(\frac{g_l - u - g_k}{T}) + \sum_{i \neq l} \exp(\frac{g_i - g_k}{T})} \\ & + \sum_{j \neq k,l} \frac{P_j}{\exp(\frac{g_j - u - g_k}{T}) + \sum_{i \neq j} \exp(\frac{g_i - g_k}{T})}. \end{aligned}$$

The first term of the right-hand side of the above equation approaches 0 at $T \to 0$ due to $\lim_{T \to 0} \frac{u + g_i - g_k}{T} = +\infty$ for $u > 0$. If there were $i \neq k, l$ such that $\lim_{T \to 0}(g_i - g_k) > 0$, we would have that the right-hand side of the above equation approaches 0 at $T \to 0$. This implies that $\lim_{T \to 0}(g_i - g_k) = 0$ and $\lim_{T \to 0}(g_i - g_l) \leq 0$ for any $i \neq l$.

On the other hand, take

$$P_l = \lim_{T \to 0} \frac{P_l}{1 + \sum_{i \neq l} \exp(\frac{u + g_i - g_l}{T})} + \lim_{T \to 0} \sum_{j \neq l} \frac{P_j}{\exp(\frac{g_j - u - g_l}{T}) + 1 + \sum_{i \neq j,l} \exp(\frac{g_i - g_l}{T})}.$$

We are able to show that $\lim_{T \to 0}(u + g_i - g_l) < 0$ cannot be satisfied, otherwise the first term is current is $P_l$ and the second term is $1 - P_l$.

**Case 1.** $P_l > 1/2$. We can also obtain that $\lim_{T \to 0}(u + g_i - g_l) > 0$ for any $i \neq l$ cannot be satisfied, because the first term of the right-hand side of the above equation is 0 and the second term is always less than 1/2 otherwise. Thus, we have $\lim_{T \to 0}(u + g_i - g_l) = 0$ for any $i \neq l$. As a result, $\lim_{T \to 0} g_l = u(m-1)/m$ and $\lim_{T \to 0} g_l i = -u/m$ for $i \neq l$ due to $\lim_{T \to 0} \sum_{i=1}^m g_i = 0$.

**Case 2.** $P_l < 1/2$. In this case, we always have $\lim_{T \to 0}(g_i - g_l) = 0$. Otherwise, the second term is $1 - P_l$ which is greater than 1/2.

## Appendix C. The properties of the multiclass $\mathcal{C}$-loss

We first prove that $C_{T,u}(\mathbf{g}(\mathbf{x}), c)$ is convex. From Appendix B.5, we can obtain the Hessian matrix of $C_{T,u}(\mathbf{g}(\mathbf{x}), c)$ w.r.t. $\mathbf{g}(\mathbf{x})$. That is,

$$\mathbf{H}_c = \frac{\partial^2 C_{T,u}(\mathbf{g}(\mathbf{x}), c)}{\partial \mathbf{g} \partial \mathbf{g}^T} = \frac{1}{T}\big(\boldsymbol{\Delta}_c - \boldsymbol{\beta}_c \boldsymbol{\beta}_c^T\big),$$

where $\boldsymbol{\Delta}_c = \operatorname{diag}(\beta_1^c, \dots, \beta_m^c)$ and $\boldsymbol{\beta}_c = (\beta_1^c, \dots, \beta_m^c)^T$. We also have from Appendix B.5 that $\mathbf{H}_c$ is positive semidefinite (in fact, it is conditionally positive definite). Thus, $C_{T,u}(\mathbf{g}(\mathbf{x}), c)$ is convex.

### *C.1. The proof of Proposition 11*

Noting that

$$\frac{\prod_{j\neq c}(1+\exp\frac{u+g_j(\mathbf{x})-g_c(\mathbf{x})}{T})}{1+\sum_{j\neq c}\exp\frac{u+g_j(\mathbf{x})-g_c(\mathbf{x})}{T}} > 1,$$

we have $C_{T,1}(\mathbf{g}(\mathbf{x}),c) < G_T(\mathbf{g}(\mathbf{x}),c)$.

Now assume that $l = \operatorname{argmax}_j\{g_j(\mathbf{x}) + u - u\mathbb{I}_{\{j=c\}}\}$. Then

$$T\log m + H_u\big(\mathbf{g}(\mathbf{x}),c\big) - C_{T,u}\big(\mathbf{g}(\mathbf{x}),c\big) = T\log\frac{m\exp\frac{u+g_l(\mathbf{x})-g_c(\mathbf{x})-u\mathbb{I}_{\{l=c\}}}{T}}{1+\sum_{j\neq c}\exp\frac{u+g_j(\mathbf{x})-g_c(\mathbf{x})}{T}} \geq 0.$$

### *C.2. The proof of Proposition 12*

First, it is easily obtained that $\lim_{T\to\infty}\omega_j^c(\mathbf{x}) = \frac{1}{m}$. Second, consider that

$$\begin{aligned}
\lim_{T\to\infty} C_{T,u}\big(\mathbf{g}(\mathbf{x}),c\big) - T\log m &= \lim_{T\to\infty}\frac{\log\frac{1+\sum_{j\neq c}\exp\frac{u+g_j(\mathbf{x})-g_c(\mathbf{x})}{T}}{m}}{\frac{1}{T}}\\
&= \lim_{\alpha\to 0}\frac{\log\frac{1+\sum_{j\neq c}\exp\alpha(u+g_j(\mathbf{x})-g_c(\mathbf{x}))}{m}}{\alpha}\\
&= \lim_{\alpha\to 0}\frac{\frac{1}{m}\sum_{j\neq c}[u+g_j(\mathbf{x})-g_c(\mathbf{x})]\exp[\alpha(u+g_j(\mathbf{x})-g_c(\mathbf{x}))]}{\frac{1+\sum_{j\neq c}\exp\alpha(u+g_j(\mathbf{x})-g_c(\mathbf{x}))}{m}}\\
&= \frac{1}{m}\sum_{j\neq c}\big[u+g_j(\mathbf{x})-g_c(\mathbf{x})\big].
\end{aligned}$$

It immediately follows from $H_u(\mathbf{g}(\mathbf{x}),c) \leq C_{T,u}(\mathbf{g}(\mathbf{x}),c) \leq H_u(\mathbf{g}(\mathbf{x}),c) + T\log(m)$ that $\lim_{T\to 0} C_{T,u}(\mathbf{g}(\mathbf{x}),c) = H_u(\mathbf{g}(\mathbf{x}),c)$.

Third, the derivative of $C_{T,u}(\mathbf{g}(\mathbf{x}),c)$ w.r.t. $T$ is given by

$$\begin{aligned}
\frac{\partial C_{T,u}}{\partial T} &= \ln\left[1+\sum_{j\neq c}\exp\frac{u+g_j(\mathbf{x})-g_c(\mathbf{x})}{T}\right] - \frac{\sum_{j\neq c}\exp\frac{u+g_j(\mathbf{x})-g_c(\mathbf{x})}{T}\frac{u+g_j(\mathbf{x})-g_c(\mathbf{x})}{T}}{1+\sum_{j\neq c}\exp\frac{u+g_j(\mathbf{x})-g_c(\mathbf{x})}{T}}\\
&= \max_j\frac{u+g_j(\mathbf{x})-g_c(\mathbf{x})-u\mathbb{I}_{\{j=c\}}}{T} - \frac{\sum_{j=1}^m\exp\frac{u+g_j(\mathbf{x})-g_c(\mathbf{x})-u\mathbb{I}_{\{j=c\}}}{T}\frac{u+g_j(\mathbf{x})-g_c(\mathbf{x})-u\mathbb{I}_{\{j=c\}}}{T}}{\sum_{j=1}^m\exp\frac{u+g_j(\mathbf{x})-g_c(\mathbf{x})-u\mathbb{I}_{\{j=c\}}}{T}}\\
&\geq 0.
\end{aligned}$$

Thus, $C_{T,u}(\mathbf{g}(\mathbf{x}),c)$ is an increasing function of $T$.

## Appendix D. The learning algorithm for MCL

For simplicity, we only consider the learning algorithm based on (17). Let

$$L = \frac{1}{2}\sum_{j=1}^m\|\mathbf{b}_j\|^2 + \frac{\gamma}{n}\sum_{i=1}^n C_T\big(\mathbf{g}(\mathbf{x}_i),c_i\big) + \sum_{i=1}^n\lambda_i\left(\sum_{j=1}^m a_j + \mathbf{b}_j^T\mathbf{x}_i\right),$$

where $\lambda_i$'s are Lagrangian multipliers, and calculate

$$\frac{\partial L}{\partial\mathbf{b}_j} = \mathbf{b}_j + \frac{\gamma}{n}\sum_{i=1}^n(w_{ij}-e_{ij})\mathbf{x}_i + \sum_{i=1}^n\lambda_i\mathbf{x}_i,$$

$$\frac{\partial L}{\partial a_j} = \frac{\gamma}{n}\sum_{i=1}^n(w_{ij}-e_{ij}) + \sum_{i=1}^n\lambda_i,$$

where $w_{ij} = w_j^{c_i}(\mathbf{x}_i)$ is defined in (13), and $e_{ij} = 1$ if $j = c_i$ and $e_{ij} = 0$ otherwise. It follows from $\sum_{j=1}^m \frac{\partial L}{\partial \mathbf{b}_j} = 0$ and $\sum_{j=1}^m \frac{\partial L}{\partial a_j} = 0$ that

$$\frac{\partial L}{\partial \mathbf{b}_j} = (\mathbf{b}_j - \bar{\mathbf{b}}) + \frac{\gamma}{n}\sum_{i=1}^n (w_{ij} - e_{ij})\mathbf{x}_i,$$
$$\frac{\partial L}{\partial a_j} = \frac{\gamma}{n}\sum_{i=1}^n (w_{ij} - e_{ij}),$$

where $\bar{\mathbf{b}} = \frac{1}{m}\sum_{j=1}^m \mathbf{b}_j$. Thus, the Lagrangian multipliers $\lambda_i$ are automatically eliminated. Denoting $\mathbf{e}_i = (e_{i1}, \dots, e_{im})^T$, $\mathbf{w}_i = (w_{i1}, \dots, w_{im})^T$, $\mathbf{a} = (a_1, \dots, a_m)^T$, $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_m]$ and $\mathrm{vec}(\mathbf{B})^T = (\mathbf{b}_1^T, \dots, \mathbf{b}_m^T)$, we have

$$\frac{\partial L}{\partial \mathbf{a}} = \frac{\gamma}{n}\sum_{i=1}^n (\mathbf{w}_i - \mathbf{e}_i) = \mathbf{0}, \tag{23}$$

$$\frac{\partial L}{\partial\, \mathrm{vec}(\mathbf{B})} = (\mathbf{I}_p \otimes \mathbf{C}_m)\, \mathrm{vec}(\mathbf{B}) + \frac{\gamma}{n}\sum_{i=1}^n (\mathbf{w}_i - \mathbf{e}_i) \otimes \mathbf{x}_i = \mathbf{0}, \tag{24}$$

where $\mathbf{A} \otimes \mathbf{B}$ denotes the Kronecker product of matrices $\mathbf{A}$ and $\mathbf{B}$, and $\mathbf{C}_m = \mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^T$ is the $m \times m$ centering matrix. We now use the Newton–Raphson method to alternatively solve the nonlinear equation systems in (23) and (24). Since the Hessian matrices are positive semidefinite, the method converges. Considering that the Newton–Raphson method requires inverting the Hessian matrix in each iteration, we employ a quadratic lower bound algorithm [2]. In particular,

$$\frac{\partial^2 L}{\partial\, \mathrm{vec}(\mathbf{B})\partial\, \mathrm{vec}(\mathbf{B})^T} = \mathbf{C}_m \otimes \mathbf{I}_p + \frac{\gamma}{n}\sum_{i=1}^n \left(\mathrm{diag}(\mathbf{w}_i) - \mathbf{w}_i\mathbf{w}_i^T\right) \otimes \mathbf{x}_i\mathbf{x}_i^T$$
$$\preceq \mathbf{C}_m \otimes \mathbf{I}_p + \frac{\gamma}{2n}\mathbf{C}_m \otimes \mathbf{X}^T\mathbf{X} = \mathbf{C}_m \otimes \left(\mathbf{I}_p + \frac{\gamma}{2n}\mathbf{X}^T\mathbf{X}\right),$$

where $\mathbf{A} \preceq \mathbf{M}$ means $\mathbf{A} - \mathbf{M}$ is positive semidefinite and we use the fact that $(\mathrm{diag}(\mathbf{w}_i) - \mathbf{w}_i\mathbf{w}_i^T) \preceq \frac{1}{2}\mathbf{C}_m$. We use the pseudoinverse of $\mathbf{C}_m$ (which is itself), and thus we need to invert $\mathbf{I}_p + \frac{\gamma}{2n}\mathbf{X}^T\mathbf{X}$ only once.

## Appendix E. The proof of Theorem 15

Consider that the first-order derivative of $C_{T,1}(\mathbf{g}(\mathbf{x}), c)$ w.r.t. $g_j(\mathbf{x})$ is

$$\frac{\partial C_{T,1}(\mathbf{g}, c)}{\partial g_j} = (\beta_j - \mathbb{I}_{\{j=c\}})$$

where

$$\beta_j = \begin{cases} \dfrac{\exp\frac{1+g_j-g_c}{T}}{1+\sum_{j\neq c}\exp\frac{1+g_j-g_c}{T}} & \text{if } j \neq c, \\ \dfrac{1}{1+\sum_{j\neq c}\exp\frac{1+g_j-g_c}{T}} & \text{if } j = c. \end{cases}$$

Given a $\bar{\mathbf{g}} = (\bar{g}_1, \dots, \bar{g}_m)^T \in \mathcal{G}$, we denote $\mathcal{J} = \{j \neq c : (1 + \bar{g}_j - \bar{g}_c) = \xi_c(\bar{\mathbf{g}}) \triangleq \max_l(1 + \bar{g}_l - \bar{g}_c - \mathbb{I}_{\{l=c\}})\}$ and $k = |\mathcal{J}|$. It is directly obtained that

$$\lim_{T\to 0}\frac{\partial C_{T,1}(\bar{\mathbf{g}}, c)}{\partial g_j} = \begin{cases} \frac{1}{k} & \text{if } j \in \mathcal{J}, \\ 0 & \text{if } j \notin \mathcal{J} \text{ and } j \neq c, \\ -1 & \text{if } j = c \end{cases}$$

if $k \neq 0$ and $\max_{l\neq c}(1 + \bar{g}_l - \bar{g}_c) > 0$, and that

$$\lim_{T\to 0}\frac{\partial C_{T,1}(\bar{\mathbf{g}}, c)}{\partial g_j} = \begin{cases} \frac{1}{k+1} & \text{if } j \in \mathcal{J}, \\ 0 & \text{if } j \notin \mathcal{J} \text{ and } j \neq c, \\ -\frac{k}{k+1} & \text{if } j = c \end{cases}$$

if $k \neq 0$ and $\max_{l \neq c}(1+\bar{g}_l-\bar{g}_c)=0$. On the other hand, let $\partial \xi_c(\bar{\mathbf{g}})$ be the subdifferential of $\xi_c$ at $\bar{\mathbf{g}}$. Assume that $\max_{l \neq c}(1+\bar{g}_l-\bar{g}_c)=0$. For any $\mathbf{z}=(z_1,\dots,z_m) \in \partial \xi_c(\bar{\mathbf{g}})$, if and only if we have $z_j \in [0,1]$ if $j \in \mathcal{J}$, $z_j = 0$ if $j \notin \mathcal{J}$ and $j \neq c$, and $z_j = -\sum_{l \in \mathcal{J}} z_l$ if $j = c$. This implies that

$$\lim_{T \to 0} \nabla C_{T,1}(\bar{\mathbf{g}}, c) \in \partial \xi_c(\bar{\mathbf{g}}).$$

Accordingly, we conclude the theorem.

## Appendix F. Derivation of multiclass GentleBoost algorithm

The empirical risk over the training data is given by

$$e(\mathbf{g}) = \frac{T}{n}\sum_{i=1}^{n} \log\left[1+\sum_{j \neq c_i} \exp \frac{1+g_j(\mathbf{x}_i)-g_{c_i}(\mathbf{x}_i)}{T}\right].$$

Let $\mathbf{h}(\mathbf{x})=(h_1(\mathbf{x}),\dots,h_m(\mathbf{x}))^T \in \mathcal{G}$ be the increments. Following the derivation of the LogitBoost algorithm, we consider the second-order Taylor expansion of $e(\mathbf{g}+\mathbf{h})$ around $\mathbf{g}$ and employ a diagonal approximation to the Hessian as

$$e(\mathbf{g}+\mathbf{h}) \approx e(\mathbf{g}) + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} h_j(\mathbf{x}_i)\big(\beta_j(\mathbf{x}_i) - \mathbb{I}_{\{j=c_i\}}\big) + \frac{1}{2nT}\sum_{i=1}^{n}\sum_{j=1}^{m} h_j^2(\mathbf{x}_i)\beta_j(\mathbf{x}_i)\big(1-\beta_j(\mathbf{x}_i)\big),$$

where

$$\beta_j(\mathbf{x}_i) = \begin{cases} \dfrac{\exp \frac{1+g_j(\mathbf{x}_i)-g_{c_i}(\mathbf{x}_i)}{T}}{1+\sum_{j \neq c_i} \exp \frac{1+g_j(\mathbf{x}_i)-g_{c_i}(\mathbf{x}_i)}{T}} & \text{if } j \neq c_i, \\ \dfrac{1}{1+\sum_{j \neq c_i} \exp \frac{1+g_j(\mathbf{x}_i)-g_{c_i}(\mathbf{x}_i)}{T}} & \text{if } j = c_i. \end{cases}$$

For each $j$, one can find $h_j(\mathbf{x})$ by minimizing

$$\sum_{i=1}^{n} h_j(\mathbf{x}_i)\big(\beta_j(\mathbf{x}_i) - \mathbb{I}_{\{j=c_i\}}\big) + \frac{1}{2T}\sum_{i=1}^{n} h_j^2(\mathbf{x}_i)\beta_j(\mathbf{x}_i)\big(1-\beta_j(\mathbf{x}_i)\big).$$

The solution is obtained by fitting the regression function $h_j(\mathbf{x})$ based on weighted least-squares of $z_{ij}$ to $\mathbf{x}_i$ with wights $w_{ij}$. We thus have the algorithm in Algorithm 1.

## References

[1] P.L. Bartlett, M.I. Jordan, J.D. McAuliffe, Convexity, classification, and risk bounds, J. Am. Stat. Assoc. 101 (473) (2006) 138–156.

[2] D. Böhning, Multinomial logistic regression algorithm, Ann. Inst. Stat. Math. 44 (1) (1992) 197–200.

[3] E.J. Bredensteiner, K.P. Bennett, Multicategory classification by support vector machines, Comput. Optim. Appl. 12 (1999) 35–46.

[4] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.

[5] K. Crammer, Y. Singer, On the algorithmic implementation of multiclass kernel-based vector machines, J. Mach. Learn. Res. 2 (2001) 265–292.

[6] M. Craven, D. Dopasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, S. Slattery, Learning to extract symbolic knowledge from the World Web Wide, in: The Fifteenth National Conference on Artificial Intelligence, 1998.

[7] Y. Freund, Boosting a weak learning algorithm by majority, Inf. Comput. 21 (1995) 256–285.

[8] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1) (1997) 119–139.

[9] J.H. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, Ann. Stat. 28 (2) (2000) 337–374.

[10] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (1) (2010) 1–22.

[11] T. Gao, D. Koller, Multiclass boosting with hinge loss based on output coding, in: Proceedings of the 22nd International Conference on Machine Learning (ICML), 2010.

[12] Y. Guermeur, Combining discriminant models with new multi-class SVMs, Pattern Anal. Appl. 5 (2) (2002) 168–179.

[13] Y. Lee, Y. Lin, G. Wahba, Multicategory support vector machines: theory and application to the classification of microarray data and satellite radiance data, J. Am. Stat. Assoc. 99 (465) (2004) 67–81.

[14] Y. Liu, Fisher consistency of multicategory support vector machines, in: The Eleventh International Conference on Artificial Intelligence and Statistics, 2007, pp. 289–296.

[15] Y. Liu, X. Shen, Multicategory $\psi$-learning, J. Am. Stat. Assoc. 101 (474) (2006) 500–509.

[16] Y. Liu, H.H. Zhang, Y. Wu, Hard or soft classification? Large-margin unified machines, J. Am. Stat. Assoc. 106 (493) (2011) 166–177.

[17] I. Mukherjee, R. Schapire, A theory of multiclass boosting, in: Advances in Neural Information Processing Systems (NIPS), vol. 24, 2010.

[18] J.M. Ortega, W.C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, SIAM, Philadelphia, 2000.

[19] X. Qiao, L. Zhang, Flexible high-dimensional classification machines and their asymptotic properties, Technical report, arXiv:1310.3004, 2013.

[20] K. Rose, E. Gurewitz, G.C. Fox, Statistical mechanics and phase transitions in clustering, Phys. Rev. Lett. 65 (1990) 945–948.

[21] M. Saberian, N. Vasconcelos, Multiclass boosting: theory and algorithms, in: Advances in Neural Information Processing Systems (NIPS), vol. 25, 2011.

[22] R.E. Schapire, Y. Singer, Improved boosting algorithms using confidence-rated predictions, Mach. Learn. 37 (1999) 297–336.

[23] I. Steinwart, C. Scovel, Fast rates for support vector machines using Gaussian kernels, Ann. Stat. 35 (2) (2007) 575–607.

[24] A. Tewari, P.L. Bartlett, On the consistency of multiclass classification methods, J. Mach. Learn. Res. 8 (2007) 1007–1025.

[25] V. Vapnik, Statistical Learning Theory, John Wiley and Sons, New York, 1998.

[26] J. Weston, C. Watkins, Support vector machines for multiclass pattern recognition, in: The Seventh European Symposium on Artificial Neural Networks, 1999, pp. 219–224.

[27] Y. Wu, Y. Liu, Robust truncated-hinge-loss support vector machines, J. Am. Stat. Assoc. 102 (479) (2007) 974–983.

[28] T. Zhang, Statistical analysis of some multi-category large margin classification methods, J. Mach. Learn. Res. 5 (2004) 1225–1251.

[29] Z. Zhang, G. Wang, D.-Y. Yeung, G. Dai, F. Lochovsky, A regularization framework for multiclass classification: a deterministic annealing approach, Pattern Recognit. 43 (7) (2010) 2466–2475.

[30] Z. Zhang, D. Liu, G. Dai, M.I. Jordan, Coherence functions with applications in large-margin classification methods, J. Mach. Learn. Res. 13 (2012) 2705–2734.

[31] J. Zhu, T. Hastie, Classification of gene microarrays by penalized logistic regression, Biostatistics 5 (3) (2004) 427–443.

[32] J. Zhu, H. Zou, S. Rosset, T. Hastie, Multi-class Adaboost, Stat. Interface 2 (2009) 349–360.

[33] H. Zou, J. Zhu, T. Hastie, New multicategory boosting algorithms based on multicategory Fisher-consistent losses, Ann. Appl. Stat. 2 (4) (2008) 1290–1306.