On the Characterization of a Class of Fisher-Consistent Loss Functions and Its Application to Boosting

Journal of Machine Learning Research 17 (2016) 1-32 Submitted 8/14; Revised 10/15; Published 4/16

On the Characterization of a Class of Fisher-Consistent Loss Functions and its Application to Boosting

Matey Neykov MNEYKOV@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08540, USA

Jun S. Liu JLIU@STAT.HARVARD.EDU
Department of Statistics
Harvard University
Cambridge, MA 02138-2901, USA

Tianxi Cai TCAI@HSPH.HARVARD.EDU
Department of Biostatistics
Harvard University
Boston, MA 02115, USA

Editor: Alexander Rakhlin

Abstract

Accurate classification of categorical outcomes is essential in a wide range of applications. Due to computational issues with minimizing the empirical 0/1 loss, Fisher consistent losses have been proposed as viable proxies. However, even with smooth losses, direct minimization remains a daunting task. To approximate such a minimizer, various boosting algorithms have been suggested. For example, with exponential loss, the AdaBoost algorithm (Freund and Schapire, 1995) is widely used for two-class problems and has been extended to the multi-class setting (Zhu et al., 2009). Alternative loss functions, such as the logistic and the hinge losses, and their corresponding boosting algorithms have also been proposed (Zou et al., 2008; Wang, 2012). In this paper we demonstrate that a broad class of losses, including non-convex functions, achieve Fisher consistency, and in addition can be used for explicit estimation of the conditional class probabilities. Furthermore, we provide a generic boosting algorithm that is not loss-specific. Extensive simulation results suggest that the proposed boosting algorithms could outperform existing methods with properly chosen losses and bags of weak learners.

Keywords: Boosting, Fisher-Consistency, Multiclass Classification, SAMME

1. Introduction

Accurate classification of multi-class outcomes is essential in a wide range of applications. To construct an accurate classifier for the outcome $C \in \{1, ..., n\}$ based on a predictor vector $\mathbf{X}$, the target is often to minimize a misclassification rate, which corresponds to a 0/1 loss. We assume that the data $(C, \mathbf{X}^T)^T$ is generated from a fixed but unknown distribution $\mathscr{P}$. Specifically, one would aim to identify $\mathbf{f} = \{f_1(\cdot), ..., f_n(\cdot)\}$ that maximizes

the misclassification rate

$$\mathbb{L}(\mathbf{f}) = \mathbb{E}\left[\mathbb{1}\{C \neq c_{\mathbf{f}}(\mathbf{X})\}\right] = \mathbb{P}\left\{C \neq c_{\mathbf{f}}(\mathbf{X})\right\}^{1}, \tag{1}$$

under the constraint $\sum_j f_j(\mathbf{X}) = 0$, where $\mathbb{1}(\cdot)$ is the indicator function and for any $\mathbf{f}$, $c_{\mathbf{f}}(\mathbf{X}) = \text{argmax}_j f_j(\mathbf{X})$. Obviously, $\mathbf{f}_{\text{Bayes}} = \{f_{\text{Bayes},j}(\cdot) = \mathbb{P}(C = j \mid \cdot) - n^{-1}, j = 1, ..., n\}$ minimizes (1). In practice, one may approximate the Bayes classifier $c_{\mathbf{f}_{\text{Bayes}}}(\cdot)$ by modeling $\mathbb{P}(C = j \mid \cdot)$ parametrically or non-parametrically. However, due to the curse of dimensionality and potential model mis-specification, such direct modeling may not work well when the underlying conditional risk functions are complex. On the other hand, due to both the discontinuity and the discrete nature of the empirical 0/1 loss, direct minimization is often computationally undesirable.

To overcome these challenges, many novel statistical procedures have been developed by replacing the 0/1 loss with a *Fisher consistent* loss ϕ such that its corresponding minimizer can be used to obtain the Bayes classifier. Lin (2004) showed that a class of smooth convex functions can achieve Fisher consistency (FC) for binary classification problems. Zou et al. (2008) further extended these results to the multi-class setting. Support vector machine (SVM) methods have been shown to yield Fisher consistent results for both binary and multi-class settings (Lin, 2002; Liu, 2007). Relying on these FC results, boosting algorithms for approximating the minimizers of the loss functions have also been proposed for specific choices of losses. Boosting algorithms search for the optimal solution by greedily aggregating a set of "weak-learners" $\mathcal{G}$ via minimization of an empirical risk, based on a loss function ϕ. The classical AdaBoost algorithm (Freund and Schapire, 1995) for example is based on the minimization of the exponential loss, $\phi(x) = e^{-x}$, using the forward stagewise additive modeling (FSAM) approach. Hastie et al. (2009) showed that the population minimizer of the AdaBoost algorithm corresponds to the Bayes rule $c_{\mathbf{f}_{\text{Bayes}}}(\cdot)$ for the two-class setting. Zhu et al. (2009) extended this algorithm and developed the Stagewise Additive Modeling using a Multi-class Exponential (SAMME) algorithm for the multi-class case.

Most existing work on Fisher consistent losses focuses on convex functions such as $\phi(x) = e^{-x}$ and $\phi(x) = |1-x|_+$. However, there are important papers advocating the usage of non-convex loss functions, which we will briefly discuss here. Inspired by Shen et al. (2003), Collobert et al. (2006) explores SVM type of algorithms with the non-convex "ramp" loss instead of the typical "hinge" loss in order to speed up computations. Bartlett et al. (2006) consider the concept of "classification calibration" in the two-class case. Classification calibration of a loss can be understood as uniform FC, along all possible conditional probabilities on the simplex. They demonstrate that non-convex losses such as $1 - \tan(kx), k > 0$ can be classification calibrated in the two class case. More generally, Tewari and Bartlett (2007) extend classification calibration to the multiclass case, and provide elegant characterization theorems. We will draw a link between our work and the work of Tewari and Bartlett (2007) in Section 2.

Asymptotically, procedures such as the AdaBoost based on FC losses would lead to the optimal Bayes classifier, provided sufficiently large space of weak learner set $\mathcal{G}$. However, in finite samples, the estimated classifiers are often far from optimal, and the choice of the loss ϕ could greatly impact the accuracy of the resulting classifier. In this paper, we consider

1. Here the expectation and probability are both with respect to the unknown true distribution $\mathscr{P}$.

a broad class of loss functions that are potentially non-convex and demonstrate that the minimizer of these losses can lead to the Bayes rules for multi-category classification, and in fact can be used to explicitly restore the conditional probabilities. Moreover, we propose a generic algorithm leading to local minimizers of these potentially non-convex losses, which as we argue, can also recover the Bayes rule. The last observation has important consequences in practice, as global minimization of non-convex losses remains a challenging problem. On the other hand, non-convex losses, although not commonly used in the existing literature, could be more robust to outliers (Masnadi-Shirazi and Vasconcelos, 2008). The rest of the paper is organized as follows. In section 2 we detail the conditions for the losses and their corresponding FC results. In settings where the cost of misclassification may not be exchangeable between classes, we generalize our FC results to a weighted loss that accounts for differential costs. In section 3, we propose a generic boosting algorithm for approximating the minimizers and study some of its numerical convergence aspects. In section 4 we present simulation results and real data analysis comparing the performance of our proposed procedures to that of some existing methods including the SAMME. In addition in Section 4 we apply our proposed algorithms to identify subtypes of diabetic neuropathy with EMR data from the Partners. These numerical studies suggest that our proposed methods, with properly chosen losses, could potentially provide more accurate classification than existing procedures. Additional discussions are given section 5. Proofs of the theorems are provided in Appendix A.

2. Fisher Consistency for a general class of loss functions

In this section we characterize a broad class of loss functions which we deem relaxed Fisher consistent. This class encompasses previous classes of loss functions, provided in Zou et al. (2008), but also admits non-convex loss functions.

2.1 Fisher Consistency for 0/1 Loss

Suppose the training data available consists of N realizations of $(C, \mathbf{X}^T)^T$ drawn from $\mathscr{P}$, and let $\mathscr{D} = \{(C_i, \mathbf{X}_i^T)^T, i = 1, \ldots, N\}$. Moreover, we assume throughout that:

$$\min_{j \in \{1,\ldots,n\}} \mathbb{P}(C = j|\mathbf{X}) > 0 : \text{almost surely in } \mathbf{X}. \tag{2}$$

Assumption (2) states that any class C has a chance to be drawn for all $\mathbf{X}$, except on a set of measure 0, where determinism in the class assignment is allowed. For a given C, define a corresponding $n \times 1$ vector $\mathbf{Y}_C = (\mathbb{1}(C = 1), \ldots, \mathbb{1}(C = n))^T$. Under this notation, clearly $\mathbf{Y}_C^T \mathbf{f}(\mathbf{X}) = f_C(\mathbf{X})$. For identifiability the following constraint is commonly used in the existing literature (Lee et al., 2004; Zou et al., 2008; Zhu et al., 2009, e.g.) :

$$\sum_{j=1}^{n} f_j(\cdot) = 0. \tag{3}$$

To identify optimal $\mathbf{f}(\cdot)$ for classifying C based on $\mathbf{f}(\mathbf{X})$, we consider continuous loss functions ϕ as alternatives to the 0/1 loss and aim to minimize

$$\mathbb{L}_\phi(\mathbf{f}) = \mathbb{E}[\phi\{\mathbf{Y}_C^T \mathbf{f}(\mathbf{X})\}] = \mathbb{E}[\phi\{f_C(\mathbf{X})\}] = \sum_{j=1}^{n} \mathbb{E}[\phi\{f_j(\mathbf{X})\}\mathbb{P}(C = j \mid \mathbf{X})], \tag{4}$$

under the constraint (3). The loss function ϕ is deemed Fisher consistent Fisher consistent if the global minimizer (assuming it exists) $\mathbf{f}_\phi = \text{argmin}_{\mathbf{f}:\sum_j f_j=0} \mathbb{L}_\phi(\mathbf{f})$ satisfies

$$c_{\mathbf{f}_\phi}(\mathbf{X}) = {}^2 c_{\mathbf{f}_{\text{Bayes}}}(\mathbf{X}). \tag{5}$$

Hence, with a Fisher consistent loss ϕ, the resulting $\text{argmax}_j \mathbf{f}_\phi(x)$ has the nice property of recovering the optimal Bayes classifier for the 0/1 loss. Clearly, the global minimizer $\mathbf{f}_\phi(x)$ also minimizes $\mathbb{E}[\phi\{f_C(\mathbf{X})\} \mid \mathbf{X} = x]$ for almost all x [3]. With a given data $\mathscr{D}$, we may approximate $\mathbf{f}_\phi$ by minimizing the empirical loss function

$$\widehat{L}_\phi(\mathbf{f}) = \frac{1}{N}\sum_{i=1}^N \phi\{\mathbf{Y}_{C_i}^T \mathbf{f}(\mathbf{X}_i)\} = \frac{1}{N}\sum_{i=1}^N \phi(f_{C_i}(\mathbf{X}_i)) = \frac{1}{N}\sum_{j=1}^n \sum_{i=1}^N \phi(f_j(\mathbf{X}_i))\mathbb{I}(C_i = j),$$

to obtain $\widehat{\mathbf{f}} = \text{argmin}_{\mathbf{f}:\sum_j f_j=0} \widehat{L}_\phi(\mathbf{f})$.

Existing literature on the choice of ϕ focuses almost entirely on convex losses other than a few important exceptions (Bartlett et al., 2006; Tewari and Bartlett, 2007, e.g.). Here, we propose a general class of ϕ to include non-convex losses and generalize the concept of FC as we defined in (5). Specifically, we consider all continuous ϕ satisfying the following properties:

$$\phi(x) - \phi(x') \geq (g(x) - g(x'))k(x') \quad \text{for all } x \in \mathbb{R}, x' \in S = \{z \in \mathbb{R} : k(z) \leq 0\}, \tag{6}$$

where g and k are both strictly increasing continuous functions, with $g(0) = 1, \inf_{x\in\mathbb{R}} g(x) = 0, \sup_{x\in\mathbb{R}} g(x) = +\infty$, $k(0) < 0$ and $\sup_{z\in\mathbb{R}} k(z) \geq 0$. This suggests[4] that $\phi\{g^{-1}(\cdot)\}$ is continuously differentiable and convex on the set $g(S) = \{g(z) : z \in S\}$. However, ϕ itself is not required to be convex or differentiable. Extensively studied convex losses such as $\phi(x) = e^{-x}$ and $\phi(x) = \log(1 + e^{-x})$ both satisfy these conditions. For $\phi(x) = e^{-x}$, (6) would hold if we let $g(x) = e^x$ and $k(x) = -e^{-2x}$. For the logistic loss $\phi(x) = \log(1 + e^{-x})$, we may let $g(x) = e^{cx}$ and $k(x) = -\{ce^{cx}(1 + e^x)\}^{-1}$ for any positive constant $c > 0$. Alternatively, $g(x) = e^x(1 + e^x)/2$ and $k(x) = -2\{e^x(1 + e^x)(1 + 2e^x)\}^{-1}$ would also satisfy (6) for the logistic loss. Our class of losses also allows non-convex functions. For example, $\phi(x) = \log(\log(e^{-x} + e))$ is a non-convex loss and (6) holds if $g(x) = e^x$ and $k(x) = -\{e^x(e^{x+1} + 1)\log(e^{-x} + e)\}^{-1}$. On an important note, we would like to mention that all three examples above can be seen to fall into the general class of classification calibrated loss functions in the two class case, as defined by Bartlett et al. (2006) and hence are FC in the two-class case. We will see a more general statement relating condition (6) to the notion of classification calibration in the two class case (see Remark 2.2 below).

Next, we extend the FC property (5), to allow for more generic classification rules. For a loss function ϕ, if there exists a functional $\mathcal{H}$ such that the minimizer of (4) has the property:

$$\underset{j\in\{1,\ldots,n\}}{\text{argmax}}\, \mathcal{H}\{f_{\phi,j}(\mathbf{X})\} = c_{\mathbf{f}_{\text{Bayes}}}(\mathbf{X}), \tag{7}$$

2. Formally the "=" in (5) should be understood as "$\subseteq$". For the sake of simplicity, we keep this slight abuse of notation consistent throughout the paper.
3. Provided that this minimizer of $\mathbb{E}[\phi\{f_C(\mathbf{X})\} \mid \mathbf{X}]$ exists on a set of probability 1, and $\mathbb{E}[|\phi\{f_C(\mathbf{X})\}|] < \infty$ so that Fubini's theorem holds.
4. We provide a formal proof of this fact in Appendix A under Lemma A.1.

then we call it *relaxed* Fisher consistent (RFC). Obviously, the RFC property would still recover the Bayes classifier. Moreover FC losses are special cases of the RFC losses with an increasing $\mathcal{H}$.

We will now point out a connection between RFC and multiclass classification calibration as defined by Tewari and Bartlett (2007). Re-casting the definition of multiclass classification calibration to our framework, it requires that for any vector $\mathbf{w}$ on the simplex, the minimizer (assuming that it exists):

$$\widehat{\mathbf{F}}(\mathbf{w}) = \underset{\mathbf{F}:\sum_j F_j=0}{\operatorname{argmin}} \sum_{i=1}^{n} \phi(F_j)w_j \text{ satisfies } \underset{j\in\{1,\ldots,n\}}{\operatorname{argmax}} \mathcal{H}(\phi(\widehat{F}_j)) = \underset{j\in\{1,\ldots,n\}}{\operatorname{argmax}} w_j, \tag{8}$$

for some functional $\mathcal{H}$. In words, classification calibration ensures that regardless of the conditional distribution of $C|\mathbf{X}$, one can recover the Bayes rule. In contrast, RFC requires this to happen for the distribution at hand $C|\mathbf{X}$, for ($\mathscr{P}$ almost) all $\mathbf{X}$. This subtle but important distinction makes a difference. Example 4 in Tewari and Bartlett (2007) shows that if ϕ is positive and convex the conditions in (8) cannot be met for all vectors $\mathbf{w}$ on the simplex, when we have at least 3 classes. On the contrary, in the present paper we argue that in fact condition (8) remains plausible for both convex and non-convex losses, provided that we require that the points $\mathbf{w}$ are not allowed to be vertexes of the simplex (i.e. $w_j > 0$ for all j), which relates back to assumption (2).

The next result justifies that the proposed losses satisfying (6) are RFC with $\mathcal{H}(x) = H_\phi(x) \equiv g(x)k(x)$. We first present in Theorem 2.1 the property of a general constrained minimization problem, which is key to establishing the RFC.

Theorem 2.1 *For a loss ϕ satisfying (6), consider the optimization problem with some given $w_j > 0$:*

$$\min_{\mathbf{F}=(F_1,\ldots,F_n)^T} \sum_{j=1}^{n} \phi(F_j)w_j \quad \textit{under the constraint} \quad \prod_{j=1}^{n} g(F_j) = 1. \tag{9}$$

Assume that there exists a minimum denoted by $\widehat{\mathbf{F}} = (\widehat{F}_1, \ldots, \widehat{F}_n)^T$. Then the minimizer $\widehat{\mathbf{F}}$ must satisfy

$$H_\phi(\widehat{F}_j)w_j = \mathcal{C} \quad \textit{for some } \mathcal{C} < 0. \tag{10}$$

Moreover, if the function $H_\phi(\cdot)$ is strictly monotone there is a unique point with the property described above.

This result indicates that $H_\phi(\widehat{F}_j)$ is inversely proportional to the weight w_j. Now, consider $g(x) = \exp(x)$, $w_j = \mathbb{P}(C = j|\mathbf{X} = x)$, and $F_j = f_j(x)$, where we hold x . Then we can recover $c_{\mathbf{f}_{\text{Bayes}}}(x)$ by classifying C according to $\operatorname{argmax}_j\{-H_\phi(\widehat{F}_j)\}^{-1} = \operatorname{argmax}_j H_\phi(\widehat{F}_j)$ (the negative sign comes in because $\mathcal{C} < 0$), which implies that ϕ is RFC. Note that when $H_\phi(\cdot)$ is not increasing, Theorem 2.1 does not immediately imply that ϕ is a Fisher consistent loss according to definition (5), because the Bayes classifier need not be recovered by $\operatorname{argmax}_j \widehat{F}_j$. Nevertheless, we have the following:

Proposition 2.2 *Assume the same conditions as in Theorem 2.1. Then in addition to (10) we have:*

$$\operatorname*{argmax}_{j\in\{1,\dots,n\}} \widehat{F}_j = \operatorname*{argmax}_{j\in\{1,\dots,n\}} w_j,$$

and hence ϕ is also FC in the sense of (5).

The validity of Proposition 2.2 can be deduced from Theorem 2.1 and an application of Lemma 4 of Tewari and Bartlett (2007), but for completeness we include a simple standalone proof in Appendix A. While Proposition 2.2 states that ϕ is FC, Theorem 2.1 suggests that in addition to classification, one can recover conditional probabilities by calculating:

$$w_j = \frac{\{H_\phi(\widehat{F}_j)\}^{-1}}{\sum_{j=1}^n \{H_\phi(\widehat{F}_j)\}^{-1}}. \tag{11}$$

It is also worth noting here that the constraint in (9), generalizes the typical identifiability constraint (3), and the two coincide when $g(\cdot) = \exp(\cdot)$. We proceed by formulating a sufficient condition for the optimization problem in Theorem 2.1 to have a minimum without requiring the convexity or differentiability of ϕ.

Theorem 2.3 *The optimization problem in Theorem 2.1 has a minimum if either of the following conditions holds:*

i. ϕ is decreasing on the whole $\mathbb{R}$ and for all $c > 0$:

$$c\phi(g^{-1}(x)) + \phi(g^{-1}(x^{1-n})) \uparrow +\infty, \ \text{as } x \downarrow 0, \tag{12}$$

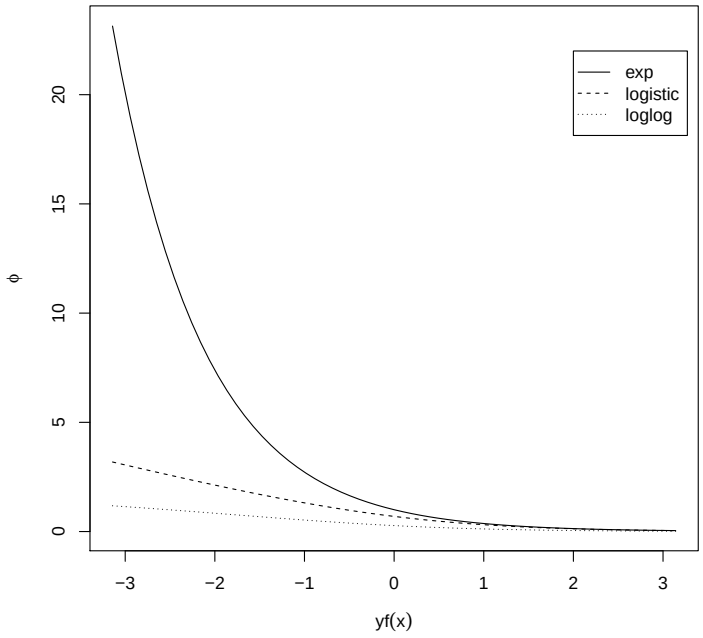
ii. ϕ is not decreasing on the whole $\mathbb{R}$.

Remark 2.1 *It follows that in any case, problem (9) has a minimum when the loss function is bounded from below and unbounded from above.*

Remark 2.2 *Take $g = \exp$ to match the constraint in (9) with the constraint considered by Bartlett et al. (2006). It turns out that a loss function obeying (6) and either i. or ii in Theorem 2.3. is classification calibrated in the two class case. See and Lemma A.2 in Appendix A for a formal proof of this fact.*

Clearly, by Remark 2.1, problem (9) would have a minimum for all three losses suggested earlier — the exponential, logistic (for both $g(x) = e^{cx}$, and $g(x) = e^{x\frac{e^x+1}{2}}$), and log-log loss.

Finally we conclude this subsection, by noting that the assumptions in both Theorem 1 and 2 in Zou et al. (2008) can be seen to imply that the assumptions in Theorems 2.1 and 2.3 hold, thus rendering these theorems as consequences of the main result shown above. For completeness we briefly recall what these conditions are. In Theorem 1, Zou et al. (2008) require a twice differentiable loss function ϕ such that $\phi'(0) < 0$ and $\phi'' > 0$. In Theorem 2 these conditions are slightly relaxed by allowing for part linear and part constant convex losses.



2.2 Fisher Consistency for Weighted 0/1 Loss

Although the expected 0/1 loss or equivalently the overall misclassification is an important summary for the overall performance of a classification, alternative measures may be preferred when the cost of misclassification is not exchangeable across outcome categories. For such settings, it would be desirable to incorporate the differential cost when evaluating the classification performance and consider a weighted misclassification rate. Consider a cost matrix $\mathcal{W} = [W(j,\jmath)]_{n\times n}$ with $W(j,\jmath)$ representing the cost in classifying the $\jmath^{\text{th}}$ class to the j^{th} class. Then, the optimal Bayes classifier is

$$c^{\mathcal{W}}_{\mathbf{f}_{\text{Bayes}}}(\mathbf{X}) = \underset{j}{\operatorname{argmin}} \sum_{\jmath=1}^{n} W(j,\jmath)\mathbb{P}(C=\jmath|\mathbf{X}). \tag{13}$$

Setting $\mathcal{W} = 1-\mathcal{I}$ corresponds to the 0/1 loss and $c^{\mathcal{W}}_{\mathbf{f}_{\text{Bayes}}} = c_{\mathbf{f}_{\text{Bayes}}}$, where $\mathcal{I}$ is the identity matrix. Without loss of generality, we assume that $W(\jmath,\jmath) \geq 0$. For ϕ satisfying (6) and the condition in Theorem 2.3, we next establish the FC results for the weighted 0/1 loss parallel to those given in Theorems 2.1 and 2.3.

Proposition 2.4 *Define the weighted loss* $\ell(F_\jmath) = \sum_{j=1}^{n} \phi(F_j)W(j,\jmath)$. *Then the optimization problem:*

$$\min_{\mathbf{F}=(F_1,\ldots,F_n)^T:\prod_{\jmath=1}^{n} g(F_\jmath)=1} \sum_{\jmath=1}^{n} \ell(F_\jmath)\mathbb{P}(C=\jmath|\mathbf{X}), \tag{14}$$

has a minimizer $\widehat{\mathbf{F}}^{\mathcal{W}} = (\widehat{F}_1^{\mathcal{W}}, ..., \widehat{F}_n^{\mathcal{W}})^T$ *which satisfies the property that:*

$$H_\phi(\widehat{F}_j^{\mathcal{W}}) w_j^{\mathcal{W}} = \widetilde{\mathcal{C}} \quad \text{for some } \widetilde{\mathcal{C}} < 0, \tag{15}$$

where $w_j^{\mathcal{W}} = \sum_{j=1}^n W(j,j)\mathbb{P}(C=j|\mathbf{X})$ *assuming that* $w_j^{\mathcal{W}} > 0$.

This proposition is a direct consequence of Theorem 2.1, after exchanging summations:

$$\sum_{j=1}^n \sum_{j=1}^n \phi(F_j) W(j,j) \mathbb{P}(C=j|\mathbf{X}) = \sum_{j=1}^n \phi(F_j) w_j^{\mathcal{W}}.$$

Remark 2.3 *Note that the above result hints on how one can relax assumption (2) by using the loss* ℓ *constructed with* $\mathcal{W} = 1 - \mathcal{I}$. *Using this particular* ℓ, *Proposition 2.4 simply requires* $w_j^{\mathcal{W}} > 0$, *which would be satisfied if we required:*

$$\max_{j \in \{1,...,n\}} \mathbb{P}(C=j|\mathbf{X}) < 1 : \text{ almost surely in } \mathbf{X}.$$

This is indeed weaker than (2). If we wanted to recover the conditional probabilities simply note that $\mathbb{P}(C=j|\mathbf{X}) = 1 - w_j^{\mathcal{W}}$, *and hence:*

$$\mathbb{P}(C=j|\mathbf{X}) = 1 - \frac{\{H_\phi(\widehat{F}_j^{\mathcal{W}})\}^{-1}}{\sum_{j=1}^n \{H_\phi(\widehat{F}_j^{\mathcal{W}})\}^{-1}}.$$

The result also suggests that using the modified loss ℓ, we can attain the optimal weighted Bayes classifier $c^{\mathcal{W}}_{\mathbf{f}_{\text{Bayes}}}(\mathbf{X})$ based on $\operatorname{argmin}_j H_\phi(\widehat{F}_j^{\mathcal{W}})$.

3. Generic Algorithm for Constructing the Classifier

In this section we provide a generic boosting algorithm, based on the explicit structure (6) that the RFC loss functions possess, and analyze certain numerical convergence aspects of the algorithm in the special case when $g = \exp$.

3.1 A Generic Boosting Algorithm

The properties of ϕ and the results in Theorem 2.1 and 2.3 also lead to a natural iterative generic boosting algorithm to attain the minimizer.

3.1.1 A Conditional Iteration

In this subsection, we provide an iterative procedure, conditional on $\mathbf{X} = x$, which eventually leeds to a generic boosting algorithm. The usefulness of this conditional iteration is based on the following result.

Theorem 3.1 *Assume that* ϕ *satisfies (6) and the condition in Theorem 2.3. Starting from* $\mathbf{F}^{(0)} \equiv 0$, *i.e.* $F_j^{(0)} = 0$ *for all* j, *define the following iterative procedure:*

$$\mathbf{F}^{(m+1)} = \underset{\mathbf{F}: \prod g(F_j)=1}{\operatorname{argmax}} \sum_{j=1}^n \{g(F_j^{(m)}) - g(F_j)\} k(F_j) w_j. \tag{16}$$

This iteration is guaranteed to converge to a point $\mathbf{F}^*$ *with the following property:*

$$g(F_j^*)k(F_j^*)w_j = H_\phi(F_j^*)w_j = \mathcal{C} < 0.$$

In the theorem above, the iterations are defined conditionally on $\mathbf{X} = x$, and F_j can be understood as $f_j(x)$. If $H_\phi(\cdot)$ turns out to be monotone, the procedure above will converge to the global minimum, as we can conclude straight from Theorem 2.1. Even if the procedure does not converge to a global minimum, because of the property of the point that it converges to, $\mathbf{F}^*$ can be used to recover the Bayes classifier. This observation is particularly useful for minimizing a non-convex loss functions as in such cases it is often hard to arrive at the global minimum. Moreover, as before the point $\mathbf{F}^*$ can be used not only for classification purposes, but also to recover the exact probabilities w_j.

In practice, the procedure described in (16) can be used to derive algorithms for boosting. However, an unconditional version of (16) is needed since w_j are unknown in general. Noting that the expectation of $\mathbb{1}(C = j)$ given $\mathbf{X}$ is w_j, we have

$$\begin{aligned}
&\mathbb{E}\left(\sum_{j=1}^{n}\Big[g\{F_j^{(m)}(\mathbf{X})\} - g\{F_j(\mathbf{X})\}\Big]k\{F_j(\mathbf{X})\}w_j\right)\\
=&\mathbb{E}\left(\sum_{j=1}^{n}\Big[g\{F_j^{(m)}(\mathbf{X})\} - g\{F_j(\mathbf{X})\}\Big]k\{F_j(\mathbf{X})\}\mathbb{1}(C = j)\right)\\
=&\mathbb{E}\left(\Big[g\{\mathbf{Y}_C^T\mathbf{F}^{(m)}(\mathbf{X})\} - g\{\mathbf{Y}_C^T\mathbf{F}(\mathbf{X})\}\Big]k\{\mathbf{Y}_C^T\mathbf{F}(\mathbf{X})\}\right), \qquad (17)
\end{aligned}$$

where recall that $\mathbf{Y}_C = (\mathbb{1}(C = 1), ..., \mathbb{1}(C = n))^T$. We next derive a boosting algorithm iterating based on an empirical version of (17).

3.1.2 The Boosting Algorithm

To derive the boosting algorithm, we let $\mathcal{G} = \{G_b(\cdot), b = 1, ..., B\}$ denote the bag of weak learners with $G(\mathbf{X}) \in \{1, ..., n\}$ denoting the predicted class based on learner G. For the bth classifier in $\mathcal{G}$, define a corresponding vectorized version of G_b, $\mathbf{F}_b = (F_{b1}, ..., F_{bn})$, with

$$F_{bj}(\mathbf{X}) = \mathcal{C}_- + \mathbb{1}\{G_b(\mathbf{X}) = j\}(\mathcal{C}_+ - \mathcal{C}_-),$$

where $\mathcal{C}_- < 0$ and $\mathcal{C}_+ > 0$ are chosen such that $\prod_{j=1}^n g(F_{bj}) = 1$. Obviously, $\mathbf{Y}_C^T\mathbf{F}_b(\cdot) = \mathcal{C}_- + \mathbb{1}(G_b(\cdot) = C)(\mathcal{C}_+ - \mathcal{C}_-)$. Let $\mathcal{G}^* = \{\mathbf{F}_b(\cdot), b = 1, ..., B\}$ denote the bag of vectorized classification functions corresponding to the classifiers in $\mathcal{G}$.

We next propose a generic iterative boosting algorithm that greedily searches for an optimal weight and for which weak learner to aggregate at each iteration. The loss function ϕ is not directly used, and instead we rely on the $g(\cdot)$ and $k(\cdot)$ functions as specified in (6). Specifically, we initialize $\mathbf{F}^{(0)} := 0$. Then for $m = 1, ..., M$ with M being the total number of desired iterations, we obtain the maximizer of

$$\sum_{i=1}^{N} g\{\mathbf{Y}_{Ci}^T\mathbf{F}^{(m-1)}(\mathbf{X}_i)\}\left[1 - g\{\mathbf{Y}_{Ci}^T\mathbf{F}(\mathbf{X}_i)\}^\beta\right]k\left(g^{-1}\left[g\{\mathbf{Y}_{Ci}^T\mathbf{F}^{(m-1)}(\mathbf{X}_i)\}g\{\mathbf{Y}_{Ci}^T\mathbf{F}(\mathbf{X}_i)\}^\beta\}\right]\right),$$

with respect to $\mathbf{F} \in \mathcal{G}^*$ and $\beta \geq 0$, denoted by $\widehat{\mathbf{F}}$ and $\widehat{\beta}$. Then we update the classifier coordinate-wise as $F_j^{(m)} = g^{-1}\{g(F_j^{(m-1)})g(\widehat{F}_j)^{\widehat{\beta}})\}$ so that we have the following

$g\{\mathbf{Y}_C^T\mathbf{F}^{(m)}\} = g\{\mathbf{Y}_C^T\mathbf{F}^{(m-1)}\}g\{\mathbf{Y}_C^T\widehat{\mathbf{F}}\}^{\widehat{\beta}}$ holding for all C. This will ensure that the property $\prod_{j=1}^n g(F_j^{(m)}) = 1$ continues to hold throughout the iterations. Thus at each iteration, we would be greedily maximizing:

$$\sum_{i=1}^N \left[g\{\mathbf{Y}_{Ci}^T\mathbf{F}^{(m-1)}(\mathbf{X}_i)\} - g\{\mathbf{Y}_{Ci}^T\mathbf{F}^{(m)}(\mathbf{X}_i)\}\right] k\left\{\mathbf{Y}_{Ci}^T\mathbf{F}^{(m)}(\mathbf{X}_i)\right\},$$

which is exactly the empirical version of (17).

For illustration, consider $g(x) = e^x$ with ϕ being differentiable and hence we may let $k(x) = \dot{\phi}(x)e^{-x}$. In this special case the update of $\mathbf{F}^{(m)}$ simplifies to $\mathbf{F}^{(m)} = \mathbf{F}^{(m-1)} + \widehat{\beta}\widehat{\mathbf{F}}$. Therefore following the iteration described above we have:

$$\underset{\mathbf{F}\in\mathcal{G}^*,\ \beta\geq 0}{\operatorname{argmin}} \sum_{i=1}^N -\{e^{-\beta\mathbf{Y}_{Ci}^T\mathbf{F}(\mathbf{X}_i)} - 1\}\dot{\phi}\{\mathbf{Y}_{Ci}^T\mathbf{F}^{(m-1)}(\mathbf{X}_i) + \beta\mathbf{Y}_{Ci}^T\mathbf{F}(\mathbf{X}_i)\}. \tag{18}$$

Note here, the apparent similarity between a coordinate descent (or gradient descent in a functional space) as proposed in Mason et al. (1999) and Friedman (2001), and the above iteration. Finally, we summarize the algorithm as follows.

Algorithm 1: Generic Boosting Algorithm

1. Set $\mathbf{F}^{(0)} = 0$
2. For $m = 1, \ldots, M$
   (a) Maximize $\sum_{i=1}^N \left\{g\{\mathbf{Y}_{Ci}^T\mathbf{F}^{(m-1)}(\mathbf{X}_i)\}\left[1 - g\{\mathbf{Y}_{Ci}^T\mathbf{F}(\mathbf{X}_i)\}^\beta\right] \times k\left(g^{-1}\left[g\{\mathbf{Y}_{Ci}^T\mathbf{F}^{(m-1)}(\mathbf{X}_i)\}g\{\mathbf{Y}_{Ci}^T\mathbf{F}(\mathbf{X}_i)\}^\beta\right]\right)\right\}$
   with respect to $\mathbf{F} \in \mathcal{G}^*$, $\beta \geq 0$ to obtain $\widehat{\mathbf{F}}$ and $\widehat{\beta}$
   (b) Update $\mathbf{F}^{(m)}$ coordinate-wise as $F_j^{(m)} = g^{-1}\{g(F_j^{(m-1)})g(\widehat{F}_j)^{\widehat{\beta}})\}$.
3. Output $\mathbf{F}^{(M)}$ and classify via $\operatorname{argmax}_j H_\phi(F_j^{(M)})$.

3.2 Numerical Convergence of the Algorithm when $g(\cdot) = \exp(\cdot)$

In this subsection we illustrate how algorithm 1 performs in finite samples, if we let it run until convergence (using potentially infinitely many iterations). More specifically, we study the properties of the iteration above in the case when $g(x) = \exp(x)$, or in other words we are concerned with the iteration given by (18). In addition, we also want to explore the relationship between iteration (18) and the following optimization problem:

$$\inf_{\mathbf{F}\in\operatorname{span}\mathcal{G}^*} \sum_{i=1}^N \phi(\mathbf{Y}_{C_i}^T\mathbf{F}(\mathbf{X}_i)). \tag{19}$$

To this end we formulate the following:

Definition 3.1 (Looping closure) *Let π be a permutation of the numbers $\{1, \ldots, n\}$ into $\{\pi_1, \ldots, \pi_n\}$. Consider the following "loop" functions, such that for all $i = 1, \ldots, n$: $l^{(0)}(\pi_i) =$*

$\pi_i, l^{(1)}(\pi_i) = \pi_{i+1}$, $l^{(k)}(\cdot) = l^{(1)}(l^{(k-1)}(\cdot))$, *where the indexing is* $\bmod n$, *and* $k = 1, \ldots, n-1$[5]. *We say that a classifier bag* $\mathcal{G}$ *is closed under "looping" if there exists a permutation* π *such that for all* $G \in \mathcal{G}$ *it follows that* $l^{(k)} \circ G \in \mathcal{G}$ *for all* $k = 0, \ldots, n-1$.

In practice, closure under looping can easily be achieved if it is not already present, by adding the missing classifiers to the bag. Similar bag closures have been considered in the two class case in Mason et al. (1999). We start our discussion with the following proposition, providing a property of the algorithm at its limiting points.

Proposition 3.2 *Suppose that the loss function* ϕ *is decreasing, continuously differentiable, bounded from below and satisfies (6) with* $g = \exp$. *Furthermore assume that, the classifier bag is closed under looping (see Definition 3.1). Then iterating (18), using possibly infinite amount of iterations until a limiting point* $\mathbf{F}^{(\infty)}$ *is reached, guarantees that the following condition holds:*

$$\sum_{i=1}^{N} \mathbf{Y}_{Ci}^T \mathbf{F}(\mathbf{X}_i) \dot{\phi}(\mathbf{Y}_{Ci}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i)) = 0, \tag{20}$$

for all $\mathbf{F} \in \mathcal{G}^*$.

Clearly, all examples of loss functions we considered satisfy the assumptions of Proposition 3.2. In the case when ϕ is convex, (20) shows that by iterating (18) we would arrive at an infimum of problem (19). This can be easily seen along the following lines. Assume that the $\widetilde{\mathbf{F}}$ is a point of infimum of (19) over the span of $\mathcal{G}^*$. By convexity of ϕ we have:

$$\sum_{i=1}^{N} \phi(\mathbf{Y}_{Ci}^T \widetilde{\mathbf{F}}(\mathbf{X}_i)) - \sum_{i=1}^{N} \phi(\mathbf{Y}_{Ci}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i)) \geq \sum_{i=1}^{N} \mathbf{Y}_{Ci}^T [\widetilde{\mathbf{F}}(\mathbf{X}_i) - \mathbf{F}^{(\infty)}(\mathbf{X}_i)] \dot{\phi}(\mathbf{Y}_{Ci}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i)) = 0.$$

The last equality follows from (20) and the fact that $\widetilde{\mathbf{F}} - \mathbf{F}^{(\infty)}$ is in the span of classifiers.

In the case when ϕ is not convex, condition (20) remains meaningful, though it doesn't guarantee convergence to the infimum. In order for us to relate condition (20) to equation (11) in the general (non-convex loss) case, and make it more intuitive, we consider a simple and illustrative example. We restrict our attention to the two class case ($n = 2$), but the example can be easily generalized.

Example 3.1 *Consider a (disjoint) partition of the predictor support:* $\mathcal{X} = \mathcal{X}_1, \ldots, \mathcal{X}_B$. *Construct classifiers based on that partition in the following manner:*

$$G_b(x) = \begin{cases} 1, & \text{if } x \in \mathcal{X}_b \\ 2 & \text{otherwise} \end{cases},$$

and close them under looping. It is easily seen that, under this framework the vector $\mathbf{F}^{(\infty)}(x)$ *is constant for* $x \in \mathcal{X}_b$ *for a fixed* b. *Denote the value of this constant with* $\mathbf{F}_b^{(\infty)}$. *Plugging in the* b^{th} *classifier in equation (20) we obtain:* $N_b \dot{\phi}(\mathbf{Y}_1^T \mathbf{F}_b^{(\infty)}) - (N - N_b) \dot{\phi}(\mathbf{Y}_2^T \mathbf{F}_b^{(\infty)}) = 0$,

5. Note that the loop functions depend on the permutation π, but we suppress this dependence for clarity of exposition.

where N_b is the number of observations correctly classified by the b^{th} classifier, or in other words:

$$\frac{\{\dot{\phi}(\mathbf{Y}_1^T\mathbf{F}_b^{(\infty)})\}^{-1}}{\{\dot{\phi}(\mathbf{Y}_1^T\mathbf{F}_b^{(\infty)})\}^{-1}+\{\dot{\phi}(\mathbf{Y}_2^T\mathbf{F}_b^{(\infty)})\}^{-1}}=\frac{N_b}{N}.^{6} \tag{21}$$

The LHS of (21) is an estimate of the probability $\mathbb{P}(C=1|\mathbf{X}\in\mathcal{X}_b)$, which in turn is a proxy to the Bayes classifier. For the RHS of (21), note that our probability recovering rule (11) with $g=\exp$ becomes:

$$\frac{\{\dot{\phi}(\mathbf{Y}_j^T\mathbf{F}^{(\infty)}(x))\}^{-1}}{\{\dot{\phi}(\mathbf{Y}_1^T\mathbf{F}^{(\infty)}(x))\}^{-1}+\{\dot{\phi}(\mathbf{Y}_2^T\mathbf{F}^{(\infty)}(x))\}^{-1}},$$

in the two class case. This expression is in complete agreement with the RHS of (21).

3.2.1 Convergence Analysis

In the convex loss function case, property (20) will be matched by a gradient descent methods in the function space such as AnyBoost (Mason et al., 1999). This motivates us to consider the question of the convergence rate of the newly suggested algorithm — is it slower, faster or the same as a gradient descent in the convex loss function case? At first glance the rate might appear to be slower as we are not using the "fastest" decrease at each iteration using simply the exp function. In the end of this subsection we establish a geometric rate of convergence under certain assumptions, which matches the convergence rate for gradient descent under similar assumptions.

As we argued in the previous subsection, in the case of a convex loss ϕ, (20) guarantees that iteration (18) converges to the infimum of problem (19). Let $\mathbf{Y}_{C_i}^T\mathbf{F}^{(\infty)}(\mathbf{X}_i)$ be the limiting (allowed to be $\pm\infty$) values achieving the infimum above. Before we formalize the convergence rate result, we will characterize the behavior of $\mathbf{Y}_{C_i}^T\mathbf{F}^{(\infty)}(\mathbf{X}_i)$. This question is of interest in its own right, as this characterization remains valid regardless of what boosting algorithm one decides to use to obtain the minimum/infimum.

For what follows we consider a loss function ϕ which satisfies a mildly strengthened condition (12). Namely, let ϕ be decreasing and for any $\alpha, c>0$ it satisfies the following condition:

$$\phi(x)+c\phi(-\alpha x)\uparrow+\infty \text{ as } x\uparrow+\infty \tag{22}$$

It is worth noting that if condition (12) is satisfied for all n (recall that $g=\exp$ here) this would imply (22). Denote with B the total number of weak learners in the bag. Let $\mathbf{D}:=\{\mathbf{Y}_{Ci}^T\mathbf{F}_j(\mathbf{X}_i)\}_{j,i}$ be a $B\times N$ matrix each entry of which is either $\mathcal{C}_+$ or $\mathcal{C}_-$. Again, we assume that the bag is closed under looping. Let $\mathbf{v}\in\mathbb{R}^N$ be a vector. Consider the equation $\mathbf{D}^T\boldsymbol{\alpha}=\mathbf{v}$ for some vector $\boldsymbol{\alpha}\in\mathbb{R}_{0,+}^B$, where $\mathbb{R}_{0,+}=[0,\infty)$. Note that because of the looping closure[7] the linear equation above has solution iff the equation $\mathbf{D}^T\boldsymbol{\alpha}=\mathbf{v}$ has a solution with $\boldsymbol{\alpha}\in\mathbb{R}^B$, since without loss of generality we can add a large positive constant

6. Here we assume that $\dot{\phi}(\mathbf{Y}_j^T\mathbf{F}_b^{(\infty)})\neq 0, j\in\{1,2\}$, which can be ensured if ϕ is unbounded from above
7. Looping closure (Definition 3.1) implies that the column sums of $\mathbf{D}$ are 0.

to the coordinates of $\boldsymbol{\alpha}$. It follows that the equation $\mathbf{D}^T\boldsymbol{\alpha} = \mathbf{v}$ has a solution $\boldsymbol{\alpha} \in \mathbb{R}^B_{0,+}$ iff $\mathbf{v} \in \text{row}(\mathbf{D})$.

To see the connection between the linear equation above and optimization problem (19) consider the following simple example:

Example 3.2 *The function $\sum_{i=1}^N \phi(\sum_{j=1}^B \alpha_j \mathbf{Y}_{Ci}^T \mathbf{F}_j(\mathbf{X}_i))$ cannot have a minimum, if there exists a vector $\mathbf{v} \in \mathbb{R}^N_+$, such that the equation $\mathbf{D}^T\boldsymbol{\alpha} = \mathbf{v}$ has a solution — $\widehat{\boldsymbol{\alpha}} \in \mathbb{R}^B_{0,+}$, where $\mathbb{R}_+ = (0, \infty)$. To see this, suppose the contrary, take an arbitrary constant $R > 0$ and note that:*

$$\sum_{i=1}^N \phi(\sum_{j=1}^B R\widehat{\alpha}_j \mathbf{Y}_{Ci}^T \mathbf{F}_j(\mathbf{X}_i)) = \sum_{i=1}^N \phi(Rv_i).$$

Take the limit $R \to \infty$, and it is clear that the infimum $N\phi(+\infty)$ is achieved.

Example 3.2 illustrates that if we want to have a solution smaller than $N\phi(+\infty)$, $\mathbf{D}$ cannot have rank N. Denote the rank of $\mathbf{D}$ with r.

More generally, our next result provides a characterization of how many (and which) of the values $\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i)$ are set to $+\infty$ at the infimum of (19). Consider the perp space of the row space of the matrix $\mathbf{D}$, $\mathbf{E} := \text{row}(\mathbf{D})^\perp$. Out of all possible bases of $\mathbf{E}$ including the 0 vector, select the one $\mathbf{e}_1, \ldots, \mathbf{e}_s$ ($s = \min(N, B) - r + 1$) for which the vector $\mathbf{e}_1$ has the most strictly positive entries at I coordinates and zeros at the rest[8]. We have the following:

Proposition 3.3 *Let ϕ be a decreasing loss function satisfying (22). Set $M := N - I$, where $I \in \{0, \ldots, N\}$. We have that:*

$$(N - M - 1)\phi(0) + (M + 1)\phi(+\infty) < \inf_{\mathbf{F} \in \text{span}\, \mathcal{G}^*} \sum_{i=1}^N \phi(\mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)) \leq (N - M)\phi(0) + M\phi(+\infty).$$

Moreover, exactly M of the values $\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i)$ (i will be corresponding to the 0 coordinates of $\mathbf{e}_1$) will be set to $+\infty$ at the infimum.

Proposition 3.3 characterizes the cases when one should expect problem (19) to have a minimum. In fact, in the cases where $I > 0$, we can simply delete the observations corresponding to the rows of $\mathbf{e}_1$ that are 0, and solve the optimization only on the set of observations left, as it can be seen from the proof.

Next we formulate the speed of the convergence of the algorithm we suggested, in the case when the function ϕ is convex. For simplicity we assume that the matrix of classifier entries, $\mathbf{D}$, is such that there is a strictly positive vector in the perp of the row space of $\mathbf{D}$. If that is not the case as argued we can delete observations that will be set to $+\infty$ at the maximum, and work with the rest. Denote with $\mathcal{S} = \{\mathbf{v} : \mathbf{D}^T\boldsymbol{\alpha} = \mathbf{v} \text{ with } \boldsymbol{\alpha} \geq 0, \sum_{i=1}^N \phi(v_i) \leq N\phi(0)\}$. Proposition 3.3 then implies that, the set $\mathcal{S}$ is bounded coordinate-wise. Next we formulate the result:

Theorem 3.4 *Let the convex, decreasing loss function ϕ be strongly convex with Lipchitz and bounded derivative on any compact subset of $\mathbb{R}$, and satisfies (22) and (6) with $g = \exp$. Furthermore, assume that there is a stricly positive vector in $\text{row}(\mathbf{D})^\perp$, and define the*

8. We allow $I = 0$, in which case $\mathbf{e}_1$ would simply represent the 0 vector.

set $\mathcal{S}$ as above. Let $\mathbf{F}^* \in \operatorname{span}\mathcal{G}^*$ achieves the minimum in problem (19). Denote with $\varepsilon_m = \sum_{i=1}^N \phi(\mathbf{Y}_{Ci}^T \mathbf{F}^{(m)}(\mathbf{X}_i)) - \sum_{i=1}^N \phi(\mathbf{Y}_{Ci}^T \mathbf{F}^*(\mathbf{X}_i))$, where $\mathbf{F}^{(m)}$ is produced iteratively using (18). Then there exists a constant $K < 1$ depending on the matrix $\mathbf{D}$, the sample size N and the set $\mathcal{S}$, such that:

$$\varepsilon_{m+1} \leq \varepsilon_m K.$$

As we can see from Theorem 3.4 if we use this algorithm in the convex loss function case, we wouldn't lose convergence speed to gradient descent (see Nesterov (2004) Theorem 2.1.14), but in the non-convex function case which still obeys (6) this algorithm will be converging to a local minimum. In the latter case we will still be capable of recovering the Bayes classifier, as indicated by equation (20).

4. Numerical Studies and Data Examples

In this section we validate empirically the performance of the generic boosting algorithm developed in the previous section with various choices of ϕ, comparing it to popular classification algorithms such as SVM and SAMME on simulated datasets, real benchmark datasets from the UCI machine learning, as well as an EMR study on diabetic neuropathy conducted at the Partners Healthcare.

For each dataset, we evaluated our proposed boosting algorithm based on (i) $\phi(x) = \log(1+e^{-x})$ (Logistic) with $g(x) = e^x$ and $k(x) = -\{e^x(1+e^x)\}^{-1}$; (ii) $\phi(x) = \log(\log(e^{-x}+e))$ (LogLog) with $g(x) = e^x$ and $k(x) = \{e^x(e^{x+1}+1)\log(e^{-x}+e)\}^{-1}$. We also experimented with $g(x) = e^{cx}$ for different values of c, and our results (not reported) were robust with respect to the choice of c. We also compare each of these algorithms to the commonly used LASSO (Friedman et al., 2010)/Multinomial Logistic Regression and SVM procedures. The SVM was trained with RBF kernel where the tuning parameter for the kernel function was chosen via the `sigest` function of `ksvm` library. The `sigest` procedure outputs three quantiles — 0.1, 0.5, 0.9 of the distribution of $\|X - X'\|_2^{-2}$ where X and X' are two predictors sampled from the matrix $\mathbf{X}$, and we take the mean of these quantiles as the tunning parameter in the RBF kernel for robustness. The fitting was performed with the `kernlab R` package implementation – `ksvm` which uses the "one-against-one"-approach to deal with multi-class problems see Hsu and Lin (2002) for example. The LASSO procedure with $\mathbf{X}$ being the predictors was based on an ℓ_1 penalized logistic regression through the `glmnet` implementation, and the tuning parameter was selected based on 5-fold CV. Throughout, the bag of classifiers for all boosting algorithms consisted of decision trees and multinomial logistic regressions weak learners, the predictors in each of which represented a subsample of the whole predictor set.

4.1 Simulation Studies

To distinguish the performance of our algorithm to the classical SAMME algorithm we chose a simulation setting with a complicated decision boundary. We borrow the idea of this simulation from the celebrated paper by Friedman et al. (2000).

The predictor matrix is generated as follows $\mathbf{X} \sim [N(0, \mathbb{I}_{10\times 10})]^{10}$. Next for each observation $\mathbf{X}_i$ it's ℓ_1 or ℓ_2 norms were calculated. The class assignment was then performed

based on:

$$C_i = \sum_{j=1}^{n} j\mathbb{1}(r_j \le \|\mathbf{X}_i\|_\ell < r_{j+1}),$$

where $0 = r_1 \le r_2 \le \ldots \le r_{n+1} = \infty$, $\ell \in \{1, 2\}$ and for $\mathbf{X} \in \mathbb{R}^p$ we have $\|\mathbf{X}\|_\ell := (\sum_{i=1}^{p} |X_i|^\ell)^{1/\ell}$. The thresholds r_j were selected in such a way so that each class got an approximately equal number of observations. Geometrically the classes $C = \{1, \ldots, k\}$ for $k < n$ were observations belonging to a 10 dimensional ℓ_1 or ℓ_2 sphere. Moreover, the more cutoffs r_j, the more classes the problem has, and hence the more complicated classification becomes.

We compared 5 algorithms – the SAMME, our algorithm with the logistic and loglog losses, the SVM and the LASSO logistic regression algorithm. Each simulation, 3200 observations were generated, 200 of which were used for training and 3000 were left out for testing purposes. We performed in total of 500 simulation for each scenario. The results are summarized in the tables 1 and 2 below:

Table 1: Percent misclassification for ℓ_1 Spheres

	$n = 3$	$n = 4$	$n = 5$	$n = 6$
SVM	17.9% (0.46)	29.4% (0.45)	37.6% (0.81)	46.6% (1.21)
LASSO	7.8% (11.6)	27.6% (5.00)	30.4% (20.0)	43.9% (15.9)
SAMME	6.9% (0.31)	10.7% (0.20)	16.4%(0.33)	28.4% (0.34)
Logistic	6.5% (22.6)	10.0% (16.4)	15.1% (22.3)	25.1% (23.6)
LogLog	6.6% (32.7)	9.8% (24.0)	15.0% (33.5)	25.2% (34.5)

Table 2: Percent misclassification for ℓ_2 Spheres

	$n = 3$	$n = 4$	$n = 5$	$n = 6$
SVM	17.0% (0.44)	28.0% (0.78)	36.6% (0.83)	44.1% (1.15)
LASSO	8.7% (11.8)	26.7% (5.70)	31.9% (19.7)	45.0% (14.6)
SAMME	8.2% (0.33)	12.9% (0.46)	19.7% (0.30)	28.9% (0.29)
Logistic	7.5% (23.4)	11.7% (25.3)	17.3% (24.6)	25.4% (23.5)
LogLog	7.7% (33.7)	11.5% (35.6)	17.9% (35.7)	25.8% (35.7)

As we can see our boosting based procedures dominated in all scenarios with the difference becoming more pronounced the more number of classes we have. We observed that our algorithms seem to combine classifiers from the bag more efficiently as compared to the SAMME algorithm in this scenario, especially so when the number of classes was large.

4.2 Experiments with UCI Benchmark Datasets

To present a more comprehensive assessment of our procedure, we analyzed 5 datasets from the UCI machine learning repository. The summaries of the characteristics of the datasets we used can be found in Table 3.

We compare results from SAMME, LogLog and Logistic from our proposed methods, SVM and a multinomial logistic regression (referred to as MLogisticReg). The MLogisticReg

9. Originally 10 classes; Extreme classes were grouped together for balanced class counts.

Table 3: Summary of datasets

	Dataset	Train Set / Test Set Size	# Features	# Classes
IS	Image Segmentation	210/2100	19	7
LED	Led Display	200/5000	7	10
RWQ[9]	Red Wine Quality	700/899	11	4
SE	Seeds	100/110	7	3
EC	Ecoli	100/236	7	8

was used instead of the LASSO procedure, as these datasets have relatively few features, and the LASSO procedure generally performs worse than the standard. As before all boosting algorithms were ran for 50 iterations.

For the analysis of all 5 datasets, the bag of weak learners we used consisted of decision trees and multinomial logistic regressions with subsampled prefixed number of predictors.

Table 4: Percent misclassifications (run time in seconds)

	IS	RWQ	SE	LED	EC
SVM	16.19% (0.44)	29.37% (0.28)	8.18% (0.08)	28.64% (1.26)	37.28% (0.27)
MLogisticReg	10.57% (0.08)	28.59% (0.09)	5.45% (0.04)	28.34% (0.06)	41.52% (0.04)
SAMME	5.09% (0.17)	30.14% (0.74)	4.55% (0.02)	29.36% (0.04)	22.45% (0.02)
Logistic	5.00% (13.6)	27.36% (35.3)	3.64% (1.61)	27.02% (1.53)	19.49% (1.22)
LogLog	4.95% (16.9)	28.36% (48.8)	3.64% (1.30)	27.00% (2.01)	19.49% (1.45)

We observe that in nearly all cases the newly suggested procedure performed better than the SAMME procedure, which typically converged too fast failing to include enough classifiers. The misclassifications rates of both the Logistic and LogLog based boosting algorithms were also better compared to the SVM and logistic regression in all 4 datasets.

4.3 Summary of Simulation Results and Benchmark Data Analyses

The results provided above demonstrate that the suggested algorithms can be used in successfully practice, and can outperform existing algorithms in many cases. The proposed boosting algorithm with LogLog loss appears to perform well across all settings considered with respect to classification accuracy. The robustness and gained accuracy of our proposed procedures come at the price of being more computationally intensive than the competitors. The slow speed is mostly due to the optimization with respect to β (see Algorithm 1) required at each iteration for each classifier in the bag. The SAMME algorithm avoids this search as an explicit solution to its corresponding optimization problem exists.

Further computational complexity in boosting procedures comes from two sources: one of them is the complexity of the classifiers in the bag, and the other is the number of classifiers in the bag. For massive datasets (with both high numbers of observations and of predictors) the optimization involved in our algorithm might be prohibitive if one decides to include a lot of weak learners. The methods we suggest could work in a reasonable time,

if one opts for sampling not too many classifiers. A relaxation of the procedure might be warranted in cases where a lot of classifiers will be used, and such relaxed algorithms are left for future work.

4.4 Application to an EMR Study of Diabetic Neuropathy

To illustrate our proposed generic boosting algorithm and demonstrate the advantage of having multiple losses, we apply our procedures to an electronic medical record (EMR) study, conducted at the Partners Healthcare, aiming to identify patients with different subtypes of diabetic neuropathy. Diabetic neuropathy (DN), a serious complication of diabetes, is the most common neuropathy in industrialized countries with an estimate of about 20-30 million people affected by symptomatic DN worldwide (Said, 2007). Increasing rates of obesity and type 2 diabetes could potentially double the number of affected individuals by the year 2030. The prevalence of DN also increases with time and poor glycemic control (Martin et al., 2006). Although many types of neuropathy can be associated with diabetes, the most common type is diabetic polyneuropathy and pain can develop as a symptom of diabetic polyneuropathy (Thomas and Eliasson, 1984; Galer et al., 2000). Pain in the feet and legs was reported to occur in 11.6% of insulin dependent diabetics and 32.1% of noninsulin dependent diabetics (Ziegler et al., 1992). Unfortunately, risk factors for developing painful diabetic neuropathy (pDN) are generally poorly understood. pDN has been reported as more prevalent in patients with type 2 diabetes and women (Abbott et al., 2011). Prior studies have also reported an association between family history and pDN, suggesting a potential genetic predisposition to pDN (Galer et al., 2000). To enable a genetic study of pDN and non-painful DN (npDN), an EMR study was performed to identify patients with these two subtypes of DN by investigators from the informatics for integrating biology to the bedside (i2b2), a National Center for Biomedical Computing based at Partners HealthCare (Murphy et al., 2006, 2010).

To identify such patients, we created a datamart compromising 20,000 patients in the Partners Healthcare with relevant ICD9 (International Classification of Diseases, version 9) codes. Two sources of information were utilized to classify patients' DN status and subtypes: (i) structured clinical data searchable in the EMR such as ICD9 codes; and (ii) variable identified using natural language processing (NLP) to identify medical concepts in narrative clinical notes. Algorithm development and validation was performed in a training set of 611 patients sampled from the datamart. To obtain the gold standard disease status for these patients, several neurologists performed chart reviews and classified them into no DN, pDN and npDN. The distribution of the classes was $64\%, 14\%, 22\%$ respectively. To train the classification algorithms, we included a total of 85 predictors most of which are NLP variables, counting mentions of medical concepts such as *"pain", "hypersensitivity"*, and *"diabetic neuropathy"*.

We trained boosting classification algorithms to classify these 3 disease classes. We used decision stumps as weak learners. They only have two nodes with the first node deciding between class C_1 vs C_2 and C_3 and the other node deciding between C_2 vs C_3, where $\{C_1, C_2, C_3\}$ is a permutation of {no DN,pDN,npDN}. In order to illustrate the algorithms we used 311 observations as a training set and the rest 300 patients we set off as a test set.

We report the percentage mis-classifications in Table 5. The boosting results show some improvement, as compared to standard methods. We can also see that the generic boosting algorithm performs slightly better than SAMME in this situation with both the logistic and the LogLog losses. It warrants further research whether picking richer tree structures as opposed to using stumps, would yield an even better performance on this dataset.

Table 5: Percent misclassifications (run time in seconds)

	% incorrect
SVM	43.67% (0.26)
LASSO	37.00% (15.8)
SAMME	33.33% (0.21)
Logistic	31.67% (9.55)
LogLog	31.67% (12.5)

5. Discussion

For multi-category classification problems, we described in this paper a class of loss functions that attain FC properties and provided theoretical justifications for how such loss functions can ultimately lead to optimal Bayes classifier. We extended the results to accommodate differential costs in misclassifying different classes. To approximate the minimizer of the empirical losses, we demonstrated that a natural iterative procedure can be used to derive generic boosting algorithms for any of the proposed losses. Numerical results suggest that non-convex losses such as LogLog could potentially lead to algorithms with better classification performance. Although the LogLog loss appears to perform well across different settings considered in the numerical studies, choosing an optimal loss for a given dataset warrants further research. Our preliminary studies (results not shown) suggest that procedures such as the cross-validation can potentially be used for loss selections.

Our proposed algorithm not only depends on the choice of ϕ but also the associated $g(\cdot)$ and $k(\cdot)$ functions as indicated in (6). We can think of g as a positive deformation of the real line and even with the same ϕ, changing g could also change the classifiers. Most existing boosting algorithms correspond to $g(x) = e^x$, in which case the constraint $\prod_{j=1}^n g(F_j) = 1$ simplifies to the commonly seen condition $\sum_j F_j = 0$. Moreover if ϕ is smooth and convex, one may let $k(x) = \dot{\phi}(x)/e^x$. Thus, under convexity, $H_\phi(x) = \dot{\phi}(x) = d\phi(x)/dx$ is an increasing function and ϕ is Fisher consistent in the traditional sense. We also saw, that even when ϕ is not convex, our suggested losses are Fisher consistent in the standard sense. Moreover, we argued that loss functions satisfying (6), can be used to recover the exact conditional probabilities. It would be interesting to develop adaptive boosting procedure where we use different g functions in the process of boosting adaptively. For example, in the suggested logistic loss boosting with $g(x) = e^{cx}$, we can adaptively select the parameter c, for better convergence results of the algorithm which will potentially result in a better classification results. We were provided a property of the limiting point of the algorithm in the case where $g = \exp$. Furthermore, we characterized when the problem has a minimum in the finite sample case under certain assumptions on ϕ. The resemblance of the proposed

generic boosting algorithm with coordinate descent, helped us to establish geometric rate of convergence in the convex loss function case. The consistency of the algorithm under conditions such as finite VC dimension of the classifier bag, warrants future research.

Acknowledgments

The authors would like to thank Alexander Rakhlin and three referees for their valuable suggestions and feedback, which led to improvements in the present manuscript. This research was partially supported by Research Grants NSF DMS1208771, NIH R01GM113242-01, NIH U54HG007963 and NIH RO1HL089778.

Appendix A. Proofs

Lemma A.1 *Assumption (6) implies that the function $\phi(g^{-1}(z))$ is continuously differentiable and convex for all $z \in g(S)$.*

Proof [Proof of Lemma A.1] Set $z := g(x), z' := g(x')$ in (6). When $x, x' \in S$ we have $z, z' \in g(S)$ and vice versa. Now (6) can be rewritten as:

$$\phi(g^{-1}(z)) - \phi(g^{-1}(z')) \geq (z - z')k(g^{-1}(z')). \tag{23}$$

Changing the roles of z and z' and using the fact that both $z, z' \in g(S)$ we obtain:

$$\phi(g^{-1}(z')) - \phi(g^{-1}(z)) \geq (z' - z)k(g^{-1}(z)).$$

The above two inequalities, combined with the fact that k and g^{-1} are increasing, give that for any $z \neq z', z, z' \in g(S)$ we have:

$$\min\{k(g^{-1}(z)), k(g^{-1}(z'))\} \leq \frac{\phi(g^{-1}(z')) - \phi(g^{-1}(z))}{z' - z} \leq \max\{k(g^{-1}(z)), k(g^{-1}(z'))\}. \tag{24}$$

By the continuity of k and g we have that the composition $k(g^{-1}(\cdot))$ is also continuous. Taking the limit $z' \to z$ in (24) shows that the function $\phi(g^{-1}(z))$ is differentiable on $g(S)$ with a continuous derivative equal to $k(g^{-1}(z))$. Now the convexity of $\phi(g^{-1}(z))$ follows from (23). ■

Proof [Proof of Theorem 2.1] To show that $H_\phi(\widehat{F}_j)w_j = \mathcal{C}$ for some $\mathcal{C} < 0$, define $\Omega = \{\mathbf{F} = (F_1, \ldots, F_n) : F_j \in S, j = 1, ..., n\}$, where recall that $S = \{z \in \mathbb{R} : k(z) \leq 0\}$. From (6),

$$\sum_{j=1}^{n} \phi(\widehat{F}_j)w_j \geq \sum_{j=1}^{n} \phi(F_j)w_j + \sum_{j=1}^{n} \{g(\widehat{F}_j) - g(F_j)\}k(F_j)w_j \quad \text{for any } \mathbf{F} \in \Omega. \tag{25}$$

Since $\widehat{\mathbf{F}}$ minimizes $\sum_{j=1}^{n} \phi(F_j)w_j$, (25) implies that

$$\sum_{j=1}^{n} g(F_j)k(F_j)w_j \geq \sum_{j=1}^{n} g(\widehat{F}_j)k(F_j)w_j \quad \text{for any } \mathbf{F} \in \Omega. \tag{26}$$

For any given constant $\mathcal{C} < 0$, let $\widetilde{F}_{\mathcal{C}j}$ be the solution to $g(\widehat{F}_j)k(\widetilde{F}_{\mathcal{C}j})w_j = \mathcal{C}$ or equivalently $\widetilde{F}_{\mathcal{C}j} = k^{-1}[\mathcal{C}/\{g(\widehat{F}_j)w_j\}]$. Obviously $\widetilde{\mathbf{F}}_{\mathcal{C}} \in \Omega$ for all $\mathcal{C} < 0$. We next show that there exists $\mathcal{C}_0 < 0$ such that $\prod_{j=1}^n g(\widetilde{F}_{\mathcal{C}_0 j}) = \prod_{j=1}^n g(k^{-1}[\mathcal{C}_0/\{g(\widehat{F}_j)w_j\}]) = 1$. Since g and k are continuous and strictly increasing functions, it suffices to show that $\prod_{j=1}^n g(\widetilde{F}_{0j}) > 1$ and $\prod_{j=1}^n g(\widetilde{F}_{\mathcal{C}j}) \le 1$ for some $\mathcal{C}$. Obviously $\prod_{j=1}^n g(\widetilde{F}_{0j}) > 1$ since $g\{k^{-1}(0)\} > g(0) = 1$. Now let $\mathcal{C}_1 = k(0)\max_j\{g(\widehat{F}_j)w_j\} < 0$. Then for all j, $\mathcal{C}_1/\{g(\widehat{F}_j)w_j\} \le k(0)$ and thus $g(k^{-1}[\mathcal{C}_1/\{g(\widehat{F}_j)w_j\}]) \le g(0) = 1$. Then by continuity of g and k, there exists $\mathcal{C}_0 \in [\mathcal{C}_1, 0)$ such that $\prod_{j=1}^n g(\widetilde{F}_{\mathcal{C}_0 j}) = 1$. Thus, the constructed $\widetilde{\mathbf{F}}_{\mathcal{C}_0}$ possesses several properties: (i) $g(\widehat{F}_j)k(\widetilde{F}_{\mathcal{C}_0 j})w_j = \mathcal{C}_0$; (ii) $\prod_{j=1}^n g(\widetilde{F}_{\mathcal{C}_0 j}) = 1$; and (iii) $k(\widetilde{F}_{\mathcal{C}_0 j}) < 0$ and hence $\widetilde{\mathbf{F}}_{\mathcal{C}_0} \in \Omega$. It then follows from the AM-GM inequality that

$$\sum_{j=1}^n g(\widetilde{F}_{\mathcal{C}_0 j})\{-k(\widetilde{F}_{\mathcal{C}_0 j})\}w_j \ge n\left[\prod_{j=1}^n g(\widetilde{F}_{\mathcal{C}_0 j})\{-k(\widetilde{F}_{\mathcal{C}_0 j})\}w_j\right]^{n^{-1}} = n\left[\prod_{j=1}^n g(\widehat{F}_j)\{-k(\widetilde{F}_{\mathcal{C}_0 j})\}w_j\right]^{n^{-1}} = -n\mathcal{C}_0,$$

where we used the fact that $\prod_{j=1}^n g(\widetilde{F}_{\mathcal{C}_0 j}) = \prod_{j=1}^n g(\widehat{F}_j) = 1$. This, together with (26), implies that

$$n\mathcal{C}_0 \ge \sum_{j=1}^n g(\widetilde{F}_{\mathcal{C}_0 j})k(\widetilde{F}_{\mathcal{C}_0 j})w_j \ge \sum_{j=1}^n g(\widehat{F}_j)k(\widetilde{F}_{\mathcal{C}_0 j})w_j = n\mathcal{C}_0$$

and hence $n\mathcal{C}_0 = \sum_{j=1}^n g(\widetilde{F}_{\mathcal{C}_0 j})k(\widetilde{F}_{\mathcal{C}_0 j})w_j$. Thus, the equality holds in the AM-GM inequality above, which also implies that $g(\widetilde{F}_{\mathcal{C}_0 j})k(\widetilde{F}_{\mathcal{C}_0 j})w_j = \mathcal{C}_0$. Since $g(\widehat{F}_j)k(\widetilde{F}_{\mathcal{C}_0 j})w_j = \mathcal{C}_0$, $k(\widetilde{F}_{\mathcal{C}_0 j}) \ne 0$ and g is strictly increasing, we have $g(\widehat{F}_j) = g(\widetilde{F}_{\mathcal{C}_0 j})$ and hence $\widehat{F}_j = \widetilde{F}_{\mathcal{C}_0 j}$. Therefore,

$$g(\widehat{F}_j)k(\widehat{F}_j)w_j = H_\phi(\widehat{F}_j)w_j = \mathcal{C}_0. \tag{27}$$

Obviously if $H_\phi(\cdot)$ is strictly monotone then $\widehat{F}_j = H_\phi^{-1}(\mathcal{C}_0/w_j)$ which is unique. ■

Proof [Proof of Proposition 2.2] The function ϕ is strictly decreasing on the set $S' := \{z : k(z) < 0\}$, as from (6) for any $x < x', x, x' \in S'$ we have:

$$\phi(x) - \phi(x') \ge (g(x) - g(x'))k(x') > 0.$$

Furthermore, it follows from Theorem 2.1, that $\widehat{F}_j \in S'$ since by (27) $k(\widehat{F}_j) < 0$ for all j. Next we show that if $w_\jmath > w_j$ we must have $\phi(\widehat{F}_\jmath) \le \phi(\widehat{F}_j)$. This observation follows since:

$$\phi(\widehat{F}_\jmath)w_\jmath + \phi(\widehat{F}_j)w_j \le \phi(\widehat{F}_\jmath)w_j + \phi(\widehat{F}_j)w_\jmath,$$

or else $\widehat{\mathbf{F}}$ cannot be a minimum of (9), as we can swap $\widehat{F}_\jmath$ and $\widehat{F}_j$ to obtain a strictly smaller value while still satisfying the constraint. Furthermore by Theorem 2.1, $w_\jmath \ne w_j$ implies that $\widehat{F}_\jmath \ne \widehat{F}_j$ because otherwise $H_\phi(\widehat{F}_\jmath) = H_\phi(\widehat{F}_j)$ and hence $w_\jmath = w_j$ by (10). Since ϕ is strictly decreasing on S' it also implies $\phi(\widehat{F}_\jmath) \ne \phi(\widehat{F}_j)$. Hence $w_\jmath > w_j$ implies $\phi(\widehat{F}_\jmath) < \phi(\widehat{F}_j)$. Finally, the last observation gives:

$$\operatorname*{argmin}_{j\in\{1,\ldots,n\}} \phi(\widehat{F}_j) = {}^{10} \operatorname*{argmax}_{j\in\{1,\ldots,n\}} w_j.$$

The fact that ϕ is decreasing on S' completes the proof. ■

Proof [Proof of Theorem 2.3] To show that a finite minimizer $\widehat{\mathbf{F}}$ exists, it suffices to show that $g(\widehat{F}_j)$ is finite and bounded away from 0, for $j = 1, ..., n$. To this end, we note that the condition (12) is equivalent to,

$$\lim_{x\downarrow 0} c_1\phi(g^{-1}(x)) + c_2\phi(g^{-1}(x^{-(n-1)})) = +\infty \quad \text{for all } c_1, c_2 > 0. \tag{28}$$

We next show that at the minimizer $\widehat{\mathbf{F}}$, $\widehat{m} = \min_j g(\widehat{F}_j) = g(\widehat{F}_{j^*})$ is bounded away from 0, where $j^* = \text{argmin}_j\, g(\widehat{F}_j)$. Since $1 = \prod_{j=1}^n g(\widehat{F}_j) \geq g(\widehat{F}_j)\widehat{m}^{n-1}$, we have $\widehat{F}_j \leq g^{-1}(\widehat{m}^{-(n-1)})$ for $j = 1, ..., n$. If ϕ is *decreasing* over $\mathbb{R}$, then

$$\begin{aligned}\phi(0)\sum_{j=1}^n w_j &\geq \sum_{j=1}^n \phi(\widehat{F}_j)w_j = w_{j^*}\phi\{g^{-1}(\widehat{m})\} + \sum_{j\neq j^*} w_j\phi(\widehat{F}_j)\\ &\geq w_{j^*}\phi\{g^{-1}(\widehat{m})\} + \sum_{j\neq j^*} w_j\phi\{g^{-1}(\widehat{m}^{-(n-1)})\}.\end{aligned}$$

From (28) with $c_1 = w_{j^*}$ and $c_2 = \sum_{j\neq j^*} w_j$, we conclude that $\widehat{m}$ must be bounded away from 0 since $\sum_{j=1}^n \phi(\widehat{F}_j)w_j \to \infty$ if $\widehat{m} \to 0$. Thus, there exists $m_0 > 0$ such that $\widehat{m} \geq m_0$ and consequently

$$0 < m_0 \leq g(\widehat{F}_j) \leq m_0^{-(n-1)} < \infty, \quad j = 1, ..., n.$$

Now, if ϕ is not decreasing on the whole $\mathbb{R}$, then there must exist $F^* < \infty$ such that $k(F^*) = 0$ since ϕ is strictly decreasing on $S' = \{z : k(z) < 0\}$.

Now we show that $\widehat{\mathbf{F}} \in \Omega \equiv \{\mathbf{F} = (F_1, \ldots, F_n) : F_j \in S, j = 1, \ldots, n\}$ as defined in Theorem 2.1. To this end, we note that ϕ is strictly decreasing on S' and $(-\infty, 0] \subset S'$. We next argue by contradiction that $\widehat{F}_j \in S$ or equivalently $\widehat{F}_j \leq F^*$ for all j. For any $F > F^*$, $\phi(F) - \phi(F^*) \geq \{g(F) - g(F^*)\}k(F^*) = 0$ by (6). Let $\mathcal{A} = \{j : \widehat{F}_j > F^*\}$ and $\widehat{F}_j^* = \mathbb{1}(j \in \mathcal{A})F^* + \mathbb{1}(j \notin \mathcal{A})\widehat{F}_j$. If $\mathcal{A}$ is not an empty set, then $\sum_{j=1}^n \phi(\widehat{F}_j^*)w_j \leq \sum_{j=1}^n \phi(\widehat{F}_j)w_j$ and $\prod_{j=1}^n g(\widehat{F}_j^*) < 1$. Since $g(F^*) > 1$, there must exist some $\widehat{F}_j^{**}$ with $F^* \geq \widehat{F}_j^{**} \geq \widehat{F}_j$ for $j \notin \mathcal{A}$ and $\widehat{F}_j^{**} = F^*$ for $j \in \mathcal{A}$ such that $\prod_{j=1}^n g(\widehat{F}_j^{**}) = 1$ and $F^* \geq \widehat{F}_j^{**} > \widehat{F}_j$ for some $j \notin \mathcal{A}$. Since ϕ is strictly decreasing on S, $\sum_{j=1}^n \phi(\widehat{F}_j^{**})w_j < \sum_{j=1}^n \phi(\widehat{F}_j^*)w_j \leq \sum_{j=1}^n \phi(\widehat{F}_j)w_j$, which contradicts that $\widehat{\mathbf{F}}$ is the minimum. Therefore, $\widehat{\mathbf{F}} \in \Omega$.

Hence $\widehat{F}_j \leq F^*$ and $g(\widehat{F}_j) \leq g(F^*) = m_1 \in (0, \infty)$. On the other hand, since $\prod_{j=1}^n g(\widehat{F}_j) = 1$, we have $g(\widehat{F}_j) \geq m_1^{-(n-1)}$ and thus $g(\widehat{F}_j)$ is also bounded away from 0 and finite.

Remark A.1 *As a useful remark we mention that the same argument shows that given any finite vector $\widehat{\mathbf{F}}$, the vectors $\mathbf{F}$ with $\sum_j \phi(F_j)w_j \leq \sum_j \phi(\widehat{F}_j)w_j$ are located on a compact set (provided that $\widehat{F}_j < F^*$ for all j in the second case).*

■

10. Recall that "=" should be interpreted as "⊆".

Lemma A.2 *Any loss function ϕ satisfying (6) with $g = \exp$, and either i. or ii. from Theorem 2.3 is classification calibrated in the two class case.*

Remark A.2 *Recall that a loss function ϕ is classification calibrated in the two class case if, for any point $w_1 + w_2 = 1$ with $w_1 \neq \frac{1}{2}$ and $w_1, w_2 > 0$, we have:*

$$\inf_{x \in \mathbb{R}} (w_1\phi(x) + w_2\phi(-x)) > \inf_{x : x(2w_1 - 1) \leq 0} (w_1\phi(x) + w_2\phi(-x)).$$

Proof [Proof of Lemma A.2] Denote the two (distinct) class probabilities with $w_1 + w_2 = 1$. Without loss of generality we distinguish two cases: $w_1 > w_2 > 0$ and $w_1 = 1, w_2 = 0$. First, consider the case when $w_1 > w_2 > 0$. Since the conditions of Theorem 2.3 hold, we know that the optimization problem (9) has a minimum, and hence by Proposition 2.2 we have that $\operatorname{argmax}_{j \in \{1,2\}} \widehat{F}_j \subseteq \{1\}$. Hence it follows that $\widehat{F}_1 > 0, \widehat{F}_2 < 0$ at the minimum. This implies that inequality in Remark A.2 is strict.

Next assume that $w_1 = 1, w_2 = 0$. This case is not covered by our results as we assume that the probabilities are bounded away from 0. As we argued earlier ϕ is strictly decreasing on the set S' and by assumption $(-\infty, 0] \subsetneq S'$. Thus by:

$$\widehat{\mathbf{F}} = \operatorname{argmin}_{\mathbf{F} : F_1 + F_2 = 0} w_1\phi(F_1) + w_2\phi(F_2) = \operatorname{argmin}_{\mathbf{F} : F_1 + F_2 = 0} \phi(F_1),$$

we must have $\widehat{F}_1 > 0$ and hence $\phi(0) > \phi(\widehat{F}_1)$. This finishes the proof. ■

Lemma A.3 *Let $\mathbf{F}^{(m)}$ be defined as in iteration (16) starting from $\mathbf{F}^{(0)} = 0$. Then we must have $\mathbf{F}^{(m)} \in \Omega$ for all m, where $\Omega = \{\mathbf{F} = (F_1, \ldots, F_n) : F_j \in S, j = 1, ..., n\}$.*

Proof [Proof of Lemma A.3] If $S = \mathbb{R}$ there is nothing to prove. Assume that there exists $F^* \in \mathbb{R}$ such that $k(F^*) = 0$. We show the statement by induction. By definition $\mathbf{F}^{(0)} \in \Omega$. Assume that $\mathbf{F}^{(m-1)} \in \Omega$ for some $m \geq 1$. We now show that $\mathbf{F}^{(m)} \in \Omega$. To arrive at a contradiction, assume the contrary. Let $\mathcal{A} = \{j : F_j^{(m)} > F^*\} \neq \varnothing$. Define $F_j^{*(m)} = \mathbb{1}(j \in \mathcal{A}) F_j^{(m-1)} + \mathbb{1}(j \notin \mathcal{A}) F_j^{(m)}$. Since $\mathbf{F}^{(m-1)} \in \Omega$, it follows that $\mathbf{F}^{*(m)} \in \Omega$ and $\prod_{j=1}^n g(\mathbf{F}_j^{*(m)}) < 1$. More importantly, observe that for all $j \in \mathcal{A}$ we have:

$$0 = (g(F_j^{(m-1)}) - g(F_j^{*(m)}))k(F_j^{*(m)})w_j > (g(F_j^{(m-1)}) - g(F_j^{(m)}))k(F_j^{(m)})w_j,$$

as $g(F_j^{(m-1)}) \leq g(F^*) < g(F_j^{(m)})$ and $k(F_j^{(m)}) > k(F^*) = 0$, and hence:

$$\sum_{j=1}^n (g(F_j^{(m-1)}) - g(F_j^{*(m)}))k(F_j^{*(m)})w_j > \sum_{j=1}^n (g(F_j^{(m-1)}) - g(F_j^{(m)}))k(F_j^{(m)})w_j.$$

Define the index set $\mathcal{B} = \{j : F_j^{*(m)} < F_j^{(m-1)}\}$. Since $\prod_{j=1}^n g(\mathbf{F}_j^{*(m)}) < 1$ and $\mathbf{F}^{(m-1)} \in \Omega$ it follows that $\mathcal{B}$ is not empty and $\mathcal{A} \cap \mathcal{B} = \varnothing$. Next for $\lambda \in [0, 1]$ define for all j:

$$F_j^{*(m),\lambda} := [\mathbb{1}(j \in \mathcal{A}) + \mathbb{1}(j \notin \mathcal{A})\mathbb{1}(j \notin \mathcal{B})]F_j^{*(m)} + \mathbb{1}(j \in \mathcal{B})((1 - \lambda)F_j^{*(m)} + \lambda F_j^{(m-1)}).$$

Note that when $\lambda = 0$, we have $F_j^{*(m),0} \equiv F_j^{*(m)}$. Now we show that for any $\lambda \in [0,1]$ the following inequality holds:

$$\sum_{j=1}^{n}(g(F_j^{(m-1)}) - g(F_j^{*(m),\lambda}))k(F_j^{*(m),\lambda})w_j \geq \sum_{j=1}^{n}(g(F_j^{(m-1)}) - g(F_j^{*(m)}))k(F_j^{*(m)})w_j. \quad (29)$$

For any $\lambda \in (0,1]$: $F_j^{*(m),\lambda} \neq F_j^{*(m),\lambda}$ iff $j \in \mathcal{B}$. Next note that for any j the function $(g(F_j^{(m-1)}) - g(x))k(x)w_j$ is an increasing function for $x \leq F_j^{(m-1)}$. The last two observations imply (29). Finally since $\prod_{j=1}^n g(F_j^{*(m),0}) = \prod_{j=1}^n g(F_j^{*(m)}) < 1$ and $\prod_{j=1}^n g(F_j^{*(m),1}) \geq \prod_{j=1}^n g(F_j^{(m-1)}) = 1$, by the continuity of g there exists a $\lambda \in (0,1]$ such that $\prod_{j=1}^n g(F_j^{*(m),\lambda}) = 1$. These facts and inequality (29) imply that $\mathbf{F}^{(m)}$ would not be a maximum in the iteration which is a contradiction. ■

Proof [Proof of Theorem 3.1] By construction we have that on the m^{th} iteration the value $\mathbf{F}^{(m)}$ satisfies $\prod_{j=1}^m g(F_j^{(m)}) = 1$, and Lemma A.3 guarantees that $\mathbf{F}^{(m)} \in \Omega$ for all m. Hence, since $F_j^{(m)}$ are viable values for $F_j^{(m+1)}$, the iteration also guarantees that:

$$\sum_{j=1}^{n}\{\phi(F_j^{(m)}) - \phi(F_j^{(m+1)})\}w_j \geq \sum_{j=1}^{n}\{g(F_j^{(m)}) - g(F_j^{(m+1)})\}k(F_j^{(m+1)})w_j \geq 0.$$

Now, from Remark A.1, $F_j^{(m+1)}$ lie on a compact set for all j, since for our starting point we have $\mathbf{F}_j^{(0)} \equiv 0 \in \Omega$. Therefore there must exist a subsequence $\{m_\ell, \ell = 1, ...\}$ such that $\mathbf{F}^{(m_\ell)}$ converges coordinate-wise on this subsequence, and denote with $\mathbf{F}^*$ its limit.

The function ϕ is continuous and hence we have that $\sum_{j=1}^n \phi(F_j^{(m_\ell)})w_j - \sum_{j=1}^n \phi(F_j^{(m_{\ell+1})})w_j \to 0$. However by the construction of our iteration, the sequences $\sum_{j=1}^n \phi(F_j^{(m_{\ell+1})})w_j$ are decreasing for all ℓ. Therefore we have that: $\sum_{j=1}^n \phi(F_j^{(m)})w_j - \sum_{j=1}^n \phi(F_j^{(m+1)})w_j \to 0$ holds for all m, not only on the subsequence. But this implies that $\sum_{j=1}^n (g(F_j^{(m)}) - g(F_j^{(m+1)}))k(F_j^{(m+1)})w_j \to 0$, which again by the construction is non-negative for all m. Take m_ℓ in place of m in the limit above, and let L be the set of all limit points $\lim_{\ell\to\infty} \mathbf{F}^{(m_\ell+1)}$. By our construction we have the following inequality holding for any point $\mathbf{F}^l \in L$:

$$0 = \sum_{j=1}^{n}\{g(F_j^*) - g(F_j^l)\}k(F_j^l)w_j \geq \sum_{j=1}^{n}\{g(F_j^*) - g(F_j)\}k(F_j)w_j, \quad (30)$$

for any $\mathbf{F} \in \Omega$ with $\prod_{j=1}^n g(F_j) = 1$. Just as in the proof of Theorem 2.1 select $\widetilde{\mathbf{F}}$ so that $g(F_j^*)k(\widetilde{F}_j)w_j = \mathcal{C}$ for all j for some $\mathcal{C} < 0$. By the AM-GM inequality we get:

$$\sum_{j=1}^{n} g(\widetilde{F}_j)\{-k(\widetilde{F}_j)\}w_j \geq n\left[\prod_{j=1}^{n} g(\widetilde{F}_j)\{-k(\widetilde{F}_j)\}w_j\right]^{n^{-1}} = n\left[\prod_{j=1}^{n} g(F_j^*)\{-k(\widetilde{F}_j)\}w_j\right]^{n^{-1}}$$
$$= \sum_{j=1}^{n} g(F_j^*)\{-k(\widetilde{F}_j)\}w_j = -n\mathcal{C}.$$

Now by (30) it follows that equality must be achieved in the preceding display, which implies that $g(F_j^*)k(\widetilde{F}_j)w_j = \mathcal{C} = g(\widetilde{F}_j)k(\widetilde{F}_j)w_j$ and yields $\widetilde{F}_j = F_j^*$ for all j. Hence $g(F_j^*)k(F_j^*)w_j = \mathcal{C}$ for all j.

Thus we showed that on subsequences the iteration converges to points satisfying the equality described above. We are left to show, that all these subsequences converge to the same point. Next, take equation (30). By what we showed it follows that for any $\mathbf{F}^l \in L$, we have that $g(F_j^l)k(F_j^l)w_j = \mathcal{C}^l$ for some $\mathcal{C}^l < 0$. Then we have:

$$\sum_{j=1}^{n} g(F_j^*)\{-k(F_j^l)\}w_j \geq n\left[\prod_{j=1}^{n} g(F_j^*)\{-k(F_j^l)\}w_j\right]^{n^{-1}} = n\left[\prod_{j=1}^{n} g(F_j^l)\{-k(F_j^l)\}w_j\right]^{n^{-1}}$$
$$= \sum_{j=1}^{n} g(F_j^l)\{-k(F_j^l)\}w_j = -n\mathcal{C}^l.$$

Equation (30) implies that the above inequality is in fact equality which shows that:

$$g(F_j^*)k(F_j^l) = g(F_j^l)k(F_j^l) \quad \text{for all } j.$$

Thus since $k(F_j^l) \neq 0$ (recall that all values on the iteration $\mathbf{F}^{(m)} \in \Omega$) we conclude that $g(F_j^*) = g(F_j^l)$, and hence $\mathbf{F}^* = \mathbf{F}^l$. This shows that for any converging subsequence m_ℓ the limiting value coincides with that of the sequence $m_\ell + 1$, which finishes our proof.

■

Proof [Proof of Proposition 3.2.] It is sufficient to show that for all $\mathbf{F} \in \mathcal{G}^*$ we have:

$$\sum_{i=1}^{N} \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)\dot{\phi}(\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i)) \geq 0.$$

The condition above is sufficient, because of the looping closure of $\mathcal{G}$. Writing the inequality for all "looped" versions of $\mathbf{F}$ and noting that they sum up to 0, gives us that the inequality is in fact an equality.

Note that with each iteration (18), we decrease the value of the target function. This can be seen by the following inequality:

$$\sum_{i=1}^{N} \phi(\mathbf{Y}_{C_i}^T \mathbf{F}^{(m-1)}(\mathbf{X}_i)) - \sum_{i=1}^{N} \phi(\mathbf{Y}_{C_i}^T \mathbf{F}^{(m)}(\mathbf{X}_i)) \geq \sum_{i=1}^{N} [\exp(-\beta \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)) - 1]\dot{\phi}(\mathbf{Y}_{C_i}^T \mathbf{F}^{(m)}(\mathbf{X}_i)) \geq 0,$$

where $\mathbf{F}^{(m)} = \mathbf{F}^{(m-1)} + \beta\mathbf{F}$. As a remark, the inequality in the preceding display holds, since ϕ is decreasing and thus by (6) we have $S = \mathbb{R}$.

Take a limiting point[11] $\mathbf{F}^{(\infty)}$ of iteration (18), where it is possible having coordinates of $\mathbf{F}^{(\infty)}(\mathbf{X}_i)$ equal to $\pm\infty$ for some i. Since ϕ is bounded from below, by our previous observation we have that for any $\beta \geq 0$ and $\mathbf{F} \in \mathcal{G}^*$:

$$\sum_{i=1}^{N} \phi(\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i)) - \sum_{i=1}^{N} \phi(\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i) + \beta \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)) \leq 0.$$

11. The existence of a limiting point is guaranteed as any sequence contains a monotone subsequence.

Let $\mathcal{A} = \{i : |\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i)| \neq \infty\}$. Then the latter inequality also implies that:

$$\sum_{i\in\mathcal{A}} \phi(\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i)) - \sum_{i\in\mathcal{A}} \phi(\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i) + \beta \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)) \leq 0.^{12}$$

Applying inequality (6) the above implies:

$$\sum_{i\in\mathcal{A}} [\exp(-\beta \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)) - 1] \dot{\phi}(\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i) + \beta \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)) \leq 0,$$

and after a Taylor expansion of the exponent, and division by $\beta \geq 0$ we get:

$$\begin{aligned}
&\sum_{i\in\mathcal{A}} -\mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i) \dot{\phi}(\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i)) \\
&+ \sum_{i\in\mathcal{A}} -\mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i) [\dot{\phi}(\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i) + \beta \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)) - \dot{\phi}(\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i))] \\
&+ O(\beta) \sum_{i\in\mathcal{A}} \dot{\phi}(\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i) + \beta \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)) \leq 0.
\end{aligned}$$

Letting $\beta \to 0$, by the continuity of $\dot{\phi}$ we get:

$$\sum_{i\in\mathcal{A}} \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i) \dot{\phi}(\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i)) \geq 0. \tag{31}$$

Next we argue that $\dot{\phi}(+\infty) = \lim_{x\to+\infty} \dot{\phi}(x) = 0$. As stated in the main text $\dot{\phi}(x) = k(x)e^x$. Let $K = \inf_{x\in\mathbb{R}} \phi(x)$. For any $\varepsilon > 0$, take a point x' such that $\phi(x') - K \leq \varepsilon$. Then for any $x \in \mathbb{R}$, by (6):

$$\varepsilon \geq \phi(x') - \phi(x) \geq (e^{x'} - e^x) k(x),$$

and thus $\varepsilon - e^{x'} k(x) \geq -\dot{\phi}(x) \geq 0$. Taking the limit $x \to +\infty$ and letting $\varepsilon \to 0$ shows that $\dot{\phi}(+\infty) = 0$.

Now consider two cases for ϕ. Suppose that ϕ is unbounded from above. We argue that $\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i) \neq -\infty$ for all i. Since we start form the point 0, and as we argued we are decreasing the target function we have that:

$$N\phi(0) \geq \sum_{i=1}^{N} \phi(\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i)) \geq \max_i \phi(\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i)) + (N-1)K,$$

and hence $\max_i \phi(\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i)) \leq N\phi(0) - (N-1)K$. Since ϕ is decreasing and unbounded from above it follows that $\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i) \neq -\infty$ for all i. In the second case suppose that ϕ is bounded from above, and let $M = \sup_{x\in\mathbb{R}} \phi(x)$. We show that $\dot{\phi}(-\infty) = 0$. For any $\varepsilon > 0$ take x so that $\varepsilon \geq M - \phi(x)$. Applying (6) for any $x' \in \mathbb{R}$ yields:

$$\varepsilon \geq \phi(x') - \phi(x) \geq (e^{x'} - e^x) k(x).$$

This gives $\varepsilon - e^{x'} k(x) \geq -\dot{\phi}(x) \geq 0$. Taking $x' \to -\infty$ gives that $\varepsilon \geq -\dot{\phi}(x) \geq 0$ for any x such that $\varepsilon \geq M - \phi(x)$. Since ϕ is decreasing we are allowed to take the limit $x \to -\infty$, and

12. Observe that since ϕ is bounded from below the values of ϕ at infinite points of the iteration i.e. $\phi(\pm\infty)$ have to be bounded.

taking $\varepsilon \to 0$ shows that $\dot{\phi}(-\infty) = 0$. In any case, all of the above arguments imply that we can substitute $\mathcal{A}$ in (31) to the whole index set $\{1, \ldots, N\}$, to finally conclude:

$$\sum_{i=1}^{N} \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i) \dot{\phi}(\mathbf{Y}_{C_i}^T \mathbf{F}^{(\infty)}(\mathbf{X}_i)) \geq 0,$$

for all $\mathbf{F} \in \mathcal{G}^*$. As argued in the beginning the looping closure gives us that in fact the "$\geq$" can be replaced with "=". This concludes the proof.

■

Proof [Proof of Proposition 3.3] First consider the case when $I = N$. Denote with $e_1^1, e_1^2, \ldots, e_1^N$ the positive coordinates of $\mathbf{e}_1$. For any vector $\mathbf{v}$ which is a solution to $\mathbf{D}^T\boldsymbol{\alpha} = \mathbf{v}$ we must have $\sum_{i=1}^N e_1^i v_i = 0$. Clearly then, if $\mathbf{v}$ is non-zero some of the v_i need to be negative. Let $l = \min v_i$. We know that $l < 0$. This immediately implies an upper bound on the maximal positive v_i — $\max_i v_i \leq |l|[\sum_{i=1}^N e_1^i / \min_j e_1^j - 1]$. We now show that $|l|$ is bounded, for all vectors $\mathbf{v}$ such that $\sum_{i=1}^N \phi(v_i) \leq N\phi(0)$. Note that $\sum_{i=1}^N \phi(v_i) \geq \phi(l) + (N-1)\phi(|l|[\sum_{i=1}^N e_1^i / \min_j e_1^j - 1])$, and thus (22) gives that $|l|$ is bounded. This in turn shows that all v_i are bounded as well.

We next consider the case where $I < N$. Without loss of generality, upon rearrangement, we can assume that the positive coordinates of $\mathbf{e}_1$ are located at the first I places. Let $\mathbf{E}_{N\times s} = (\mathbf{e}_1, \ldots, \mathbf{e}_s)$. Let $\widetilde{\mathbf{E}}_{(N-I)\times(s-1)}$ denote the sub-matrix corresponding to the 0 entries of $\mathbf{e}_1$ (excluding $\mathbf{e}_1$) (see (32) for a visualization). Note that the matrix $\widetilde{\mathbf{E}}$ cannot be of full column rank, because otherwise we would have that a vector with positive coordinates is inside the column space which is a contradiction (we can always scale it by a small number and add to $\mathbf{e}_1$). Thus we can eliminate all extra columns that do not contribute to the rank of $\widetilde{\mathbf{E}}$, by doing a linear manipulation on the columns of the whole matrix $\mathbf{E}$ (see (32)). In doing so, we can eliminate extra columns of the matrix $\widetilde{\mathbf{E}}$ so that we end up with a $\widetilde{\mathbf{E}}$ matrix where the number of non-zero columns matches the rank, and some columns of $\mathbf{E}$ have 0 coefficients on the lower part. Here, observe that the columns of $\mathbf{E}$ with 0 sub-columns in $\widetilde{\mathbf{E}}$, are part of the space $\text{row}(\mathbf{D}_-)^\perp$, where $\mathbf{D}_-$ corresponds to the matrix $\mathbf{D}$ with observations corresponding to 0's of $\mathbf{e}_1$ removed.

We next note that if we discard the observations corresponding to 0 coordinates in e_1, and optimize the problem on the rest of the observations we will obtain some optimal solution $\mathbf{v} = (\hat{v}_1, \ldots, \hat{v}_I)^T$, the entries of which are bounded as argued in the first case. We next show that we can populate the vector $\mathbf{v}$ with positive numbers $p_1, \ldots, p_{N-I}$ to $\boldsymbol{\nu} = (\hat{v}_1, \ldots, \hat{v}_I, p_1, \ldots, p_{N-I})^T$, so that $\boldsymbol{\nu}$ is "perpendicular" to the matrix $\mathbf{E}$ (i.e. $\mathbf{E}^T\boldsymbol{\nu} = 0$), and thus can be written in the form $\mathbf{D}^T\boldsymbol{\alpha}$. Moreover, we will show that $p_1, \ldots, p_{N-I}$, can become arbitrarily large, which will complete the proof.

$$\mathbf{E} = \begin{array}{c} \\ I \\ \\ N-I \end{array} \begin{array}{|c|c|} \hline \mathbf{e}_1 & \mathbf{e}_2 \ \ldots \ \mathbf{e}_s \\ \hline e_1^1 \\ \vdots \\ e_1^I & \\ \hline 0 \\ \vdots \\ 0 & \widetilde{\mathbf{E}} \\ \hline \end{array} \quad \rightarrow \quad \begin{array}{|c|c|c|} \hline \mathbf{e}_1 & \ldots \ \widetilde{\mathbf{e}}_{l+1} & \widetilde{\mathbf{e}}_{l+2} \ \ldots \ \widetilde{\mathbf{e}}_s \\ \hline e_1^1 \\ \vdots & \mathbf{G} & \\ e_1^I & & \\ \hline 0 & & 0 \ \ldots \ 0 \\ \vdots & \widetilde{\widetilde{\mathbf{E}}} & \vdots \ \ldots \ \vdots \\ 0 & & 0 \ \ldots \ 0 \\ \hline \end{array} \tag{32}$$

Note that the only part of the matrix $\mathbf{E}$ that would be potentially non-zero upon multiplication by $\boldsymbol{\nu}$ would be the part corresponding to the non-zero parts of $\widetilde{\mathbf{E}}$, because as we argued earlier the columns of $\mathbf{E}$ with 0 sub-columns in $\widetilde{\mathbf{E}}$ belong to $\text{row}(\mathbf{D}_-)^\perp$ and on the other hand $\mathbf{v} \in \text{row}(\mathbf{D}_-)$. Denote with $\widetilde{\widetilde{\mathbf{E}}}_{(N-I)\times l}$ the full-rank sub-matrix of $\widetilde{\mathbf{E}}$, where l is the rank of $\widetilde{\mathbf{E}}$, and let $\mathbf{G}_{I\times l}$ be the sub-matrix of $\mathbf{E}$ above $\widetilde{\widetilde{\mathbf{E}}}$ (see (32)). Clearly $l < N - I$ as otherwise there is a positive vector in the column space, and we argued previously that would be a contradiction with the maximality property of $\mathbf{e}_1$. We need to find a positive vector $\mathbf{p}$ such that $(\widetilde{\widetilde{\mathbf{E}}}_{(N-I)\times l})^T \mathbf{p}_{(N-I)\times 1} = -(\mathbf{G}_{I\times l})^T \mathbf{v}_{I\times 1} = \mathbf{K}_{l\times 1}$. Therefore the proof will be completed, if we can find arbitrary large positive vectors $\mathbf{p}$ solving the system $\widetilde{\widetilde{\mathbf{E}}}^T \mathbf{p} = \mathbf{K}$, where $l < N - I$ and $\widetilde{\widetilde{\mathbf{E}}}^T$ has the property that any non-zero linear combination of its rows results into a vector with at least one positive and one negative entry.

Since $\widetilde{\widetilde{\mathbf{E}}}$ is full-rank and $l < N - I$, the linear system $\widetilde{\widetilde{\mathbf{E}}}^T \mathbf{p} = \mathbf{K}$ has a solution. Consider the homogeneous system $\widetilde{\widetilde{\mathbf{E}}}^T \mathbf{p} = 0$. We will show that the homogeneous equation admits arbitrary large positive solutions, which would complete the proof. Fix the value of the i^{th} parameter to be 1. The system then becomes $\widetilde{\widetilde{\mathbf{E}}}^T_{-i} \mathbf{p}_{-i} = -\widetilde{\widetilde{\mathbf{e}}}_i$, where by indexing with $-i$ we mean removing the i^{th} column or element and $\widetilde{\widetilde{\mathbf{e}}}_i$ is the i^{th} column of $\widetilde{\widetilde{\mathbf{E}}}^T$. Next we apply Farkas's lemma to show that the last equation has a non-negative solution. Assume that there is a vector $\mathbf{y}_{l\times 1}$ such that $\widetilde{\widetilde{\mathbf{E}}}_{-i}\mathbf{y} \geq 0$ (coordinate-wise) and $-\widetilde{\widetilde{\mathbf{e}}}_i^T \mathbf{y} < 0$. This is clearly a violation with the property that $\widetilde{\widetilde{\mathbf{E}}}$ satisfies. Therefore by Farkas's lemma the equation $\widetilde{\widetilde{\mathbf{E}}}^T_{-i} \mathbf{p}_{-i} = -\widetilde{\widetilde{\mathbf{e}}}_i$ has a non-negative solution. Since we can achieve this for any index i, averaging these solutions yields a positive solution to the homogeneous system $\widetilde{\widetilde{\mathbf{E}}}^T \mathbf{p} = 0$, and thus this system admits arbitrarily large positive solutions.

■

Proof [Proof of Theorem 3.4] Without loss of generality for the purposes of the proof we will consider $\mathcal{C}_+ = 1$ and $\mathcal{C}_- = -1/(n-1)$ (it's equivalent to rescaling the β in the iteration).

By the iteration's construction we know:

$$\varepsilon_m - \varepsilon_{m+1} \geq \max_{\beta\geq 0, \mathbf{F}\in\mathcal{G}^*} \sum_{i=1}^{N} \{e^{-\beta \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)} - 1\} \dot{\phi}\{\mathbf{Y}_{C_i}^T \mathbf{F}^{(m)}(\mathbf{X}_i) + \beta \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)\}.$$

Note that we have the following simple inequality holding for $\exp(-x) \leq 1 - x + x^2$ for values of $-1/2 \leq x \leq 1/2$. Since $|\mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)| \leq \mathcal{C}_+ = 1$ and ϕ is decreasing, for values of $0 \leq \beta \leq 1/2$ we have that:

$$\begin{aligned}
&\sum_{i=1}^{N} \{e^{-\beta \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)} - 1\} \dot{\phi}\{\mathbf{Y}_{C_i}^T \mathbf{F}^{(m)}(\mathbf{X}_i) + \beta \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)\} \\
&\geq -\sum_{i=1}^{N} (\beta \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i) - \beta^2) \dot{\phi}\{\mathbf{Y}_{C_i}^T \mathbf{F}^{(m)}(\mathbf{X}_i) + \beta \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)\}.
\end{aligned}$$

Let L denote the Lipschitz constant of $\dot{\phi}$ on the set $\mathcal{S}$. Consequently we have:

$$-\sum_{i=1}^{N}(\beta\mathbf{Y}_{C_i}^T\mathbf{F}(\mathbf{X}_i)-\beta^2)\dot{\phi}\{\mathbf{Y}_{C_i}^T\mathbf{F}^{(m)}(\mathbf{X}_i)+\beta\mathbf{Y}_{C_i}^T\mathbf{F}(\mathbf{X}_i)\}=\sum_{i=1}^{N}-(\beta\mathbf{Y}_{C_i}^T\mathbf{F}(\mathbf{X}_i)-\beta^2)\dot{\phi}\{\mathbf{Y}_{C_i}^T\mathbf{F}^{(m)}(\mathbf{X}_i)\}$$
$$-\sum_{i=1}^{N}(\beta\mathbf{Y}_{C_i}^T\mathbf{F}(\mathbf{X}_i)-\beta^2)[\dot{\phi}\{\mathbf{Y}_{C_i}^T\mathbf{F}^{(m)}(\mathbf{X}_i)+\beta\mathbf{Y}_{C_i}^T\mathbf{F}(\mathbf{X}_i)\}-\dot{\phi}\{\mathbf{Y}_{C_i}^T\mathbf{F}^{(m)}(\mathbf{X}_i)\}]\geq$$
$$\sum_{i=1}^{N}-(\beta\mathbf{Y}_{C_i}^T\mathbf{F}(\mathbf{X}_i)-\beta^2)\dot{\phi}\{\mathbf{Y}_{C_i}^T\mathbf{F}^{(m)}(\mathbf{X}_i)\}-L\left|\beta\mathbf{Y}_{C_i}^T\mathbf{F}(\mathbf{X}_i)-\beta^2\right|\left|\beta\mathbf{Y}_{C_i}^T\mathbf{F}(\mathbf{X}_i)\right|\geq$$
$$\sum_{i=1}^{N}-(\beta\mathbf{Y}_{C_i}^T\mathbf{F}(\mathbf{X}_i)-\beta^2)\dot{\phi}\{\mathbf{Y}_{C_i}^T\mathbf{F}^{(m)}(\mathbf{X}_i)\}-\frac{3}{2}LN\beta^2.$$

Thus we have established:

$$\varepsilon_m-\varepsilon_{m+1}\geq\beta\left(\sum_{i=1}^{N}(-\mathbf{Y}_{C_i}^T\mathbf{F}(\mathbf{X}_i)+\beta)\dot{\phi}\{\mathbf{Y}_{C_i}^T\mathbf{F}^{(m)}(\mathbf{X}_i)\}-\frac{3}{2}LN\beta\right),$$

for any $0\leq\beta\leq 1/2$. We select β so that we maximize the RHS in the expression above. It turns out that this happens for:

$$\beta_0=\frac{\frac{1}{2}\sum_{i=1}^{N}\mathbf{Y}_{C_i}^T\mathbf{F}(\mathbf{X}_i)\dot{\phi}\{\mathbf{Y}_{C_i}^T\mathbf{F}^{(m)}(\mathbf{X}_i)\}}{-\frac{3}{2}LN+\sum_{i=1}^{N}\dot{\phi}\{\mathbf{Y}_{C_i}^T\mathbf{F}^{(m)}(\mathbf{X}_i)\}}.$$

Since $\dot{\phi}$ is always negative and as we mentioned $\left|\mathbf{Y}_{C_i}^T\mathbf{F}(\mathbf{X}_i)\right|\leq 1$, provided that the numerator is ≤ 0, we have that $0\leq\beta_0\leq 1/2$. Then we would have:

$$\varepsilon_m-\varepsilon_{m+1}\geq-\frac{1}{2}\beta_0\sum_{i=1}^{N}\mathbf{Y}_{C_i}^T\mathbf{F}(\mathbf{X}_i)\dot{\phi}\{\mathbf{Y}_{C_i}^T\mathbf{F}^{(m)}(\mathbf{X}_i)\}. \tag{33}$$

Next we show that there exists a classifier, such that the above expression is strictly positive, which will also ensure that $0\leq\beta_0\leq 1/2$ is in the correct range. Denote with B the total number of classifiers in the bag. Consider the representation $\mathbf{F}^*(\cdot)-\mathbf{F}^{(m)}(\cdot)=\sum_{j=1}^{B}\alpha_j\mathbf{F}_j(\cdot)$. Here the $\boldsymbol{\alpha}$ vector is any vector that yields a correct representation (note that we will have many possible $\boldsymbol{\alpha}$ vectors, in the case when $B>N$).

By convexity of ϕ we have:

$$-\varepsilon_m=\sum_{i=1}^{N}\phi(\mathbf{Y}_{C_i}^T\mathbf{F}^*(\mathbf{X}_i))-\phi(\mathbf{Y}_{C_i}^T\mathbf{F}^{(m)}(\mathbf{X}_i))\geq\sum_{i=1}^{N}[\mathbf{Y}_{C_i}^T\mathbf{F}^*(\mathbf{X}_i)-\mathbf{Y}_{C_i}^T\mathbf{F}^{(m)}(\mathbf{X}_i)]\dot{\phi}(\mathbf{Y}_{C_i}^T\mathbf{F}^{(m)}(\mathbf{X}_i))$$
$$=\sum_{j=1}^{B}\sum_{i=1}^{N}\alpha_j\mathbf{Y}_{C_i}^T\mathbf{F}_j(\mathbf{X}_i)\dot{\phi}(\mathbf{Y}_{C_i}^T\mathbf{F}^{(m)}(\mathbf{X}_i)).$$

By the pigeonhole principle it is clear that there exists an index $j\in\{1,\ldots,B\}$ such that:

$$\frac{\varepsilon_m}{B\max_j|\alpha_j|}\leq\frac{\varepsilon_m}{B|\alpha_j|}\leq-\operatorname{sign}(\alpha_j)\sum_{i=1}^{N}\mathbf{Y}_{C_i}^T\mathbf{F}_j(\mathbf{X}_i)\dot{\phi}(\mathbf{Y}_{C_i}^T\mathbf{F}^{(m)}(\mathbf{X}_i)).$$

Now if $\text{sign}(\alpha_j) = 1$ we already have a "decent" lower bound. Otherwise if $\text{sign}(\alpha_j) = -1$, using the fact that the loop closed classifiers wrt to $\mathbf{F}_j$ sum up to 0, we can claim that for one of the looped classifiers $\mathbf{F}_j^l$ we would have a bound:

$$\frac{\varepsilon_m}{B(n-1)\max_j|\alpha_j|} \leq -\sum_{i=1}^{N} \mathbf{Y}_{C_i}^T \mathbf{F}_j^l(\mathbf{X}_i)\dot{\phi}(\mathbf{Y}_{C_i}^T \mathbf{F}^{(m)}(\mathbf{X}_i)).$$

So that in both cases we established the existence of a classifier such that $\mathbf{F} \in \mathcal{G}^*$ and:

$$\frac{\varepsilon_m}{B(n-1)\max_j|\alpha_j|} \leq -\sum_{i=1}^{N} \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)\dot{\phi}(\mathbf{Y}_{C_i}^T \mathbf{F}^{(m)}(\mathbf{X}_i))$$

We then know from (33) that:

$$\begin{aligned}\varepsilon_m - \varepsilon_{m+1} &\geq -\frac{1}{2}\beta_0 \sum_{i=1}^{N} \mathbf{Y}_{C_i}^T \mathbf{F}(\mathbf{X}_i)\dot{\phi}\{\mathbf{Y}_{C_i}^T \mathbf{F}^{(m)}(\mathbf{X}_i)\} \\ &\geq \frac{1}{4}\frac{\varepsilon_m^2}{B^2(n-1)^2\max_j \alpha_j^2(\frac{3}{2}LN - \sum_{i=1}^{N}\dot{\phi}\{\mathbf{Y}_{C_i}^T \mathbf{F}^{(m)}(\mathbf{X}_i)\})}.\end{aligned}$$

Notice that the derivative is bounded on the set $\mathcal{S}$ and therefore collapsing all constants above into one constant say T we get the following:

$$\varepsilon_m - \varepsilon_{m+1} \geq \frac{\varepsilon_m^2}{T\max_j \alpha_j^2}.$$

Here T depends on the number of classifiers, number of classes, and the bound on the first derivative $\dot{\phi}$ on the set $\mathcal{S}$. We will proceed to bound the $\max_j \alpha_j^2$ for some of the representations from above.

Because on the set $\mathcal{S}$, ϕ is also strongly convex (with a constant say l), we have the following:

$$\begin{aligned}\varepsilon_m &= \sum_{i=1}^{N} \phi(\mathbf{Y}_{C_i}^T \mathbf{F}^{(m)}(\mathbf{X}_i)) - \sum_{i=1}^{N} \phi(\mathbf{Y}_{C_i}^T \mathbf{F}^*(\mathbf{X}_i)) \\ &\geq \sum_{i=1}^{N} [\mathbf{Y}_{C_i}^T \mathbf{F}^{(m)}(\mathbf{X}_i) - \mathbf{Y}_{C_i}^T \mathbf{F}^*(\mathbf{X}_i)]\dot{\phi}(\mathbf{Y}_{C_i}^T \mathbf{F}^*(\mathbf{X}_i)) \\ &+ l\sum_{i=1}^{N}\left(\sum_{j=1}^{B} \alpha_j \mathbf{Y}_{C_i}^T \mathbf{F}_j(\mathbf{X}_i)\right)^2.\end{aligned}$$

The expression $\sum_{i=1}^{N}[\mathbf{Y}_{C_i}^T \mathbf{F}^{(m)}(\mathbf{X}_i) - \mathbf{Y}_{C_i}^T \mathbf{F}^*(\mathbf{X}_i)]\dot{\phi}(\mathbf{Y}_{C_i}^T \mathbf{F}^*(X_i))$ is 0, as $\mathbf{F}^*$ is the minimum, ϕ is convex and the classifier bag is closed under looping. Let $\mathbf{D} = \{\mathbf{Y}_{C_i}^T \mathbf{F}_j(\mathbf{X}_i)\}_{j,i}$ is the $B \times N$ matrix, each entry of which is either $\mathcal{C}_+$ or $\mathcal{C}_-$. Let the rank of $\mathbf{D}$ is $r \leq \min(N, B)$. We then have $\sum_{i=1}^{N}\left(\sum_{j=1}^{B} \alpha_j \mathbf{Y}_{C_i}^T \mathbf{F}_j(\mathbf{X}_i)\right)^2 = \boldsymbol{\alpha}^T \mathbf{D}\mathbf{D}^T \boldsymbol{\alpha}$. Since, all the bounds above are true for any of the $\boldsymbol{\alpha}$ representations, we could have picked the representation corresponding to the $r \times N$ sub matrix of $\mathbf{D}$, $\mathbf{D}_r$ with rank r for which the smallest eigenvalue of $\mathbf{D}_r\mathbf{D}_r^T$

is the largest. Let this eigenvalue be $\lambda_r > 0$. For this eigenvalue and this choice of $\boldsymbol{\alpha}$ we clearly have $\boldsymbol{\alpha}^T\mathbf{D}\mathbf{D}^T\boldsymbol{\alpha} = \boldsymbol{\alpha}^T\mathbf{D}_r\mathbf{D}_r^T\boldsymbol{\alpha} \geq \lambda_r\|\boldsymbol{\alpha}\|_2^2 \geq \lambda_r \max_j \alpha_j^2$. (in the second equality we abuse notation deleting zeros from the $\boldsymbol{\alpha}$). Consequently we get:

$$\varepsilon_m \geq l\lambda_r \max_j \alpha_j^2.$$

Thus:

$$\begin{aligned}\varepsilon_{m+1} &\leq \varepsilon_m - \frac{\varepsilon_m^2}{T\max_j \alpha_j^2} \\ &\leq \varepsilon_m\left(1 - \frac{l\lambda_r}{T}\right).\end{aligned}$$

Since both $\varepsilon_{m+1}, \varepsilon_m \geq 0$ we must have $1 - \frac{l\lambda_r}{T} \geq 0$. Furthermore, by construction we have $\frac{l\lambda_r}{T} > 0$, which concludes the proof.

■

Remark A.3 *Since $\mathcal{S}$ is bounded we did not need to require that the first derivative — $\dot{\phi}$ is bounded since we already assumed its Lipschitz continuity.*

References

Caroline A Abbott, Rayaz A Malik, Ernest RE van Ross, Jai Kulkarni, and Andrew JM Boulton. Prevalence and characteristics of painful diabetic neuropathy in a large community-based diabetic population in the uk. *Diabetes care*, 34(10):2220–2224, 2011.

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

Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou. Trading convexity for scalability. In *Proceedings of the 23rd international conference on Machine learning*, pages 201–208. ACM, 2006.

Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337–407, 2000.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, pages 1189–1232, 2001.

Bradley S Galer, Ann Gianas, and Mark P Jensen. Painful diabetic polyneuropathy: epidemiology, pain description, and quality of life. *Diabetes research and clinical practice*, 47(2):123–128, 2000.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2009.

Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

Yi Lin. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6:259–275, 2002.

Yi Lin. A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68(1):73 – 82, 2004. ISSN 0167-7152.

Yufeng Liu. Fisher consistency of multicategory support vector machines. In *International Conference on Artificial Intelligence and Statistics*, pages 291–298, 2007.

Catherine L Martin, James Albers, William H Herman, Patricia Cleary, Barbara Waberski, Douglas A Greene, Martin J Stevens, and Eva L Feldman. Neuropathy among the diabetes control and complications trial cohort 8 years after trial completion. *Diabetes care*, 29 (2):340–344, 2006.

Hamed Masnadi-Shirazi and Nuno Vasconcelos. On the Design of Loss Functions for Classification: theory, robustness to outliers, and SavageBoost. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors, *NIPS*, pages 1049–1056. Curran Associates, Inc., 2008.

Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent in function space. *NIPS*, 1999.

Shawn N Murphy, Michael E Mendis, David A Berkowitz, Isaac Kohane, and Henry C Chueh. Integration of clinical and genetic data in the i2b2 architecture. In *AMIA Annual Symposium Proceedings*, volume 2006, page 1040. American Medical Informatics Association, 2006.

Shawn N Murphy, Griffin Weber, Michael Mendis, Vivian Gainer, Henry C Chueh, Susanne Churchill, and Isaac Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, 17(2):124–130, 2010.

Yurii Nesterov. *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media, 2004.

Gerard Said. Diabetic neuropathya review. *Nature Clinical Practice Neurology*, 3(6):331–340, 2007.

Xiaotong Shen, George C Tseng, Xuegong Zhang, and Wing Hung Wong. On ψ-learning. *Journal of the American Statistical Association*, 98(463):724–734, 2003.

Ambuj Tewari and Peter L Bartlett. On the consistency of multiclass classification methods. *The Journal of Machine Learning Research*, 8:1007–1025, 2007.

PK Thomas and SG Eliasson. Diabetic neuropathy. *Peripheral neuropathy*, 2:1773–1810, 1984.

Zhu Wang. Multi-class hingeboost. method and application to the classification of cancer types using gene expression data. *Methods of information in medicine*, 51(2):162–167, 2012.

Ji Zhu, Saharon Rosset, Hui Zou, and Trevor Hastie. Multi-class adaboost. *Statistics and Its Interface*, 2:349–360, 2009.

D Ziegler, FA Gries, M Spüler, and F Lessmann. The epidemiology of diabetic neuropathy. *Journal of diabetes and its complications*, 6(1):49–57, 1992.

Hui Zou, Ji Zhu, and Trevor Hastie. New multicategory boosting algorithms based on multicategory fisher-consistent losses. *The Annals of Applied Statistics*, pages 1290–1306, 2008.