Cost-sensitive Multiclass Classification Risk Bounds

Bernardo Avila Pires, Mohammad Ghavamzadeh, Csaba Szepesvari

Cost-sensitive Multiclass Classification Risk Bounds

Bernardo Ávila Pires BPIRES@UALBERTA.CA
Department of Computing Science
University of Alberta

Mohammad Ghavamzadeh MOHAMMAD.GHAVAMZADEH@INRIA.FR
Team SequeL
INRIA Lille - Nord Europe

Csaba Szepesvári SZEPESVA@UALBERTA.CA
Department of Computing Science
University of Alberta

Abstract

A commonly used approach to multiclass classification is to replace the $0-1$ loss with a convex surrogate so as to make empirical risk minimization computationally tractable. Previous work has uncovered sufficient and necessary conditions for the consistency of the resulting procedures. In this paper, we strengthen these results by showing how the $0-1$ excess loss of a predictor can be upper bounded as a function of the excess loss of the predictor measured using the convex surrogate. The bound is developed for the case of cost-sensitive multiclass classification and a convex surrogate loss that goes back to the work of Lee, Lin and Wahba. The bounds are as easy to calculate as in binary classification. Furthermore, we also show that our analysis extends to the analysis of the recently introduced "Simplex Coding" scheme.

1. Introduction

A common technique to reduce the computational cost of learning a classifier is to define a convex "surrogate loss", such as the hinge loss in binary classification (e.g., Cortes & Vapnik, 1995). Although the computational problem that results may be more amenable to efficient optimization, it is unclear whether minimizing the surrogate loss will still result in a good accuracy. In fact, whether this happens is clearly a property of the surrogate loss (Rosasco et al., 2004). If L_2 is the surrogate loss and L_1 is the loss of primary interest, the question can be reduced to studying how small the suboptimality gap measured in L_2 should be to achieve a suboptimality gap measured in L_1 of a given size ε (Bartlett et al., 2006; Steinwart, 2007). If it suffices to keep the suboptimality gap measured in L_2 below $\delta(\varepsilon)$ to achieve a suboptimality gap as measured in L_1 below ε then δ is called a *calibration function* (Steinwart, 2007). When a positive-valued calibration function exists, the surrogate loss is said to be *calibrated* w.r.t. the primary loss. While the existence of a positive-valued calibration function ensures consistency, the knowledge of a calibration function allows one to derive finite-sample bounds for the primary loss given such bounds for the surrogate loss.

Calibration is well-understood in the case of binary classification losses where the surrogate loss is based on some convex function φ. Bartlett et al. (2006) fully characterized the calibration functions for the losses that are based on convex φ, for binary cost-insensitive classification (cf. Lemma 3.3). Steinwart (2007) provides a general treatment of calibration and calibration functions, and as an example recovers and extends the results of Bartlett et al. (2006) for the binary cost-sensitive scenario.

The central contribution of this paper is an analytic expression for a calibration function for the loss derived from that of Lee et al. (2004), where the hinge loss is replaced in their definition by φ. Here, calibration is meant w.r.t. a 0-1-like cost-sensitive loss. The conditions we impose on φ are convexity and no non-positive subdifferentials at zero. We also require φ to

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

be such that the loss of Lee et al. (2004) to be lower-bounded over the set of "centered" vectors. The calibration function we derive has the same form as that given by Bartlett et al. (2006) and Steinwart (2007) for binary classification (up to modifications for cost-sensitivity), meaning that the effort needed to calculate a calibration function for multiclass classification is the same as that for a binary classification problem with the same loss.

Our general-form calibration function also applies to the loss of Lee et al. (2004) with "rotated" inputs. We show that in K-class classification problems we can apply input transformations to convert a minimization with a sum-to-zero-constraint over vectors in $\mathbb{R}^K$ into an unconstrained minimization over $\mathbb{R}^{K-1}$. Moreover, we show that a particular case of these rotations is the Simplex Coding studied by Mroueh et al. (2012) (who derived calibration functions for the cases when φ is the hinge loss and the squared loss). To the best of our knowledge, the only work that has reported similar analytic expressions for calibration functions in multiclass classification is the one by Mroueh et al. (2012). As a secondary consequence of one of our results, we show that the calibration function is the same for certain well-known $\varphi(s)$ and their "truncated" version, *e.g.*, $(1+s)^2$ and $[(1+s)_+]^2$, or $|1+s|$ and $(1+s)_+$.

The remainder of this paper is organized as follows. In Section 2, we describe the cost-sensitive multiclass classification problem, and present our risk bounds, our general-form calibration function for the surrogate loss of Lee et al. (2004), along with a few examples of choices of φ and respective calibration functions. In Section 3, we present the proof of our core Theorem together with the intermediate results and the sketches of their proofs. The proofs themselves are presented in the supplementary material (Appendix A) due to space constraints. In Section 4, we generalize our results to encompass the Simplex Coding losses of Mroueh et al. (2012). Finally, in Section 5, we present a brief overview of the work that provides the foundation and the context for our results.

2. Risk Bounds for Cost-sensitive Multiclass Classification

Let (X, Y) be jointly distributed random variables taking values in the measurable spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\mathcal{K} = \{1, \dots, K\}$ be the label space and fix a measurable function $c : \mathcal{X} \times \mathcal{K} \times \mathcal{Y} \to \mathbb{R}$. The multiclass cost-sensitive classification problem finds a (measurable) mapping $g : \mathcal{X} \to \mathcal{K}$ that achieves the smallest expected label cost, or in short, risk

$$\mathcal{R}(g) = \mathbb{E}\left[c(X, g(X), Y)\right], \tag{1}$$

where $(X, Y) \sim P$ are jointly distributed random variables. This means that $\mathbb{E}\left[c(x, k, Y)|X = x\right]$ is the expected cost of assigning the label k to input x, while $c(x, k, Y)$ is the random cost of this labelling. If $K = 2$, we have *binary* cost-sensitive classification, and when $\mathcal{Y} = \mathcal{K}$ and $c(X, k, Y) = \mathbb{I}_{\{k \neq Y\}}$, the problem is *ordinary* multiclass classification.

In the learning problem, the distribution P is unknown and we are only given a finite, i.i.d. sample $(X_i, Y_i) \sim P$, $i = 1, \dots, n$. A natural approach to this problem is to minimize the empirical risk

$$\mathcal{R}_n(g) = \frac{1}{n} \sum_{i=1}^{n} C_i(X_i, g(X_i)),$$

together with an optional penalty term to prevent overfitting. Here we use the shorthand notation $C_i(x, k) = c(x, k, Y_i)$ as the role of Y_i will not be important, usually, only the random costs $C_i(X_i, \cdot)$ are observed, and not Y_i.

In practice, one often starts with some specific form for the *classifiers* $g : \mathcal{X} \to \mathcal{K}$, restricting the search to $\mathcal{G}' \subset \mathcal{G}$, where $\mathcal{G}$ is the set of measurable functions $X \to \mathbb{R}$. Unfortunately, for many function classes $\mathcal{G}'$, even simple ones, calculating such a minimizer is a hard computational problem (Höffgen et al., 1995; Steinwart, 2007), and the common practice is to introduce a so-called *surrogate* objective that depends on the data and the mapping g so as to make the optimization problem convex and, in particular, tractable. What we need to investigate, then, is how minimizing this surrogate loss will help us in minimizing the risk $\mathcal{R}(g)$.

2.1. The Surrogate Loss of Lee et al. (2004)

Let $\mathcal{C} = \left\{t \in \mathbb{R}^K : \mathbf{1}_K^\top t = 0\right\}$ be the set of *balanced (or centered) score-vectors*. A (balanced) *scoring function* h maps $\mathcal{X}$ to $\mathcal{C}$ and is measurable. The surrogate loss of Lee et al. (2004) assumes that the classifier makes its decision based on such a scoring function. In particular, the idea is to assign scores $h_i(x)$ to each label $i \in \mathcal{K}$, and then, given the scores, choose the label that maximizes the score $g(x; h) = \operatorname{argmax}_{i \in \mathcal{K}} h_i(x)$. There is no loss in generality by assuming this form for the classifiers, because given g, any measurable classifier can be constructed by selecting an appropriate scoring function. The surrogate loss then assigns scores to scoring functions and not to classifiers, but again, generality is not harmed by this constraint. To

define this loss, we first choose some convex function $\varphi : \mathbb{R} \to \mathbb{R}$. The expected surrogate loss, or surrogate risk, of $h : \mathcal{X} \to \mathcal{C}$ is defined as

$$\mathcal{R}_\varphi(h) = \mathbb{E}\left[L_\varphi\big(X, h(X)\big)\right], \tag{2}$$

where the surrogate loss of the balanced score vector $t \in \mathcal{C}$ at $x \in \mathcal{X}$ is defined by[1]

$$L_\varphi(x, t) = \sum_{k=1}^{K} c(x, k)\, \varphi(t_k)\,. \tag{3}$$

Here with a slight abuse of notation, we define $c(x, k) = \mathbb{E}\left[c(X, k, Y)|X = x\right]$ as the expected cost of input-label pair.

To understand (3), consider the case when φ is the hinge loss, $\varphi(s) = (1 + s)_+$,[2] $t \in \mathcal{C}$ incurs larger loss when either t_k or the corresponding label cost is large. Intuitively, minimizing the surrogate loss should push for smaller scores for labels with larger cost (in particular, when φ is the hinge loss, scores below -1 do not incur any cost). As a result, the classifier that selects the label with maximal score should incur a small cost.

When the scoring function is a linear function of some weights, the above surrogate loss is convex, and thus, can be used as the basis for designing efficient algorithms to minimize the empirical surrogate loss. What remains now is to investigate how choice of φ affects the performance of the surrogate-risk minimizer w.r.t. the true risk.

2.2. Risk Bounds

In what follows, we let $\partial f(x)$ denote the set of subdifferentials of the convex function $f : \mathbb{R} \to \mathbb{R}$. With a slight abuse of notation, we denote by $\varphi'(0)$ *an arbitrary subdifferential of* φ *at zero*. In the rest of the paper, we will assume that φ satisfies the following two assumptions:

Assumption 2.1. *The function $\varphi : \mathbb{R} \to \mathbb{R}$ is convex with $\partial\varphi(0) \subset (0, \infty)$.*

Assumption 2.2. $\inf_{t\in\mathcal{C}} \sum_{k=1}^{K} c(X, k)\varphi(t_k) > -\infty$ *almost surely (a.s.).*

The effort needed to verify Assumption 2.2 may change on a case-by-case basis. However, since we are ultimately interested in not making any assumption about the expected costs of the classes (besides non-negativity), we may want to select φ that satisfies $\inf_{t\in\mathcal{C}} \sum_{k=1}^{K} \rho_k\varphi(t_k) > -\infty$ for any non-negative constants $\rho_1, \ldots, \rho_K$. Fortunately, verifying whether this condition holds for a given φ is straightforward as we need the condition to hold for $\rho_1 = \cdots = \rho_{K-1} = 0$ and $\rho_K = 1$ (this is the case in binary classification when we have $x \in \mathcal{X}$ with $\mathbb{P}(Y = K|X = x) = 1$). In this case, we have $\inf_{t\in\mathcal{C}} \sum_{k=1}^{K} \rho_k\varphi(t_k) = \inf_s \rho_K\varphi(s)$, and so $\inf_s \varphi(s) > -\infty$ is necessary. It is also easy to show that it is sufficient.

Definition 2.1. *The function $f : \mathcal{C} \to \mathcal{K}$ is called a* maximum selector *if $f(t) \in \operatorname{argmax}_{k\in\mathcal{K}} t_k$ for all $t \in \mathcal{C}$.*

In what follows, we choose any maximum selector function f. For such an f, $f \circ h : x \mapsto f(h(x))$ is the classifier obtained given a scoring function h.

Let $\mathcal{H}$ be the space of measurable balanced scoring functions. The following theorem reveals how the excess risk of a classifier of the form $f \circ h$ for $h \in \mathcal{H}$ is related to the excess risk of the scoring function measured using the surrogate loss. In particular, the theorem provides an explicit way to calculate excess risk bounds for classifiers of the stated form in terms of excess risk bounds. For this theorem, we will need the following extra assumption:

Assumption 2.3. *Assume that there exist* measurable *functions $g^* : \mathcal{X} \to \mathcal{K}$ and $h^* \in \mathcal{H}$ such that $c(x, g^*(x)) = \min_{k\in\mathcal{K}} c(x, k)$ and $L_\varphi(x, h^*(x)) = \min_{t\in\mathcal{C}} L_\varphi(x, t)$ hold for any $x \in \mathcal{X}$.*

Theorem 2.1 (Risk bound for cost-sensitive multiclass classification)**.** *Select any measurable $c : \mathcal{X} \times \mathcal{K} \times \mathcal{Y} \to \mathbb{R}$ and distribution P over $\mathcal{X} \times \mathcal{Y}$, $(X, Y) \sim P$. Let $\mathcal{R}$ and $\mathcal{R}_\varphi$ be the risk and surrogate risk defined according to Equations (1) and (2), respectively, f be a maximum selector function, and assume that $c(X, k) \geq 0$ holds a.s. for all $k \in \mathcal{K}$.*[3] *For $\varepsilon \geq 0$, $x \in \mathcal{X}$, define*

$$\delta(x, \varepsilon) = \big(c_\varepsilon(x) + c_0(x)\big)\,\varphi(0) - \inf_{s\in\mathbb{R}} \left\{c_\varepsilon(x)\varphi(s) + c_0(x)\varphi(-s)\right\},$$

where

$$c_\varepsilon(x) = \min\left\{c(x, j) : c(x, j) \geq \varepsilon + \min_k c(x, k), 1 \leq j \leq K\right\}.$$

[1] Our loss is a mild generalization of the one by Lee et al. (2004). While we allow the costs assigned to the labels to depend on the inputs, in Lee et al. (2004) the cost of a label only depends on the label identity and not the instance.

[2] Note the difference between our definition and the common formulation of the hinge loss, i.e., $\varphi^{\text{common}}(s) = (1 - s)_+$. Since the original formulation of $L_\varphi(x, t)$ uses $\varphi^{\text{common}}(-t_k)$, by deviating from the common formulation we avoid frequent flipping of signs.

[3] For this result, we only require $c(X, k) \geq 0$ a.s., but for the empirical approximation of the surrogate loss to be convex in practice, we also need $c(X, Y, k) \geq 0$ a.s. for all $k \in \mathcal{K}$.

Then, under Assumptions 2.1, 2.2, and 2.3, we have that $\delta(X,\varepsilon) > 0$ holds a.s. for all $\varepsilon > 0$, and for any $h \in \mathcal{H}$ and $\varepsilon > 0$, $\mathcal{R}_\varphi(h) - \min_{h\in\mathcal{H}} \mathcal{R}_\varphi(h) < \mathbb{E}[\delta(X,\varepsilon)]$ implies

$$\mathcal{R}(f\circ h) - \min_{g\in\mathcal{G}} \mathcal{R}(g) < \varepsilon.$$

For specific convex functions, this result is easy to interpret. For example, for the hinge loss, it is not hard to show that $\varepsilon \leq \delta(x,\varepsilon)$. Further examples are given in Table 1 below.

Following the argument of Steinwart (2007), it turns out that Assumption 2.3 allows one to study the relationship between risks by first considering conditional risks. The proof of Theorem 2.1 follows immediately from the "pointwise" result stated below, which we will prove in Section 3.

Theorem 2.2. *Let $\rho_1 \leq \cdots \leq \rho_K, K > 0$ be non-negative numbers. Consider a maximum selector function f, and also a convex function φ satisfying Assumption 2.1 and s.t.*

$$\inf_{t\in\mathcal{C}} \sum_{k=1}^K \rho_k \varphi(t_k) > -\infty.$$

For $\varepsilon \geq 0$, define

$$\begin{aligned} j_\varepsilon &= \min\{j : \rho_j \geq \varepsilon + \rho_1\}, \\ \delta(\varepsilon) &= (\rho_{j_\varepsilon} + \rho_{j_0})\varphi(0) - \inf_{s\in\mathbb{R}} \{\rho_{j_\varepsilon}\varphi(s) + \rho_{j_0}\varphi(-s)\}. \end{aligned}$$

Then, for all $\varepsilon > 0$ and $t \in \mathcal{C}$, if

$$\sum_{k=1}^K \rho_k \varphi(t_k) - \inf_{t'\in\mathcal{C}} \sum_{k=1}^K \rho_k \varphi(t'_k) < \delta(\varepsilon)$$

then $\rho_{f(t)} - \inf_k \rho_k < \varepsilon$. Moreover, $\delta(\varepsilon) > 0$ holds for all $\varepsilon > 0$.

An equivalent way of stating the conclusion of this theorem is to say that δ is a calibration function for L_2^φ w.r.t. L_1^f, where $L_1^f(t) = \rho_{f(t)}$ and $L_2^\varphi(t) = \sum_{k=1}^K \rho_k \varphi(t_k)$.

Proof of Theorem 2.1. We first prove that the bound holds pointwise, *i.e.*, for any $\varepsilon > 0$ and any $h \in \mathcal{H}$, we have

$$\begin{aligned} &\mathbb{E}[L_\varphi(X,h(X))|X] - \min_{h'\in\mathcal{H}} \mathbb{E}[L_\varphi(X,h'(X))|X] < \delta(\varepsilon) \\ &\Rightarrow c(X, f\circ h(X)) - \min_{g\in\mathcal{G}} c(X,g(X)) < \varepsilon. \end{aligned} \tag{4}$$

This pointwise bound is the direct application of Theorem 2.2 with $t = h(X)$ and $\rho_k = c(X,k)$. Assumption 2.2 ensures that $\inf_{t\in\mathcal{C}} \sum_{k=1}^K \rho_k\varphi(t_k) > -\infty$ in all cases that Theorem 2.2 is used. Note that with this definition of t and ρ_k we have $\sum_{k=1}^K \rho_k\varphi(t_k) = \mathbb{E}[L_\varphi(X,h(X))|X]$ and $f(t) = (f\circ h)(X)$. To conclude the proof we need to take the expectation (w.r.t. X) of both sides of Equation 4, and the result follows because from Assumption 2.3 we know that both $f\circ h^*$ and g^* are minimizers of $\mathcal{R}(g)$ over $g \in \mathcal{G}$, and that $h^* \in \operatorname{argmin}_{h\in\mathcal{H}} \mathcal{R}_\varphi(h)$. Thus, we may write

$$\begin{aligned} \min_{h\in\mathcal{H}} \mathbb{E}[L_\varphi(X,h(X))] &= \mathbb{E}\left[\min_{t\in\mathcal{C}} \mathbb{E}[L_\varphi(t,Y)|X]\right], \\ \min_{g\in\mathcal{G}} \mathbb{E}[c(X,g(X))] &= \mathbb{E}\left[\min_{k\in\mathcal{K}} \mathbb{E}[c(X,k)|X]\right]. \end{aligned}$$

□

Next, we present a few examples of $\delta(\varepsilon)$ for different choices of φ, so as to illustrate the type of risk bounds that can be obtained using Theorem 2.1. The examples are presented in Table 1, while the derivations are given in Appendix C.

These results can also be found in Table 2 of Steinwart (2007), except for $\varphi(s) = \ln(1+e^s)$ and $\varphi(s) = |1+s|$. The main difference to the work of Steinwart (2007), however, is that here we prove that these are calibration for *multiclass* classification, whereas Steinwart (2007) proved that they are calibration functions in the *binary* case.

$\varphi(s)$	$\delta(\varepsilon)$ lower bound
$(1+s)_+$	ε
$\lvert 1+s\rvert$	ε
$(1+s)^2$	$\frac{\varepsilon^2}{\rho_{j_\varepsilon}+\rho_1}$
$[(1+s)_+]^2$	$\frac{\varepsilon^2}{\rho_{j_\varepsilon}+\rho_1}$
e^s	$\left(\sqrt{\rho_{j_\varepsilon}} - \sqrt{\rho_1}\right)^2$
$\ln(1+e^s)$	$(\rho_1+\rho_{j_\varepsilon})\left[H(\rho_1,\rho_{j_\varepsilon}) - H\left(\frac{1}{2},\frac{1}{2}\right)\right]$

Table 1. Examples of the choices of φ and lower bounds on the corresponding calibration functions. In the table, $H(a,b) = \frac{a}{a+b}\ln\left(\frac{a}{a+b}\right) + \frac{b}{a+b}\ln\left(\frac{b}{a+b}\right)$, for $a,b>0$. The calibration functions presented for e^s and $\ln(1+e^s)$ require $\rho_1 > 0$. When this is not the case, both are equal to $\delta(\varepsilon) = \rho_{j_\varepsilon}$.

3. Proof of Theorem 2.2

In this section, we give the proof of Theorem 2.2, which is broken into a series of lemmas. The proofs of some of these intermediate results are delegated to Appendix

A. We start with an adaptation of Definition 2.7 of Steinwart (2007) which will give key concepts for our derivations: *calibration*, and *calibration functions*.

Definition 3.1 (Calibration function, calibration)**.** *Consider a set $\mathcal{I}$ and two functions $L_1, L_2 : \mathcal{I} \to \mathbb{R}$. We call any positive-valued function $\delta : (0,\infty) \to (0,\infty]$ a* calibration function *for the pair (L_1, L_2) if for all $\varepsilon > 0$ and $i \in \mathcal{I}$ it holds that*

$$L_2(i) < \inf_{i' \in \mathcal{I}} L_2(i') + \delta(\varepsilon) \Rightarrow L_1(i) < \inf_{i' \in \mathcal{I}} L_1(i') + \varepsilon.$$

If there exists a calibration function for the pair (L_1, L_2), then L_2 is said to be calibrated *w.r.t. L_1.*

Calibration means that minimizers of L_2 are also minimizers of L_1. As a result, a calibration function expresses an upper-bound on the rate of convergence of the target loss (L_1) in terms of the convergence rate of the surrogate loss (L_2).

Before proceeding, we will make a boundedness assumption. We will discuss its suitability when we make our choices of L_1 and L_2 for multiclass classification. For now it suffices to say that if $\inf_{i \in \mathcal{I}} L_2(i) = -\infty$, then L_2 has no rate of convergence, but it is still, by definition, calibrated w.r.t. L_1, regardless of what L_1 is. This case requires a different treatment of calibration and is beyond the scope of this paper.

Assumption 3.1. *Assume that*

$$-\infty < \inf_{i \in \mathcal{I}} L_1(i) < \infty \quad , \quad -\infty < \inf_{i \in \mathcal{I}} L_2(i) < \infty.$$

An equivalent and useful way to characterize calibration under Assumption 3.1 is as follows:

Proposition 3.1. *For each $\varepsilon \geq 0$, let*

$$\mathcal{M}(\varepsilon) = \left\{ i \in \mathcal{I} : L_1(i) < \inf_{i' \in \mathcal{I}} L_1(i') + \varepsilon \right\}$$

and for $\varepsilon > 0$, let

$$\delta_{\max}(\varepsilon) = \inf_{i \in \mathcal{I} \setminus \mathcal{M}(\varepsilon)} L_2(i) - \inf_{i \in \mathcal{I}} L_2(i).$$

Under Assumption 3.1, it holds that L_2 is calibrated w.r.t. L_1 iff $\delta_{\max}(\varepsilon)$ is positive for all $\varepsilon > 0$. Furthermore, for any $\delta : (0,\infty) \to (0,\infty]$ s.t. $0 < \delta(\varepsilon) \leq \delta_{\max}(\varepsilon)$ for all $\varepsilon > 0$, δ is also a calibration function.

This proposition is derived from Lemma 2.9 of Steinwart (2007) and its proof is included in the appendix for completeness. In what follows, we refer to $\delta_{\max}$ as the *maximum calibration function*. It follows from the second statement in Proposition 3.1 that characterizing a calibration function can be done by "meaningfully" lower-bounding, rather than explicitly calculating, $\delta_{\max}$. This is the strategy that we will use to prove Theorem 2.2.

For the rest of this section, unless otherwise stated, we let $\mathcal{I} = \mathcal{C}$ (*i.e.*, the set of zero-mean vectors in $\mathbb{R}^K$) and fix the non-negative numbers $\rho_1, \ldots, \rho_K$ and $K > 0$, such that $\rho_1 \leq \rho_2 \leq \ldots \leq \rho_K$. We define the loss functions

$$L_1^f(t) = \rho_{f(t)}, \qquad L_2^\varphi(t) = \sum_{k=1}^{K} \rho_k \varphi(t_k),$$

with f a maximum selector and $\varphi : \mathbb{R} \to \mathbb{R}$. We also use the following assumption:

Assumption 3.2. $\inf_{t \in \mathcal{C}} \sum_{k=1}^{K} \rho_k \varphi(t_k) > -\infty$.

Clearly, $-\infty < \inf_{t \in \mathcal{C}} L_1^f(t) < \infty$ and $\inf_{t \in \mathcal{C}} L_2^\varphi(t) < \infty$, whenever φ satisfies Assumption 2.1. Assumption 3.2 ensures that $\inf_{t \in \mathcal{C}} L_2^\varphi(t) > -\infty$.

Unfortunately, in order to verify that Assumption 3.2 holds for constants $\rho_1, \ldots, \rho_K$ in general, it does not suffice to verify that it holds for pairs of ρ_i, ρ_j with $1 \leq i < j \leq K$, or for any strict subset of the constants, as shown by the following proposition.

Proposition 3.2. *For $\varphi : \mathbb{R} \to \mathbb{R}$ with $-\infty < \varphi(0) < \infty$ and constants $\rho_1, \ldots, \rho_K$, $\inf_{t \in \mathcal{C}} \sum_{k=1}^{K} \rho_k \varphi(t_k) > -\infty$ holds iff for every $\mathcal{I} \subseteq \{1, \ldots, K\}$, we have $\inf_{\substack{t \in \mathcal{C}: \\ t_i = 0, i \notin \mathcal{I}}} \sum_{i \in \mathcal{I}} \rho_i \varphi(t_i) > -\infty$.*

Our goal is to find a function δ, which may depend on our choice of φ and $\rho_1, \ldots, \rho_K$, such that under Assumptions 2.1 and 3.2, it is a calibration function for (L_1^f, L_2^φ) for all maximum selectors f. Therefore, it would be enough to find a calibration function for the "worst" maximum selector:

Definition 3.2 (Oracle "worst" maximum selector)**.** *Given non-negative numbers $\rho_1 \leq \rho_2 \leq \ldots \leq \rho_K$, let $\bar{f}(t) = \max \{\operatorname{argmax}_k t_k\}$ for all $t \in \mathbb{R}^K$.*

This means that $\bar{f}$ is a maximum selector function that breaks ties so that for any maximum selector function f and vector $t \in \mathbb{R}^K$, $\rho_{\bar{f}(t)} \geq \rho_{f(t)}$. As a result, we may write $L_1^f \leq L_1^{\bar{f}}$ pointwise, and since $\inf_{t \in \mathcal{C}} L_1^f(t) = \inf_{t \in \mathcal{C}} L_1^{\bar{f}}(t) = \rho_1$, we immediately see that if δ is a calibration function for $(L_1^{\bar{f}}, L_2^\varphi)$, then it is also a calibration function for (L_1^f, L_2^φ). In what follows, we fix f to $\bar{f}$. The set $\mathcal{M}(\varepsilon)$ of Proposition 3.1 takes the form $\mathcal{M}(\varepsilon) = \left\{ t \in \mathcal{C} : \rho_{\bar{f}(t)} - \rho_1 < \varepsilon \right\}$. From Proposition 3.1, we have

$$\delta_{\max}(\varepsilon) = \inf_{t \in \mathcal{C} \setminus \mathcal{M}(\varepsilon)} L_2^\varphi(t) - \inf_{t \in \mathcal{C}} L_2^\varphi(t). \tag{5}$$

Unlike the binary ($K = 2$) case (see Lemma 3.3), this expression itself does not lead to an analytic form for

the calibration function that is straightforward to calculate for different choices of φ. Thus, in order to obtain a calibration function that is as straightforward to calculate as a binary classification calibration function, we lower-bound $\delta_{\max}$. To do so, we find subsets of $\mathcal{C}\backslash\mathcal{M}(\varepsilon)$ where the infimum does not change, and upper bound the second term in (5) in a way that parts of this upper-bound cancel out parts of the first term.

Before developing our lower-bounds, we state a result that specifies $\delta_{\max}$ for the case of $K = 2$. This result is essentially extracted from Theorem 2 in Bartlett et al. (2006) (though we added Assumption 3.2) and its proof is only included for completeness. The result itself will be used to ensure that the calibration functions we present are positive.

Lemma 3.3. *Let $K = 2$ and φ be a convex function satisfying Assumption 3.2 and such that $\partial\varphi(0) \subset [0, \infty)$. Then for*

$$\delta(\rho_1, \rho_2) = (\rho_1 + \rho_2)\varphi(0) - \inf_{s\in\mathbb{R}} \{\rho_1\varphi(s) + \rho_2\varphi(-s)\}$$

it holds that

$$\delta_{\max}(\varepsilon) \geq \delta(\rho_1, \rho_2). \tag{6}$$

Moreover, $\delta(\rho_1, \rho_2) > 0$ for all $0 \leq \rho_1 < \rho_2$, iff additionally φ satisfies Assumption 2.1, i.e., *$\partial\varphi(0) \subset (0, \infty)$.*

It follows from the proof of Lemma 3.3 that while for $\varepsilon \leq \rho_2 - \rho_1$ we have equality in (6), for $\varepsilon > \rho_2 - \rho_1$ we have $\delta_{\max}(\varepsilon) = \infty$ (it is trivial to see this case considering the definition of calibration and Proposition 3.1). The lemma is important because it gives a closed-form expression for $\delta_{\max}$ as a function of $(\varepsilon, \rho_1, \rho_2)$ for many commonly used convex functions φ satisfying Assumptions 2.1 and 3.2 for $\rho_1, \ldots, \rho_K$ (cf. the examples in Table 1).

Let us now return to bounding $\delta_{\max}$ from below when $K \geq 2$. We start by rewriting the first term in (5) as an (equal) infimum over a subset of $\mathcal{C}\backslash\mathcal{M}(\varepsilon)$. We first prove the results for a *non-decreasing* function $\sigma : \mathbb{R} \to \mathbb{R}$ with $\sigma'(0) > 0$. Recall that $0 \leq \rho_1 \leq \ldots \leq \rho_K$. For $\varepsilon > 0$ and $t \in \mathcal{C}$, we define

$$j_\varepsilon = \min\{j : \rho_j - \rho_1 \geq \varepsilon\} \quad \text{and} \quad \theta_t^\varepsilon = -\frac{1}{j_\varepsilon} \sum_{k>j_\varepsilon} t_k.$$

Lemma 3.4. *If $\sigma : \mathbb{R} \to \mathbb{R}$ is convex and non-decreasing, and $\rho_1 \leq \cdots \leq \rho_K$, for all $\varepsilon > 0$, we have*

$$\inf_{t\in\mathcal{C}\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^K \rho_k\sigma(t_k) = \inf_{\theta\geq 0} \inf_{\substack{t\in\mathcal{C}\backslash\mathcal{M}(\varepsilon):\\ \theta_t^\varepsilon=\theta \\ t_k\leq 0, \forall k>j_\varepsilon}} \sum_{i\leq j_\varepsilon} \rho_i\sigma(\theta) + \sum_{k>j_\varepsilon} \rho_k\sigma(t_k).$$

Sketch of the proof of Lemma 3.4. In order to prove this lemma, we first show that

$$\inf_{t\in\mathcal{C}\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^K \rho_k\sigma(t_k) = \inf_{t\in\mathcal{C}\backslash\mathcal{M}(\varepsilon):t_{j_\varepsilon}\geq 0} \sum_{k=1}^K \rho_k\sigma(t_k)$$

by showing that for any $t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)$ with $t_{j_\varepsilon} < 0$, we can construct a $t' \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)$ with $t'_{j_\varepsilon} \geq 0$, whose loss is no smaller than that of t. We construct t' by swapping the coordinates $\bar{f}(t)$ and j_ε (since $t_{\bar{f}(t)} \geq 0$ and $\bar{f}(t) > j_\varepsilon$). The next step is to observe that by Jensen's inequality the infimum above is lower-bounded by the infimum over $t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)$ s.t. $t_1 = \cdots = t_{j_\varepsilon}$. We then show that for any of these vectors t, if $t_i > 0$ for some $i > j_\varepsilon$, we can construct a vector t' still in that set but with $t'_i = 0$ and such that $L_2^\varphi(t') \leq L_2^\varphi(t)$. This means that the infimum can be taken over $\{t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon) : t_1 = \cdots = t_{j_\varepsilon}, \max_{i>j_\varepsilon} t_i \leq 0\}$. □

This result reveals the structure of the minimizers of $L_2^\sigma(t)$ within $\mathcal{C}\backslash\mathcal{M}(\varepsilon)$, and in fact, it can be used to show that this same structure carries on for $L_2^\varphi(t)$ with φ satisfying Assumption 2.1. This is the result of the following lemma.

Lemma 3.5. *Let $0 \leq \rho_1 \leq \ldots \leq \rho_K$ and φ satisfy Assumption 2.1. Let $\sigma = \varphi$ if φ does not have a minimizer, otherwise let*

$$\sigma(s) = \begin{cases} \varphi(s), & s \geq s^*; \\ \varphi(s^*), & s < s^*, \end{cases}$$

where $s^* = \max\{\operatorname{argmin}_s \varphi(s)\}$. Then, for any $\varepsilon \geq 0$, we have

$$\inf_{t\in\mathcal{C}\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^K \rho_k\varphi(t_k) = \inf_{t\in\mathcal{C}\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^K \rho_k\sigma(t_k).$$

Note that if a minimizer of φ exists, then s^* is well-defined thanks to Assumption 2.1.

An interesting corollary of this lemma is that using the absolute value loss, $\varphi(s) = |1 + s|$, gives the same calibration function as using the hinge loss. The same conclusion holds for $\varphi(s) = (1 + s)^2$ and $\varphi(s) = [(1 + s)_+]^2$, etc.

Sketch of the proof of Lemma 3.5. The result is trivial in case $\sigma = \varphi$. In the other case, the argument to prove the lemma is based on the observation that there will be a minimizer of the loss L_2^φ whose coordinates all are larger than or equal to s^* (*i.e.*, at the coordinates, φ is increasing). We show the weaker but sufficient statement that for any vector of the form $t_1 = \cdots = t_{j_\varepsilon}$, s.t. $t_i \leq 0$ for $i > j_\varepsilon$, if $t_i < s^*$ for some i, we can construct another vector t' with $t'_1 = \cdots = t'_{j_\varepsilon} \geq 0$ s.t. $s^* \leq t'_i \leq 0$ for $i > j_\varepsilon$. We observe that $\varphi(t_k) = \sigma(t_k)$ for all k and for all these t, which are clearly in $\mathcal{C}\backslash\mathcal{M}(\varepsilon)$. We then use Lemma 3.4 to show the desired result. □

We are now almost ready to connect all these results in order to prove Theorem 2.2. All that remains is to present the next two lemmas.

Lemma 3.6. *Let ρ_1, ρ_2 be two non-negative numbers and φ satisfy Assumptions 2.1 and 3.2. Then we have*

$$\begin{aligned}&\inf_{\theta\geq 0}(\rho_1+\rho_2)\varphi(\theta) - \inf_{s\in\mathbb{R}}\{\rho_1\varphi(\theta+s)+\rho_2\varphi(\theta-s)\}\\ &= (\rho_1+\rho_2)\varphi(0) - \inf_{s\in\mathbb{R}}\{\rho_1\varphi(s)+\rho_2\varphi(-s)\}\,.\end{aligned}$$

Lemma 3.7. *Let $0 \leq \rho_1 \leq \ldots \leq \rho_K$ and φ satisfy Assumptions 2.1 and 3.2. For $\varepsilon \geq 0$, define*

$$\begin{aligned}j_\varepsilon &= \min\{j : \rho_j \geq \varepsilon + \rho_1\}\,,\\ \delta(\varepsilon) &= (\rho_{j_\varepsilon}+\rho_{j_0})\varphi(0) - \inf_{s\in\mathbb{R}}\{\rho_{j_\varepsilon}\varphi(s)+\rho_{j_0}\varphi(-s)\}\,.\end{aligned}$$

Then for all $\varepsilon > 0$ and $t \in \mathcal{C}$, it also holds that

$$\begin{aligned}\delta_{\max}(\varepsilon) &= \inf_{t'\in\mathcal{C}\backslash\mathcal{M}(\varepsilon)}\sum_{k=1}^K \rho_k\varphi(t'_k) - \inf_{t\in\mathcal{C}}\sum_{k=1}^K\rho_k\varphi(t_k)\\ &\geq \delta(\varepsilon).\end{aligned}$$

The essence of the proof of Lemma 3.7 lies in properly manipulating the constraints of the infima involved in the calculations, in order to "meaningfully" lower-bound $\delta_{\max}$. We conclude this section with the proof of Theorem 2.2.

Proof of Theorem 2.2. The proof follows from Proposition 3.1 combined with Lemma 3.7, and from the observation that $\delta(\varepsilon) = \delta(\rho_{j_\varepsilon}, \rho_1)$, as defined in Lemma 3.3. From Lemma 3.3, we have that the calibration function δ is positive for all $\varepsilon > 0$. This is because by definition of ρ_{j_ε}, $\rho_{j_\varepsilon} > \rho_1$ for all $\varepsilon > 0$.

Since δ is a positive lower-bound to $\delta_{\max}$, from Proposition 3.1, it is a calibration function, and thus, we obtain the calibration result for $\bar{f}$. All that remains is to see that for all $t \in \mathcal{C}$ and $\varepsilon > 0$, $\rho_{\bar{f}(t)} - \rho_1 < \varepsilon$ implies $\rho_{f(t)} - \rho_1 < \varepsilon$. □

4. Calibration Functions for Simplex Coding Losses

One issue with using the loss of Lee et al. (2004) is that it requires enforcing, in practice, the sum-to-zero constraint on the hypothesis class over which we perform the surrogate loss minimization and because of this the running time of the resulting empirical risk-minimization procedure may grow linearly with the number of classes. Mroueh et al. (2012) address this problem through a multiclass classification approach called Simplex Coding. In this section, we demonstrate that the "Simplex Coding loss" is a special type of the loss of Lee et al. (2004) with a particular rotation of the set $\mathcal{C}$. Therefore, we generalize Theorem 2.2 to obtain calibration functions for these "rotated losses", some of which are the Simplex Coding loss. The calibration functions calculated are the same as those of the corresponding problems with "unrotated" inputs (*i.e.*, with the losses defined over $\mathcal{C}$). Due to space constraints, the proofs are given in Appendix B.

We begin with the calibration-function result for general "rotated losses".

Theorem 4.1. *Let $\rho_1 \leq \cdots \leq \rho_K$ be $K > 0$ non-negative numbers, $Q \in \mathbb{R}^{K\times K-1}$ be s.t. the columns of Q span $\mathcal{C}$, and φ satisfy Assumptions 2.1 and 3.2. Denote by q_k the k-th row of Q as a column vector. For $\varepsilon \geq 0$ and $w \in \mathbb{R}^{K-1}$, define*

$$\begin{aligned}&f_Q(w) \in \operatorname*{argmax}_k q_k^\top w\,, \qquad j_\varepsilon = \min\{j : \rho_j \geq \varepsilon + \rho_1\}\,,\\ &\delta(\varepsilon) = (\rho_{j_\varepsilon}+\rho_{j_0})\varphi(0) - \inf_{s\in\mathbb{R}}\{\rho_{j_\varepsilon}\varphi(s)+\rho_{j_0}\varphi(-s)\}\,.\end{aligned}$$

Then, for all $\varepsilon > 0$ and $w \in \mathbb{R}^{K-1}$, it holds that

$$\sum_{k=1}^K \rho_k\varphi(q_k^\top w) - \inf_{w'\in\mathcal{C}}\sum_{k=1}^K\rho_k\varphi(q_k^\top w') < \delta(\varepsilon)$$

implies $\rho_{f_Q(w)} - \rho_1 < \varepsilon$. Furthermore, δ is a calibration function for these losses.

Now, all that remains is to show that the Simplex Coding matrix given by Definition 4.1 below (originally defined by Mroueh et al. 2012) is a matrix Q satisfying the conditions for Theorem 4.1. We show this in Proposition 4.2.

Definition 4.1 (Simplex Coding)**.** *A simplex-coding matrix $C \in \mathbb{R}^{K\times(K-1)}$ is a matrix such that each of its rows c_k satisfy (i) $\|c_k\|_2^2 = 1$; (ii) $c_i^\top c_j = -\frac{1}{K-1}$, for $i \neq j$, $i, j \in \mathcal{K}$; and (iii) $\sum_{k=1}^K c_k = \mathbf{0}_{K-1}$.*

Proposition 4.2. *The columns of a simplex-coding matrix C span $\mathcal{C}$.*

5. Related Work

Calibration functions are well characterized for convex binary classification losses. Bartlett et al. (2006) fully characterized calibration functions for losses based on convex φ, in binary cost-insensitive classification. In contrast to our presentation of their main result (Lemma 3.3), and also in contrast to our results, Bartlett et al. (2006) characterize binary classification calibration even when $\inf_s \rho_1\varphi(s) + \rho_2\varphi(-s) = -\infty$. A similar construction seems to be harder to deal with in multiclass classification, and, thus, we leave relaxing Assumption 3.2 for future work.

The work of Steinwart (2007) provides a general treatment of calibration and calibration functions, and as an example recovers and extends the results of Bartlett et al. (2006) for the binary cost-sensitive scenario. Proposition 3.1, which is the starting point of our work on calibration functions for multiclass classification, is a special case of Lemma 2.9 of Steinwart (2007). Some of the examples of calibration functions in Table 1 were originally presented in Steinwart (2007).

For multiclass classification, Tewari & Bartlett (2007) furthered the work of Zhang (2004) and showed that the surrogate loss we study in this paper is *consistent* (*i.e.*, calibrated) in ordinary multiclass classification. They presented an asymptotic convergence result, guaranteeing that a minimizer of $\mathcal{R}_\varphi(h)$ w.r.t. h also minimizes $\mathcal{R}(f \circ h)$ (an *existential* proof for a calibration function). Liu (2007) also provided calibration and non-calibration existential proofs. Along these lines, Guruprasad & Agarwal (2012) investigated conditions for the existence of surrogate losses that are calibrated w.r.t. a generalized notion of a true loss that encompasses the cost-sensitive 0-1-like labelling cost. Their results are also proofs of calibration without explicit calculation of calibration functions.

The work of Chen & Sun (2006) is the closest in nature to our results. However, their results are for ordinary multiclass classification and the assumptions they make are stronger. They require φ to be differentiable, convex, and increasing with $\lim_{s\to-\infty}\varphi(s) = 0$ and $\lim_{s\to\infty}\varphi(s) = \infty$. Moreover, the minimizer of $t \mapsto \sum_{k=1}^K \rho_k\varphi(t_k)$ must exist over $\mathcal{C}$. It is possible to show that their Condition (3), which is essential to their main result via Condition (6), is equivalent to assuming that there exist $\alpha \geq 1$ and $\beta \geq 0$ s.t. for all $\varepsilon > 0$, $\inf_{\substack{t\in\mathcal{C}\\ t_{j_\varepsilon}=t_1}} \sum_{k=1}^K \rho_k\varphi(t_k) - \inf_{t\in\mathcal{C}} \sum_{k=1}^K \rho_k\varphi(t_k) \geq \beta(\rho_{j_\varepsilon} - \rho_1)^\alpha$. From our Lemma 3.4, we know that this corresponds to assuming that $\delta(\varepsilon) = \beta(\rho_{j_\varepsilon} - \rho_1)^\alpha$ is a calibration function for $(L_1^{\bar{f}}, L_2^\varphi)$. As a result, verifying Condition (6) of Chen & Sun (2006) is not simpler than lower-bounding $\delta_{\max}$ directly. On the other hand, in our approach we managed to avoid these complications by lower-bounding $\delta_{\max}$ with a binary classification calibration function without fixing φ in advance.

Mroueh et al. (2012) provided risk bounds for minimizers of simplex-coding-based losses and studied classification algorithms based on minimizing these losses. Our results in Section 4 show that minimizing the loss of Lee et al. (2004) over the set $\mathcal{C}$ is equivalent to minimizing a simplex coding loss over $\mathbb{R}^{K-1}$. This essentially generalizes our Theorem 2.2 to simplex coding losses, and is the statement of Theorem 4.1. By using this theorem with the calibration function examples in Table 2 of Steinwart (2007), one can easily recover the bounds in Theorem 1 of Mroueh et al. (2012) for the SC-SVM. However, more investigation is required to explicitly recover two other losses studied in Mroueh et al. (2012): SH-SVM and S-LS.

There is some overlap between results in proper loss functions in density estimation (cf. Reid & Williamson 2009; Vernet et al. 2011; Reid et al. 2012) and in calibration. Reid & Williamson (2009) recover the calibration function for binary classification of Bartlett et al. (2006), but the works in "multiclass" density estimation, in particular those by Vernet et al. (2011); Reid et al. (2012), contain results that are not immediately related to calibration functions. The authors, however, only investigate how their results relate to *existential* results in cost-insensitive multiclass calibration.

Ben-David et al. (2012) studied surrogate loss risk bounds for *linear* binary classifiers, which is related but complementary to our results. This is because their problem concerns classification loss guarantees for classifiers in a specific class, rather than considering all measurable classifiers. It remains an interesting direction to extend our results to a similar case.

6. Acknowledgements

We would like to thank Yaoliang Yu and András György for the help provided in discussing and verifying the proofs of some of our lemmas. This work was supported by Alberta Innovates Technology Futures, NSERC and INRIA.

References

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

Ben-David, Shai, Loker, David, Srebro, Nathan, and Sridharan, Karthik. Minimizing the misclassification error

rate using a surrogate convex loss. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 1863–1870, 2012.

Chen, Di-Rong and Sun, Tao. Consistency of multiclass empirical risk minimization methods based on convex loss. *The Journal of Machine Learning Research*, 7: 2435–2447, 2006.

Cortes, Corinna and Vapnik, Vladimir. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Guruprasad, Harish and Agarwal, Shivani. Classification calibration dimension for general multiclass losses. In *Advances in Neural Information Processing Systems 25*, pp. 2087–2095. 2012.

Höffgen, Klaus-U, Simon, Hans-U, and Horn, Kevin S Van. Robust trainability of single neurons. *Journal of Computer and System Sciences*, 50:114–125, Jan 1995.

Lee, Yoonkyung, Lin, Yi, and Wahba, Grace. 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.

Liu, Yufeng. Fisher consistency of multicategory support vector machines. *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, 2:289–296, 2007.

Mroueh, Youssef, Poggio, Tomaso, Rosasco, Lorenzo, and Slotine, Jean-Jacques. Multiclass learning with simplex coding. In *Advances in Neural Information Processing Systems 25*, pp. 2798–2806. 2012.

Reid, Mark, Williamson, Robert, and Sun, Peng. The convexity and design of composite multiclass losses. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 687–694, 2012.

Reid, Mark D. and Williamson, Robert C. Surrogate regret bounds for proper losses. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML-09)*, pp. 897–904, 2009.

Rosasco, Lorenzo, Vito, Ernesto De, Caponnetto, Andrea, Piana, Michele, and Verri, Alessandro. Are loss functions all the same? *Neural Computation*, 16(5):1063–107, 2004.

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

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

Vernet, Elodie, Williamson, Robert C., and Reid, Mark D. Composite multiclass losses. In *Advances in Neural Information Processing Systems 24*, pp. 1224–1232. 2011.

Zhang, Tong. Statistical analysis of some multi-category large margin classification methods. *The Journal of Machine Learning Research*, 5:1225–1251, 2004.

APPENDIX—SUPPLEMENTARY MATERIAL

This supplementary material covers the proofs omitted from the main text due to space constraints. The respective results are re-stated for ease of reference.

A. Proofs

Proposition 3.1. *For each $\varepsilon \geq 0$, let*

$$\mathcal{M}(\varepsilon) = \left\{ i \in \mathcal{I} : L_1(i) < \inf_{i' \in \mathcal{I}} L_1(i') + \varepsilon \right\}$$

and for $\varepsilon > 0$, let

$$\delta_{\max}(\varepsilon) = \inf_{i \in \mathcal{I} \setminus \mathcal{M}(\varepsilon)} L_2(i) - \inf_{i \in \mathcal{I}} L_2(i).$$

Under Assumption 3.1, it holds that L_2 is calibrated w.r.t. L_1 iff $\delta_{\max}(\varepsilon)$ is positive for all $\varepsilon > 0$. Furthermore, for any $\delta : (0, \infty) \to (0, \infty]$ s.t. $0 < \delta(\varepsilon) \leq \delta_{\max}(\varepsilon)$ for all $\varepsilon > 0$, δ is also a calibration function.

Proof of Proposition 3.1. First, assume that $\delta_{\max}$ is positive-valued for $\varepsilon > 0$. We claim that from this follows that L_2 is calibrated w.r.t. L_1. In fact, we claim that in this case $\delta_{\max}$ is a calibration function as per Definition 3.1. By assumption, $\delta_{\max}$ is positive-valued for all $\varepsilon > 0$. Now, let us check the last part of Definition 3.1. For this choose $\varepsilon > 0$ and take any $i \in \mathcal{I}$ such that $L_2(i) < \inf_{i' \in \mathcal{I}} L_2(i') + \delta_{\max}(\varepsilon)$. By the definition of $\delta_{\max}$, it follows that $L_2(i) < \inf_{i' \in \mathcal{I} \setminus \mathcal{M}(\varepsilon)} L_2(i')$.[4] Therefore, i must be in $\mathcal{M}(\varepsilon)$, and so $L_1(i) < \inf_{i' \in \mathcal{I}} L_1(i') + \varepsilon$. This shows that $\delta_{\max}$ is indeed a calibration function.

Now, suppose that L_2 is calibrated w.r.t. to L_1. Then there exists a calibration function δ s.t., by contrapositive, for all $\varepsilon > 0, i \in \mathcal{I}$

$$L_1(i) - \inf_{i' \in \mathcal{I}} L_1(i') \geq \varepsilon \Rightarrow L_2(i) - \inf_{i' \in \mathcal{I}} L_2(i') \geq \delta(\varepsilon) > 0$$

Since this holds for any $i \in \mathcal{I}$, we have, by the definition of infimum, that for any $\varepsilon > 0$, $\delta_{\max}(\varepsilon) = \inf_{i \in \mathcal{I} \setminus \mathcal{M}(\varepsilon)} L_2(i) - \inf_{i' \in \mathcal{I}} L_2(i') \geq \delta(\varepsilon) > 0$, which proves our statement about $\delta_{\max}$. [5]

All that remains to be shown is that any $\delta : (0, \infty) \to (0, \infty]$ s.t. $\delta(\varepsilon) > 0$ and $\delta(\varepsilon) \leq \delta_{\max}(\varepsilon)$ is also a calibration function. Indeed, δ is positive, and for all $i \in \mathcal{I}$

$$\begin{aligned} L_2(i) < \inf_{i' \in \mathcal{I}} L_2(i') + \delta(\varepsilon) &\Rightarrow L_2(i) < \inf_{i' \in \mathcal{I}} L_2(i') + \delta_{\max}(\varepsilon) \\ &\Rightarrow L_1(i) < \inf_{i' \in \mathcal{I}} L_1(i') + \varepsilon, \end{aligned}$$

which means δ satisfies the definition of a calibration function, according to Definition 3.1.

□

Proposition 3.2. *For $\varphi : \mathbb{R} \to \mathbb{R}$ with $-\infty < \varphi(0) < \infty$ and constants $\rho_1, \ldots, \rho_K$, $\inf_{t \in \mathcal{C}} \sum_{k=1}^K \rho_k \varphi(t_k) > -\infty$ holds iff for every $\mathcal{I} \subseteq \{1, \ldots, K\}$, we have $\inf_{\substack{t \in \mathcal{C}: \\ t_i = 0, i \notin \mathcal{I}}} \sum_{i \in \mathcal{I}} \rho_i \varphi(t_i) > -\infty$.*

Proof of Proposition 3.2. For any $\mathcal{I} \subseteq \{1, \ldots, K\}$, we have

$$\begin{aligned} \inf_{\substack{t \in \mathcal{C}: \\ t_i = 0, i \notin \mathcal{I}}} \sum_{i \in \mathcal{I}} \rho_i \varphi(t_i) &= -\sum_{i \notin \mathcal{I}} \rho_i \varphi(0) + \inf_{\substack{t \in \mathcal{C}: \\ t_i = 0, i \notin \mathcal{I}}} \sum_{i \in \mathcal{I}} \rho_i \varphi(t_i) + \sum_{i \notin \mathcal{I}} \rho_i \varphi(t_i) \\ &\geq -\sum_{i \notin \mathcal{I}} \rho_i \varphi(0) + \inf_{t \in \mathcal{C}} \sum_{k=1}^K \rho_k \varphi(t_k). \end{aligned}$$

[4] This is where we use the convention that concerns the case when $\inf_{i \in \mathcal{I}} L_2(i) = -\infty$.

[5] This also shows that no calibration function is larger than $\delta_{\max}$.

Since $-\infty < \varphi(0) < \infty$, if there exists $\mathcal{I} \subseteq \{1, \dots, K\}$ s.t.

$$\inf_{\substack{t \in \mathcal{C}: \\ t_i = 0, i \notin \mathcal{I}}} \sum_{i \in \mathcal{I}} \rho_i \varphi(t_i) = -\infty,$$

then

$$\inf_{t \in \mathcal{C}} \sum_{k=1}^{K} \rho_k \varphi(t_k) = -\infty.$$

Conversely, if for all $\mathcal{I} \subseteq \{1, \dots, K\}$ we have

$$\inf_{\substack{t \in \mathcal{C}: \\ t_i = 0, i \notin \mathcal{I}}} \sum_{i \in \mathcal{I}} \rho_i \varphi(t_i) > -\infty,$$

then, in particular, we have it for $\mathcal{I} = \{1, \dots, K\}$, and so

$$\inf_{t \in \mathcal{C}} \sum_{k=1}^{K} \rho_k \varphi(t_k) > -\infty.$$

Note that it is nevertheless possible to have $\inf_s \varphi(s) = -\infty$, since if $|\mathcal{I}| = 1$ then

$$\inf_{\substack{t \in \mathcal{C}: \\ t_i = 0, i \notin \mathcal{I}}} \sum_{i \in \mathcal{I}} \rho_i \varphi(t_i) = \sum_{i \in \mathcal{I}} \rho_i \varphi(0).$$

□

Lemma 3.3. *Let $K = 2$ and φ be a convex function satisfying Assumption 3.2 and such that $\partial\varphi(0) \subset [0, \infty)$. Then for*

$$\delta(\rho_1, \rho_2) = (\rho_1 + \rho_2)\varphi(0) - \inf_{s \in \mathbb{R}} \{\rho_1 \varphi(s) + \rho_2 \varphi(-s)\}$$

it holds that

$$\delta_{\max}(\varepsilon) \geq \delta(\rho_1, \rho_2). \tag{6}$$

Moreover, $\delta(\rho_1, \rho_2) > 0$ for all $0 \leq \rho_1 < \rho_2$, iff additionally φ satisfies Assumption 2.1, i.e., *$\partial\varphi(0) \subset (0, \infty)$.*

Proof of Lemma 3.3. Let us start by showing that (6) holds. Let $\delta = \delta(\rho_1, \rho_2)$. Remember that by assumption $\rho_1 \leq \rho_2$. For $\varepsilon > \rho_2 - \rho_1$, the result is trivially true, since $\delta < \infty$ and in this case $\mathcal{M}(\varepsilon) = \mathcal{C}$ and therefore $\delta_{\max}(\varepsilon) = \infty$. Hence, it remains to consider $0 < \varepsilon \leq \rho_2 - \rho_1$. In this case, we must have $\rho_2 > \rho_1$, and we have $\mathcal{M}(\varepsilon) = \{(s, -s) : s > 0\}$, so

$$\delta_{\max}(\varepsilon) = \inf_{s \leq 0} \{\rho_1 \varphi(s) + \rho_2 \varphi(-s)\} - \inf_{s} \{\rho_1 \varphi(s) + \rho_2 \varphi(-s)\}.$$

Using linear lower bounds for $s \mapsto \varphi(s)$, $s \mapsto \varphi(-s)$ at 0,

$$\begin{aligned}
(\rho_1 + \rho_2)\varphi(0) &\geq \inf_{s \leq 0} \{\rho_1 \varphi(s) + \rho_2 \varphi(-s)\} \\
&\geq \inf_{s \leq 0} \{\rho_1 \varphi(0) + \rho_1 \varphi'(0)s + \rho_2 \varphi(0) - \rho_2 \varphi'(0)s\} \\
&= (\rho_1 + \rho_2)\varphi(0) + \inf_{s \leq 0} \{(\rho_1 - \rho_2)\varphi'(0)s\} \\
&\geq (\rho_1 + \rho_2)\varphi(0),
\end{aligned}$$

because $(\rho_1 - \rho_2)\varphi'(0)s \geq 0$ for all $s \leq 0$ thanks to our assumption that $\varphi'(0) \geq 0$ and $\rho_1 < \rho_2$. Hence,

$$\inf_{s \leq 0} \{\rho_1 \varphi(s) + \rho_2 \varphi(-s)\} = (\rho_1 + \rho_2)\varphi(0),$$

which shows that $\delta_{\max}(\varepsilon) = \delta$ in this case, finishing the proof of the first part.

To prove the second part, first assume that $\partial\varphi(0) \subset (0, \infty)$, which we write[6] as $\varphi'(0) > 0$. The function $\psi(s) = \rho_1\varphi(s) + \rho_2\varphi(-s)$ is decreasing at zero since $\psi'(0) = (\rho_1 - \rho_2)\varphi'(0) < 0$ thanks to $\rho_1 < \rho_2$. Hence,

$$(\rho_1 + \rho_2)\varphi(0) > \inf_s \{\rho_1\varphi(s) + \rho_2\varphi(-s)\}. \tag{7}$$

Now, assume that (7) holds for all $\rho_1 < \rho_2$. Our goal is to show that $\partial\varphi(0) \subset (0, \infty)$, which, under the assumption that $\partial\varphi(0) \subset [0, \infty)$, reduces to showing that $0 \notin \partial\varphi(0)$. We prove this by contradiction. Assume $0 \in \partial\varphi(0)$. Using linear lower bounds for $s \mapsto \varphi(s)$ and $s \mapsto \varphi(-s)$ at 0, we have that, for all $s \in \mathbb{R}$ and for any $z \in \partial\varphi(0)$

$$\rho_1\varphi(s) + \rho_2\varphi(-s) \geq (\rho_1 + \rho_2)\varphi(0) + z(\rho_1 - \rho_2)s.$$

In particular, this holds for $z = 0$, so

$$(\rho_1 + \rho_2)\varphi(0) = \inf_s \rho_1\varphi(s) + \rho_2\varphi(-s),$$

which contradicts (7), and so $0 \notin \partial\varphi(0)$ □

Lemma 3.4. *If $\sigma : \mathbb{R} \to \mathbb{R}$ is convex and non-decreasing, and $\rho_1 \leq \cdots \leq \rho_K$, for all $\varepsilon > 0$, we have*

$$\inf_{t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k\sigma(t_k) = \inf_{\theta \geq 0} \inf_{\substack{t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon): \\ \theta_t^\varepsilon = \theta \\ t_k \leq 0, \forall k > j_\varepsilon}} \sum_{i \leq j_\varepsilon} \rho_i\sigma(\theta) + \sum_{k > j_\varepsilon} \rho_k\sigma(t_k).$$

Proof of Lemma 3.4. Consider $t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon) : t_{j_\varepsilon} \leq 0$; then $t_{\bar{f}(t)} \geq 0 \geq t_{j_\varepsilon}$. Define t' s.t.

$$t'_k = \begin{cases} t_{\bar{f}(t)}, & k = j_\varepsilon, \\ t_{j_\varepsilon}, & k = \bar{f}(t), \\ t_k & \text{otherwise.} \end{cases}$$

This essentially "swaps" coordinates $\bar{f}(t)$ and j_ε in t to create t'. Then we have

$$\begin{aligned} \sum_{k=1}^{K} \rho_k\sigma(t_k) - \sum_{k=1}^{K} \rho_k\sigma(t'_k) &= \rho_{\bar{f}(t)} \left(\sigma(t_{\bar{f}(t)}) - \sigma(t_{j_\varepsilon})\right) + \rho_{j_\varepsilon} \left(\sigma(t_{j_\varepsilon}) - \sigma(t_{\bar{f}(t)})\right) \\ &= (\rho_{\bar{f}(t)} - \rho_{j_\varepsilon}) \left(\sigma(t_{\bar{f}(t)}) - \sigma(t_{j_\varepsilon})\right) \\ &\geq 0, \end{aligned}$$

because $\rho_{\bar{f}(t)} \geq \rho_{j_\varepsilon}$, $t_{\bar{f}(t)} \geq t_{j_\varepsilon}$ and σ is non-decreasing. Therefore, since $t' \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)$,

$$\inf_{t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k\sigma(t_k) \geq \inf_{\substack{t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon): \\ t_{j_\varepsilon} \geq 0}} \sum_{k=1}^{K} \rho_k\sigma(t_k).$$

[6] Recall our convention that $\varphi'(s)$ means an arbitrary element of $\partial\varphi(s)$.

For $t \in \mathcal{C}$, let $\theta_t^\varepsilon = -\frac{1}{j_\varepsilon}\sum_{k>j_\varepsilon} t_k$. We have

$$\begin{aligned}
\inf_{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon)} \sum_{k=1}^K \rho_k\sigma(t_k) &\geq \inf_{\substack{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon):\\ t_{j_\varepsilon}\geq 0}} \sum_{k=1}^K \rho_k\sigma(t_k) \\
&= \inf_{\substack{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon):\\ t_{j_\varepsilon}\geq 0}} \sum_{i\leq j_\varepsilon} \rho_i \cdot \sum_{k\leq j_\varepsilon} \frac{\rho_k}{\sum_{i\leq j_\varepsilon}\rho_i}\sigma(t_k) + \sum_{k>j_\varepsilon}\rho_k\sigma(t_k) \\
&\geq \inf_{\substack{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon):\\ t_{j_\varepsilon}\geq 0}} \sum_{i\leq j_\varepsilon} \rho_i\sigma\left(\sum_{k\leq j_\varepsilon} \frac{\rho_k}{\sum_{i\leq j_\varepsilon}\rho_i}\cdot t_k\right) + \sum_{k>j_\varepsilon}\rho_k\sigma(t_k) \\
&= \inf_{\substack{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon):\\ t_{j_\varepsilon}\geq 0}} \sum_{i\leq j_\varepsilon} \rho_i\sigma(\theta_t^\varepsilon) + \sum_{k>j_\varepsilon}\rho_k\sigma(t_k)
\end{aligned}$$

The third line follows by Jensen's inequality, and the last line follows because σ is non-decreasing, $t_{j_\varepsilon} \geq 0$ and $\sum_{k\leq j_\varepsilon} t_k = -\sum_{k>j_\varepsilon} t_k = \theta_t^\varepsilon \cdot j_\varepsilon$.

Consider $t \in \mathcal{C}\setminus\mathcal{M}(\varepsilon) : t_k = \theta_t^\varepsilon, t_i > 0$ for $k \leq j_\varepsilon$, $i > j_\varepsilon$. Also, given one such t, let $t' \in \mathcal{C}\setminus\mathcal{M}(\varepsilon) : t'_k = \theta_{t'}^\varepsilon, t_i = 0, t'_m = t_m$ for $k \leq j_\varepsilon$ and $m > j_\varepsilon$, $m \neq i$. Then

$$\theta_t^\varepsilon = -\frac{1}{j_\varepsilon}\sum_{k>j_\varepsilon} t_k > -\frac{1}{j_\varepsilon}\sum_{\substack{k>j_\varepsilon\\ k\neq i}} t_k = \theta_{t'}^\varepsilon,$$

and so, because σ is non-decreasing,

$$\begin{aligned}
\sum_{k=1}^K \rho_k\sigma(t_k) - \sum_{k=1}^K \rho_k\sigma(t'_k) &= \left(\sum_{k<j_\varepsilon}\rho_k\right)(\sigma(\theta_t^\varepsilon) - \sigma(\theta_{t'}^\varepsilon)) + \rho_i(\sigma(t_i) - \sigma(0)) \\
&\geq 0.
\end{aligned}$$

From this we conclude that

$$\begin{aligned}
\inf_{\substack{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon):\\ t_{j_\varepsilon}\geq 0}} \sum_{i\leq j_\varepsilon} \rho_i\sigma(\theta_t^\varepsilon) + \sum_{k>j_\varepsilon}\rho_k\sigma(t_k) &\geq \inf_{\substack{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon):\\ t_{j_\varepsilon}\geq 0,\\ \theta_t^\varepsilon\geq 0\\ t_k\leq 0,\forall k>j_\varepsilon}} \sum_{i\leq j_\varepsilon} \rho_i\sigma(\theta_t^\varepsilon) + \sum_{k>j_\varepsilon}\rho_k\sigma(t_k) \\
&= \inf_{\substack{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon):\\ \bar{f}(t)=j_\varepsilon,\\ \theta_t^\varepsilon\geq 0\\ t_k\leq 0,\forall k>j_\varepsilon}} \sum_{i\leq j_\varepsilon} \rho_i\sigma(\theta_t^\varepsilon) + \sum_{k>j_\varepsilon}\rho_k\sigma(t_k) \\
&= \inf_{\theta\geq 0} \inf_{\substack{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon):\\ \bar{f}(t)=j_\varepsilon,\\ \theta_t^\varepsilon=\theta\\ t_k\leq 0,\forall k>j_\varepsilon}} \sum_{i\leq j_\varepsilon} \rho_i\sigma(\theta) + \sum_{k>j_\varepsilon}\rho_k\sigma(t_k) \\
&= \inf_{\theta\geq 0} \inf_{\substack{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon):\\ \theta_t^\varepsilon=\theta\\ t_k\leq 0,\forall k>j_\varepsilon}} \sum_{i\leq j_\varepsilon} \rho_i\sigma(\theta) + \sum_{k>j_\varepsilon}\rho_k\sigma(t_k).
\end{aligned}$$

Since

$$\inf_{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon)} \sum_{k=1}^K \rho_k\sigma(t_k) \leq \inf_{\theta\geq 0} \inf_{\substack{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon):\\ \theta_t^\varepsilon=\theta\\ t_k\leq 0,\forall k>j_\varepsilon}} \sum_{i\leq j_\varepsilon} \rho_i\sigma(\theta_t^\varepsilon) + \sum_{k>j_\varepsilon}\rho_k\sigma(t_k),$$

the equality in the lemma follows.

□

Lemma 3.5. *Let $0 \leq \rho_1 \leq \ldots \leq \rho_K$ and φ satisfy Assumption 2.1. Let $\sigma = \varphi$ if φ does not have a minimizer, otherwise let*

$$\sigma(s) = \begin{cases} \varphi(s), & s \geq s^*; \\ \varphi(s^*), & s < s^*, \end{cases}$$

where $s^* = \max\{\operatorname{argmin}_s \varphi(s)\}$. Then, for any $\varepsilon \geq 0$, we have

$$\inf_{t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k \varphi(t_k) = \inf_{t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k \sigma(t_k).$$

Proof of Lemma 3.5. The result is trivial when $\sigma = \varphi$. Assume the complementary case, and, w.l.o.g., that $\rho_1 \leq \cdots \leq \rho_K$. Then φ does have a minimum, and $\varphi'(0) > 0$ implies that $s^* < 0$. For $\varepsilon \geq 0$ and $t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)$, let $\theta_t^\varepsilon = -\frac{1}{j_\varepsilon} \sum_{k>j_\varepsilon} t_k$.

From $\sigma(s) \leq \varphi(s)$ for all $s \in \mathbb{R}$, σ non-decreasing, and Lemma 3.4, we have for all $\varepsilon \geq 0$

$$\begin{aligned} \inf_{t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k \varphi_t(t_k) &\geq \inf_{t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k \sigma(t_k) \\ &= \inf_{\theta \geq 0} \inf_{\substack{t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon): \\ \theta_t^\varepsilon = \theta \\ t_k \leq 0, \forall k > j_\varepsilon}} \sum_{i \leq j_\varepsilon} \rho_i \sigma(\theta) + \sum_{k > j_\varepsilon} \rho_k \sigma(t_k). \end{aligned}$$

Consider $t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon) : t_k = \theta_t^\varepsilon, t_i < s^*, t_m \leq 0$ for $k \leq j_\varepsilon$, for $m > j_\varepsilon$ and for some i (evidently, $i > j_\varepsilon$), and let $t' \in \mathcal{C}\backslash\mathcal{M}(\varepsilon) : t'_k = \theta_{t'}^\varepsilon, t'_i = 0, t'_m = t_m$, for $k \leq j_\varepsilon$ and $m > j_\varepsilon, m \neq i$. Then $t'_m \leq 0$ for $m > j_\varepsilon$, $\sigma(t'_i) = \sigma(t_i)$, and

$$\theta_t^\varepsilon = -\frac{1}{j_\varepsilon} \sum_{k>j_\varepsilon} t_k > -\frac{1}{j_\varepsilon} \sum_{\substack{k>j_\varepsilon \\ k \neq i}} t_k + \frac{s^*}{j_\varepsilon} = \theta_{t'}^\varepsilon.$$

Because σ is non-decreasing in (s^*, ∞), $s^* < 0$ and $0 \leq \theta_{t'}^\varepsilon < \theta_t^\varepsilon$, it follows that

$$\begin{aligned} \sum_{k=1}^{K} \rho_k \sigma(t_k) &= \sum_{i \leq j_\varepsilon} \rho_i \sigma(\theta_t^\varepsilon) + \sum_{k > j_\varepsilon} \rho_k \sigma(t_k) \\ &= \sum_{i \leq j_\varepsilon} \rho_i \sigma(\theta_t^\varepsilon) + \sum_{k > j_\varepsilon} \rho_k \sigma(t'_k) \\ &\geq \sum_{i \leq j_\varepsilon} \rho_i \sigma(\theta_{t'}^\varepsilon) + \sum_{k > j_\varepsilon} \rho_k \sigma(t'_k) \\ &= \sum_{k=1}^{K} \rho_k \sigma(t'_k). \end{aligned}$$

Hence,

$$\begin{aligned}
\inf_{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon)}\sum_{k=1}^{K}\rho_k\sigma(t_k) &\geq \inf_{\theta\geq 0}\ \inf_{\substack{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon):\\ \theta_t^\varepsilon=\theta\\ t_k\leq 0,\forall k>j_\varepsilon}}\ \sum_{i\leq j_\varepsilon}\rho_i\sigma(\theta)+\sum_{k>j_\varepsilon}\rho_k\sigma(t_k)\\
&\geq \inf_{\theta\geq 0}\ \inf_{\substack{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon):\\ \theta_t^\varepsilon=\theta\\ t_k\leq 0,\forall k>j_\varepsilon\\ t_k\geq s^*,\forall k>j_\varepsilon}}\ \sum_{i\leq j_\varepsilon}\rho_i\sigma(\theta)+\sum_{k>j_\varepsilon}\rho_k\sigma(t_k)\\
&= \inf_{\theta\geq 0}\ \inf_{\substack{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon):\\ \theta_t^\varepsilon=\theta\\ t_k\leq 0,\forall k>j_\varepsilon\\ t_k\geq s^*,\forall k>j_\varepsilon}}\ \sum_{i\leq j_\varepsilon}\rho_i\varphi(\theta)+\sum_{k>j_\varepsilon}\rho_k\varphi(t_k)\\
&\geq \inf_{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon)}\sum_{k=1}^{K}\rho_k\varphi_t(t_k)\\
&\geq \inf_{t\in\mathcal{C}\setminus\mathcal{M}(\varepsilon)}\sum_{k=1}^{K}\rho_k\sigma(t_k).
\end{aligned}$$

and the result follows.

□

Lemma 3.6. *Let ρ_1,ρ_2 be two non-negative numbers and φ satisfy Assumptions 2.1 and 3.2. Then we have*

$$\begin{aligned}
&\inf_{\theta\geq 0}(\rho_1+\rho_2)\varphi(\theta)-\inf_{s\in\mathbb{R}}\left\{\rho_1\varphi(\theta+s)+\rho_2\varphi(\theta-s)\right\}\\
&\quad=(\rho_1+\rho_2)\varphi(0)-\inf_{s\in\mathbb{R}}\left\{\rho_1\varphi(s)+\rho_2\varphi(-s)\right\}.
\end{aligned}$$

Proof of Lemma 3.6. Consider a function σ as in Lemma 3.5, *i.e.*, if $\min_s\varphi(s)$ exists, let $s^*=\max\{\operatorname{argmin}_s\varphi(s)\}$ and define

$$\sigma(s)=\begin{cases}\varphi(s) & s\geq s^*,\\ \varphi(s^*) & s<s^*;\end{cases}$$

otherwise, let $\sigma=\varphi$. Note that σ is non-decreasing, $\sigma(s)=\varphi(s)$ for all $s\geq 0$, and as a consequence of Lemma 3.5

$$\inf_{s\in\mathbb{R}}\left\{\rho_1\varphi(s)+\rho_2\varphi(-s)\right\}=\inf_{s\in\mathbb{R}}\left\{\rho_1\sigma(s)+\rho_2\sigma(-s)\right\}.$$

Then

$$\inf_{\theta\geq 0}(\rho_1+\rho_2)\varphi(\theta)-\inf_{s\in\mathbb{R}}\left\{\rho_1\varphi(\theta+s)+\rho_2\varphi(\theta-s)\right\}=\inf_{\theta\geq 0}(\rho_1+\rho_2)\sigma(\theta)-\inf_{s\in\mathbb{R}}\left\{\rho_1\sigma(\theta+s)+\rho_2\sigma(\theta-s)\right\}.$$

We are going to connect three facts to obtain the desired result. First,

$$\inf_{s\in\mathbb{R}}\left\{\rho_1\sigma(\theta+s)+\rho_2\sigma(\theta-s)\right\}$$

is increasing for $\theta\geq 0$. To see this, notice that this quantity is bounded from above, and, by Assumption 3.2, it is also bounded from below. If the infimum is not taken at any finite s[7] then

$$\inf_{s\in\mathbb{R}}\left\{\rho_1\sigma(\theta+s)+\rho_2\sigma(\theta-s)\right\}=\inf_{s\in\mathbb{R}}\left\{\rho_1\sigma(s)+\rho_2\sigma(s)\right\}.$$

[7] This happens, *e.g.*, when $K=2$, $\varphi(s)=e^s$ and $0=\rho_1<\rho_2$.

Otherwise, for each $\theta \geq 0$, the infimum is taken at some finite s^*_θ, so for $0 \leq \theta < \theta'$, we have

$$\begin{aligned}\inf_{s\in\mathbb{R}} \{\rho_1\sigma(\theta'+s) + \rho_2\sigma(\theta'-s)\} &= \{\rho_1\sigma(\theta'+s^*_{\theta'}) + \rho_2\sigma(\theta'-s^*_{\theta'})\} \\ &> \{\rho_1\sigma(\theta+s^*_{\theta'}) + \rho_2\sigma(\theta-s^*_{\theta'})\} \\ &\geq \inf_{s\in\mathbb{R}} \{\rho_1\sigma(\theta+s) + \rho_2\sigma(\theta-s)\}.\end{aligned}$$

Second, $(\rho_1+\rho_2)\sigma(\theta)$ is also increasing in θ. Third, for every $\theta \geq 0$,

$$(\rho_1+\rho_2)\sigma(\theta) \geq \inf_{s\in\mathbb{R}} \{\rho_1\sigma(\theta+s) + \rho_2\sigma(\theta-s)\}$$

Therefore, we obtain that

$$(\rho_1+\rho_2)\sigma(\theta) - \inf_{s\in\mathbb{R}} \{\rho_1\sigma(\theta+s) + \rho_2\sigma(\theta-s)\}$$

is increasing in θ, so that the infimum over non-negative θ is attained at $\theta = 0$, and the final statement of the lemma can be obtained from the above through Lemma 3.5. □

Lemma 3.7. *Let $0 \leq \rho_1 \leq \ldots \leq \rho_K$ and φ satisfy Assumptions 2.1 and 3.2. For $\varepsilon \geq 0$, define*

$$\begin{aligned}j_\varepsilon &= \min\{j : \rho_j \geq \varepsilon + \rho_1\}, \\ \delta(\varepsilon) &= (\rho_{j_\varepsilon} + \rho_{j_0})\varphi(0) - \inf_{s\in\mathbb{R}} \{\rho_{j_\varepsilon}\varphi(s) + \rho_{j_0}\varphi(-s)\}.\end{aligned}$$

Then for all $\varepsilon > 0$ and $t \in \mathcal{C}$, it also holds that

$$\begin{aligned}\delta_{\max}(\varepsilon) &= \inf_{t'\in\mathcal{C}\setminus\mathcal{M}(\varepsilon)} \sum_{k=1}^K \rho_k\varphi(t'_k) - \inf_{t\in\mathcal{C}} \sum_{k=1}^K \rho_k\varphi(t_k) \\ &\geq \delta(\varepsilon).\end{aligned}$$

Proof of Lemma 3.7. We have

$$\delta_{\max}(\varepsilon) = \inf_{t'\in\mathcal{C}\setminus\mathcal{M}(\varepsilon)} \sum_{k=1}^K \rho_k\varphi(t'_k) - \inf_{t\in\mathcal{C}} \sum_{k=1}^K \rho_k\varphi(t_k)$$

for all $\varepsilon > 0$. If φ is non-decreasing, let $\sigma = \varphi$, otherwise take σ as in Lemma 3.5, so that in either case

$$\delta_{\max}(\varepsilon) = \inf_{t'\in\mathcal{C}\setminus\mathcal{M}(\varepsilon)} \sum_{k=1}^K \rho_k\sigma(t'_k) - \inf_{t\in\mathcal{C}} \sum_{k=1}^K \rho_k\sigma(t_k).$$

Given that σ is non-decreasing, from Lemma 3.4 we get that

$$\begin{aligned}\delta_{\max}(\varepsilon) = \inf_{\theta\geq 0} \inf_{\substack{t'\in\mathcal{C}\setminus\mathcal{M}(\varepsilon): \\ \theta^\varepsilon_{t'}=\theta \\ t'_k\leq 0, \forall k>j_\varepsilon}} &\left\{\sum_{i\leq j_\varepsilon} \rho_i\sigma(\theta) + \sum_{k>j_\varepsilon} \rho_k\sigma(t'_k)\right\} \\ &- \inf_{t\in\mathcal{C}} \sum_{k=1}^K \rho_k\sigma(t_k).\end{aligned}$$

Thus, by simple manipulation of the constraints, and lower-bounding the resulting supremum (third line below)

by a particular choice of the argument t (so as to "align" some of its coordinates to t'),

$$\begin{aligned}
\delta_{\max}(\varepsilon) &= \inf_{\theta \geq 0} \inf_{\substack{t' \in \mathcal{C} \setminus \mathcal{M}(\varepsilon): \\ \theta^{\varepsilon}_{t'} = \theta \\ t'_k \leq 0, \forall k > j_\varepsilon}} \sup_{t \in \mathcal{C}} \\
&\qquad \left\{ \sum_{i \leq j_\varepsilon} \rho_i \sigma(\theta) + \sum_{k > j_\varepsilon} \rho_k \sigma(t'_k) - \sum_{k=1}^{K} \rho_k \sigma(t_k) \right\} \\
&\geq \inf_{\theta \geq 0} \inf_{\substack{t' \in \mathcal{C} \setminus \mathcal{M}(\varepsilon): \\ \theta^{\varepsilon}_{t'} = \theta \\ t'_k \leq 0, \forall k > j_\varepsilon}} \sup_{\substack{t \in \mathcal{C}: \\ t_k = t'_k, \forall k > j_\varepsilon}} \\
&\qquad \left\{ \sum_{k \leq j_\varepsilon} \rho_k \sigma(\theta) - \sum_{k \leq j_\varepsilon} \rho_k \sigma(t_k) \right\} \\
&\geq \inf_{\theta \geq 0} \inf_{\substack{t' \in \mathcal{C} \setminus \mathcal{M}(\varepsilon): \\ \theta^{\varepsilon}_{t'} = \theta \\ t'_k \leq 0, \forall k > j_\varepsilon}} \sup_{\substack{t \in \mathcal{C}: \\ t_k = t'_k, \forall k > j_\varepsilon \\ t_k = \theta, \forall 1 < k < j_\varepsilon}} \\
&\qquad \{(\rho_{j_\varepsilon} + \rho_1)\sigma(\theta) - \rho_{j_\varepsilon}\sigma(\theta - t_{j_\varepsilon}) - \rho_1 \sigma(\theta + t_1)\} \\
&= \inf_{\theta \geq 0} \Big\{ (\rho_{j_\varepsilon} + \rho_1)\sigma(\theta) \\
&\qquad - \inf_{s \in \mathbb{R}} \rho_{j_\varepsilon}\sigma(\theta - s) + \rho_1 \sigma(\theta + s) \Big\}.
\end{aligned}$$

By applying Lemma 3.5 to the inequality above, we can replace σ with φ and get

$$\begin{aligned}
\delta_{\max}(\varepsilon) \geq \inf_{\theta \geq 0} \Big\{ &(\rho_{j_\varepsilon} + \rho_1)\varphi(\theta) \\
&- \inf_{s \in \mathbb{R}} \rho_{j_\varepsilon}\varphi(\theta - s) + \rho_1 \varphi(\theta + s) \Big\}.
\end{aligned}$$

Proposition 3.2 allows us to apply Lemma 3.6 to the above, so we conclude that

$$\begin{aligned}
\delta_{\max}(\varepsilon) &\geq (\rho_{j_\varepsilon} + \rho_1)\varphi(0) - \inf_{s \in \mathbb{R}} \rho_{j_\varepsilon}\varphi(-s) + \rho_1 \varphi(s) \\
&= \delta(\varepsilon).
\end{aligned}$$

□

B. Proofs of results for Simplex Coding losses

In this section we state the proofs of the results given in Section 4. We start with a related lemma, showing that we do not need the sum-to-zero constraint for our results to hold: a sum-to-non-negative constraint suffices. We show through Lemma B.2 that we can transform the original surrogate loss minimization problem, which ultimately requires a constrained minimization over $\mathcal{C}$, into one that requires an unconstrained minimization over $\mathbb{R}^{K-1}$.

Lemma B.1. *Consider the set $\mathcal{C}' = \left\{ t \in \mathbb{R}^K : \mathbf{1}_K^\top t \geq 0 \right\}$, non-negative constants $\rho_1, \ldots, \rho_K$, as well as φ satisfying Assumption 2.1. Then for all $\varepsilon \geq 0$*

$$\inf_{t \in \mathcal{C}' \setminus \mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k \varphi_t(t_k) = \inf_{t \in \mathcal{C} \setminus \mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k \varphi_t(t_k).$$

Proof of Lemma B.1. In what follows, fix some $\varepsilon \geq 0$. Clearly, we know that $\mathcal{C} \subseteq \mathcal{C}'$ implies

$$\inf_{t \in \mathcal{C}' \setminus \mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k \varphi_t(t_k) \leq \inf_{t \in \mathcal{C} \setminus \mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k \varphi_t(t_k).$$

Consider a function σ as in Lemma 3.5, *i.e.*, if $\min_s \varphi(s)$ exists, let $s^* = \max\{\operatorname{argmin}_s \varphi(s)\}$ and define

$$\sigma(s) = \begin{cases} \varphi(s) & s \geq s^*, \\ \varphi(s^*) & s < s^*; \end{cases}$$

otherwise, let $\sigma = \varphi$. Note that σ is non-decreasing, $\varphi(s) \geq \sigma(s)$ for all $s \in \mathbb{R}$, and, from Lemma 3.5,

$$\inf_{t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k \varphi(t_k) = \inf_{t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k \sigma(t_k).$$

For any $t \in \mathcal{C}'\backslash\mathcal{M}(\varepsilon)$ s.t. $\mathbf{1}_K^\top t > 0$, let $i = \operatorname{argmin}_k t_k$ and construct t' s.t. $t'_k = t_k$ for $k \neq i$, and $t'_i = -\sum_{k \neq i} t_k$, so that $\mathbf{1}_K^\top t' = 0$, and $\bar{f}(t') = \bar{f}(t)$, *i.e.*, $t' \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)$. Furthermore, $t_i > t'_i$, and, since σ is non-decreasing,

$$\begin{aligned} \sum_{k=1}^{K} \rho_k \sigma(t_k) &= \sigma(t_i) + \sum_{k \neq i} \rho_k \sigma(t'_k) \\ &\geq \sigma(t'_i) + \sum_{k \neq i} \rho_k \sigma(t'_k) \\ &= \sum_{k=1}^{K} \rho_k \sigma(t'_k). \end{aligned}$$

Hence,

$$\begin{aligned} \inf_{t \in \mathcal{C}'\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k \varphi_t(t_k) &\geq \inf_{t \in \mathcal{C}'\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k \sigma(t_k) \\ &\geq \inf_{t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k \sigma(t_k) \\ &= \inf_{t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k \varphi_t(t_k) \end{aligned}$$

and our result follows.

□

We need a few definitions for Lemma B.2 and Theorem 4.1. Let $Q \in \mathbb{R}^{K \times K-1}$ s.t. its columns span $\mathcal{C}$, non-negative constants $\rho_1 \leq \cdots \leq \rho_K$, and for all $w \in \mathbb{R}^{K-1}, \varepsilon \geq 0$,

$$\begin{aligned} f_Q(w) &\in \operatorname*{argmax}_k q_k^\top w, \\ \bar{f}_Q(w) &= \max\left\{ j : j \in \operatorname*{argmax}_k q_k^\top w \right\}, \\ \mathcal{M}_Q(\varepsilon) &= \left\{ t \in \mathbb{R}^{K-1} : \rho_{\bar{f}_Q(w)} - \min_k \rho_k < \varepsilon \right\}, \end{aligned}$$

where q_k is the k-th row of Q.

Lemma B.2. *Consider $Q \in \mathbb{R}^{K \times K-1}$, s.t. the columns of Q span $\mathcal{C}$, as well as non-negative constants $\rho_1 \leq \cdots \leq \rho_K$. Let φ be a function satisfying Assumption 2.1. Denote by q_k the k-th row of Q. Then for all $\varepsilon \geq 0$,*

$$\inf_{w \in \mathbb{R}^{K-1}\backslash\mathcal{M}_Q(\varepsilon)} \sum_{k=1}^{K} \rho_k \varphi(q_k^\top w) = \inf_{t \in \mathcal{C}\backslash\mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k \varphi(t_k).$$

Proof of Lemma B.2. Since Q is orthogonal and its columns span $\mathcal{C}$, for every $t \in \mathcal{C}$ there exists exactly one $w \in \mathbb{R}^{K-1}$ s.t. $Qw = t$, and vice-versa, so $\left\{Q^\top w : w \in \mathbb{R}^{K-1}\right\} = \mathcal{C}$ and for every $t \in \mathcal{C}, w \in \mathbb{R}^{K-1}$ s.t. $Qw = t$,

$$\sum_{k=1}^{K} \rho_k \varphi(q_k^\top w) = \sum_{k=1}^{K} \rho_k \varphi(t_k),$$

$$\bar{f}_Q(w) = \max\left\{\operatorname*{argmax}_k q_k^\top w\right\} = \max\left\{\operatorname*{argmax}_k t_k\right\} = \bar{f}(t).$$

This immediately implies that for any $\varepsilon \geq 0$, $\mathcal{M}(\varepsilon) = \mathcal{M}_Q(\varepsilon)$ and that

$$\inf_{w \in \mathbb{R}^{K-1} \setminus \mathcal{M}_Q(\varepsilon)} \sum_{k=1}^{K} \rho_k \varphi(q_k^\top w) = \inf_{t \in \mathcal{C} \setminus \mathcal{M}(\varepsilon)} \sum_{k=1}^{K} \rho_k \varphi(t_k).$$

□

Theorem 4.1. *Let $\rho_1 \leq \cdots \leq \rho_K$ be $K > 0$ non-negative numbers, $Q \in \mathbb{R}^{K \times K-1}$ be s.t. the columns of Q span $\mathcal{C}$, and φ satisfy Assumptions 2.1 and 3.2. Denote by q_k the k-th row of Q as a column vector. For $\varepsilon \geq 0$ and $w \in \mathbb{R}^{K-1}$, define*

$$f_Q(w) \in \operatorname*{argmax}_k q_k^\top w\,, \qquad j_\varepsilon = \min\left\{j : \rho_j \geq \varepsilon + \rho_1\right\},$$

$$\delta(\varepsilon) = (\rho_{j_\varepsilon} + \rho_{j_0})\varphi(0) - \inf_{s \in \mathbb{R}} \left\{\rho_{j_\varepsilon}\varphi(s) + \rho_{j_0}\varphi(-s)\right\}.$$

Then, for all $\varepsilon > 0$ and $w \in \mathbb{R}^{K-1}$, it holds that

$$\sum_{k=1}^{K} \rho_k \varphi(q_k^\top w) - \inf_{w' \in \mathcal{C}} \sum_{k=1}^{K} \rho_k \varphi(q_k^\top w') < \delta(\varepsilon)$$

implies $\rho_{f_Q(w)} - \rho_1 < \varepsilon$. Furthermore, δ is a calibration function for these losses.

Proof of Theorem 4.1. It suffices to combine Lemma B.2 and Lemma 3.7 to obtain a lower-bound on $\delta_{\max}(\varepsilon)$. Since $\delta(\varepsilon) = \delta(\rho_{j_\varepsilon}, \rho_1)$ as defined in Lemma 3.3, we get that δ is positive for all $\varepsilon > 0$. This is true because $\rho_{j_\varepsilon} > \rho_1$ for all $\varepsilon > 0$, by definition of ρ_{j_ε}.

Since δ is positive and a lower-bound to $\delta_{\max}$, we get from Proposition 3.1 that it is a calibration function and we obtain the calibration result for $\bar{f}_Q$. All that remains is to note that for all $w \in \mathbb{R}^{K-1}$

$$\rho_{\bar{f}_Q(w)} - \rho_1 < \varepsilon \Rightarrow \rho_{f_Q(w)} - \rho_1 < \varepsilon.$$

□

Proposition 4.2. *The columns of a simplex-coding matrix C span $\mathcal{C}$.*

Proof of Proposition 4.2. Since $\sum_{k=1}^{K} c_k = C^\top \mathbf{1}_K = \mathbf{0}_{K-1}$, we have that the columns of C lie in $\mathcal{C}$. All we have to show now is that the column span of C contains $\mathcal{C}$, which is the nullspace of $\mathbf{1}_K \mathbf{1}_K^\top$. We prove this fact by showing that

$$\frac{(K-1)}{2} CC^\top + \frac{1}{2}\mathbf{1}_K \mathbf{1}_K^\top = \mathbf{I}_K.$$

To see that the above is true, we use that $(CC^\top)_{i,j} = -\frac{1}{K-1}$ for $i \neq j$, and that $(CC^\top)_{i,i} = 1$ for $1 \leq i \leq K$, so for $i \neq j$

$$\begin{aligned}\left(\frac{(K-1)}{2} CC^\top + \frac{1}{2}\mathbf{1}_K \mathbf{1}_K^\top\right)_{i,j} &= -\frac{K-1}{2(K-1)} + \frac{1}{2} \\ &= 0,\end{aligned}$$

and for $1 \leq i \leq K$

$$\begin{aligned}\left(\frac{(K-1)}{2}CC^\top + \frac{1}{2}\mathbf{1}_K\mathbf{1}_K^\top\right)_{i,i} &= \frac{1}{2} + \frac{1}{2} \\ &= 1.\end{aligned}$$

□

C. Calibration function calculations for the examples

In this section, we provide the calculations for the calibration functions presented in Table 1. To this end, let

$$\delta_\varphi(\rho_1, \rho_2) = (\rho_1 + \rho_2)\varphi(0) - \inf_s \{\rho_1\varphi(s) + \rho_2\varphi(-s)\},$$

for non-negative constants $\rho_1 < \rho_2$ and φ satisfying Assumption 2.1. If Assumption 3.2 is satisfied, then we have that for all $\varepsilon > 0$ s.t. j_ε exists, $\delta(\varepsilon) = \delta_\varphi(\rho_1, \rho_{j_\varepsilon})$ (cf. the definitions of j_ε and δ in Theorem 2.2). If j_ε does not exist, then we can take $\delta(\varepsilon) = \infty$ or $\delta(\varepsilon) = \delta_\varphi(\rho_1, \rho_K)$, for example. We need not concern ourselves with calculating $\delta_\varphi(\rho_1, \rho_2)$ when $\rho_1 = \rho_2$ because $\rho_{j_\varepsilon} > \rho_1$ for all $\varepsilon > 0$ s.t. j_ε exists.

Proposition C.1 (Hinge loss calibration function)**.** *When* $\varphi(s) = (1+s)_+$, $\delta_\varphi(\rho_1, \rho_2) = \rho_2 - \rho_1$.

Proof of Proposition C.1. Clearly, φ is convex, lower-bounded and $\partial\varphi(0) = \{1\} \subset (0, \infty)$, so Assumptions 2.1 and 3.2 are satisfied.

We have that

$$\inf_s \{\rho_1\varphi(s) + \rho_2\varphi(-s)\} = \min\{c_1, c_2, c_3\},$$

where

$$\begin{aligned}c_1 &= \inf_{s \leq -1} \{\rho_1\varphi(s) + \rho_2\varphi(-s)\} \\ c_2 &= \inf_{-1 \leq s \leq 1} \{\rho_1\varphi(s) + \rho_2\varphi(-s)\} \\ c_3 &= \inf_{s \geq 1} \{\rho_1\varphi(s) + \rho_2\varphi(-s)\}.\end{aligned}$$

So

$$\begin{aligned} c_1 &= \inf_{s\le -1} \{\rho_1\varphi(s) + \rho_2\varphi(-s)\} \\ &= \inf_{s\le -1} \{\rho_1(1+s)_+\rho_2(1-s)_+\} \\ &= \inf_{s\le -1} \rho_2(1-s) \\ &= 2\rho_2, \end{aligned}$$

$$\begin{aligned} c_2 &= \inf_{-1\le s\le 1} \{\rho_1\varphi(s) + \rho_2\varphi(-s)\} \\ &= \inf_{-1\le s\le 1} \{\rho_1(1+s)_+ + \rho_2(1-s)_+\} \\ &= \inf_{-1\le s\le 1} \{\rho_1(1+s) + \rho_2(1-s)\} \\ &= 2\rho_1, \end{aligned}$$

$$\begin{aligned} c_3 &= \inf_{s\ge 1} \{\rho_1\varphi(s) + \rho_2\varphi(-s)\} \\ &= \inf_{s\ge 1} \{\rho_1(1+s)_+ + \rho_2(1-s)_+\} \\ &= \inf_{s\ge 1} \rho_1(1+s) \\ &= 2\rho_1. \end{aligned}$$

Since $\varphi(0) = 1$, $\delta_\varphi(\rho_1, \rho_2) = \rho_2 - \rho_1$.

□

Proposition C.2 (Absolute-value loss calibration function)**.** *When* $\varphi(s) = |1+s|$, $\delta_\varphi(\rho_1, \rho_2) = \rho_2 - \rho_1$.

Proof of Proposition C.2. Clearly, φ is convex, lower-bounded and $\partial\varphi(0) = \{1\} \subset (0, \infty)$, so Assumptions 2.1 and 3.2 are satisfied.

The result follows from Lemma 3.5, if we choose $\varphi(s) = |1+s|$. In that case, we obtain $\sigma(s) = (1+s)_+$, so that $\delta_\varphi(\rho_1, \rho_2) = \delta_\sigma(\rho_1, \rho_2) = \rho_2 - \rho_1$, from Proposition C.1. □

Proposition C.3 (Squared loss calibration function)**.** *When* $\varphi(s) = (1+s)^2$, $\delta_\varphi(\rho_1, \rho_2) = \frac{(\rho_2-\rho_1)^2}{\rho_2+\rho_1}$.

Proof of Proposition C.3. Clearly, φ is convex, lower-bounded and $\partial\varphi(0) = \{2\} \subset (0, \infty)$, so Assumptions 2.1 and 3.2 are satisfied.

Because Assumptions 2.1 and 3.2 are satisfied, and because $\lim_{s\to\infty} \{\rho_1\varphi(s) + \rho_2\varphi(-s)\} = \lim_{s\to-\infty} \{\rho_1\varphi(s) + \rho_2\varphi(-s)\} = \infty$, we know $\inf_s \{\rho_1\varphi(s) + \rho_2\varphi(-s)\}$ is taken at some s^*. Because φ is differentiable, we know that

$$\frac{d}{ds} \{\rho_1\varphi(s) + \rho_2\varphi(-s)\}\Big|_{s^*} = 0,$$

so

$$2\rho_1(1+s^*) = 2\rho_2(1-s^*),$$

which implies $s^* = \frac{\rho_2-\rho_1}{\rho_2+\rho_1}$. Hence

$$\begin{aligned} \delta_\varphi(\rho_1, \rho_2) &= (\rho_1+\rho_2) - \rho_1(1+s^*)^2 - \rho_2(1-s^*)^2 \\ &= \frac{(\rho_2-\rho_1)^2}{\rho_2+\rho_1}. \end{aligned}$$

□

Proposition C.4 (Truncated squared loss calibration function). *When* $\varphi(s) = [(1+s)_+]^2$, $\delta_\varphi(\rho_1, \rho_2) = \frac{(\rho_2-\rho_1)^2}{\rho_2+\rho_1}$.

Proof of Proposition C.4. Clearly, φ is convex, lower-bounded and $\partial\varphi(0) = \{2\} \subset (0, \infty)$, so Assumptions 2.1 and 3.2 are satisfied.

The result follows from Lemma 3.5, if we choose $\varphi(s) = (1+s)^2$. In that case, we obtain $\sigma(s) = [(1+s)_+]^2$, so that $\delta_\sigma(\rho_1, \rho_2) = \delta_\varphi(\rho_1, \rho_2) = \frac{(\rho_2-\rho_1)^2}{\rho_2+\rho_1}$, from Proposition C.3.

□

Proposition C.5 (Exponential loss calibration function). *When* $\varphi(s) = e^s$, *if* $\rho_1 = 0$, *then* $\delta_\varphi(\rho_1, \rho_2) = \rho_2$, *and, if* $\rho_1 > 0$, *then* $\delta_\varphi(\rho_1, \rho_2) = (\sqrt{\rho_2} - \sqrt{\rho_1})^2$.

In particular, if $\rho_1 < \rho_2$ *and* $\rho_1 + \rho_2 = 1$, *then* $\delta_\varphi(\rho_1, \rho_2) = 1 - \sqrt{1 - (\rho_2 - \rho_1)^2}$.

Proof of Proposition C.5. Clearly, φ is convex, and $\partial\varphi(0) = \{1\} \subset (0, \infty)$, so Assumption 2.1 is satisfied. When $\rho_1 = 0$, we have $\inf_s \{\rho_1\varphi(s) + \rho_2\varphi(-s)\} = 0$ (no minimum exists), so $\delta_\varphi(\rho_1, \rho_2) = (\rho_1 + \rho_2)e^0 = \rho_2$. Otherwise, *i.e.*, if $\rho_1 > 0$, then $\lim_{s\to\infty} \{\rho_1\varphi(s) + \rho_2\varphi(-s)\} = \lim_{s\to-\infty} \{\rho_1\varphi(s) + \rho_2\varphi(-s)\} = \infty$, Assumption 3.2 is satisfied and the infimum is taken at some s^*. Since e^s is differentiable,

$$\frac{d}{ds} \{\rho_1\varphi(s) + \rho_2\varphi(-s)\}\Big|_{s^*} = 0,$$

so

$$\rho_1 e^{s^*} = \rho_2 e^{-s^*},$$

which implies $s^* = \frac{1}{2} \ln \frac{\rho_2}{\rho_1}$. Hence

$$\begin{aligned}\delta_\varphi(\rho_1, \rho_2) &= (\rho_1 + \rho_2) - \rho_1\sqrt{\frac{\rho_2}{\rho_1}} - \rho_2\sqrt{\frac{\rho_1}{\rho_2}} \\ &= (\sqrt{\rho_2} - \sqrt{\rho_1})^2.\end{aligned}$$

The second statement is easy to verify from the above. □

Proposition C.6 (Logistic loss calibration function). *When* $\varphi(s) = \ln(1 + e^s)$, *if* $\rho_1 = 0$, *then* $\delta_\varphi(\rho_1, \rho_2) = \rho_2$, *and, if* $\rho_1 > 0$, *then* $\delta_\varphi(\rho_1, \rho_2) = (\rho_1 + \rho_2)\left[H(\rho_1, \rho_2) - H\left(\frac{1}{2}, \frac{1}{2}\right)\right]$, *with* $H(a, b) = \frac{a}{a+b} \ln\left(\frac{a}{a+b}\right) + \frac{b}{a+b} \ln\left(\frac{b}{a+b}\right)$ *for* $a, b > 0$.

Proof of Proposition C.6. Clearly, φ is convex, and $\partial\varphi(0) = \{1\} \subset (0, \infty)$, so Assumption 2.1 is satisfied. When $\rho_1 = 0$, we have $\inf_s \{\rho_1\varphi(s) + \rho_2\varphi(-s)\} = 0$ (no minimum exists), so $\delta_\varphi(\rho_1, \rho_2) = (\rho_1 + \rho_2)\ln(1 + e^0) = \rho_2$. Otherwise, *i.e.*, if $\rho_1 > 0$, then $\lim_{s\to\infty} \{\rho_1\varphi(s) + \rho_2\varphi(-s)\} = \lim_{s\to-\infty} \{\rho_1\varphi(s) + \rho_2\varphi(-s)\} = \infty$, Assumption 3.2 is satisfied and the infimum is taken at some s^*. Since $\ln(1 + e^s)$ is differentiable,

$$\frac{d}{ds} \{\rho_1\varphi(s) + \rho_2\varphi(-s)\}\Big|_{s^*} = 0,$$

so

$$\rho_1 \frac{1}{1 + e^{-s^*}} = \rho_2 \frac{1}{1 + e^{s^*}},$$

which implies $s^* = \ln \frac{\rho_2}{\rho_1}$. Hence,

$$\begin{aligned}\delta_\varphi(\rho_1, \rho_2) &= (\rho_1 + \rho_2) \ln 2 - \rho_1 \ln\left(1 + \frac{\rho_2}{\rho_1}\right) - \rho_2 \ln\left(1 + \frac{\rho_1}{\rho_2}\right) \\ &= (\rho_1 + \rho_2)\left[H(\rho_1, \rho_2) - H\left(\frac{1}{2}, \frac{1}{2}\right)\right].\end{aligned}$$

□