Proceedings of Machine Learning Research vol 125:1–43, 2020 33rd Annual Conference on Learning Theory

Calibrated Surrogate Losses for Adversarially Robust Classification

Han Bao* TSUTSUMI@MS.K.U-TOKYO.AC.JP
The University of Tokyo
RIKEN AIP

Clayton Scott CLAYSCOT@UMICH.EDU
University of Michigan

Masashi Sugiyama SUGI@K.U-TOKYO.AC.JP
RIKEN AIP
The University of Tokyo

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

Adversarially robust classification seeks a classifier that is insensitive to adversarial perturbations of test patterns. This problem is often formulated via a minimax objective, where the target loss is the worst-case value of the 0-1 loss subject to a bound on the size of perturbation. Recent work has proposed convex surrogates for the adversarial 0-1 loss, in an effort to make optimization more tractable. In this work, we consider the question of which surrogate losses are *calibrated* with respect to the adversarial 0-1 loss, meaning that minimization of the former implies minimization of the latter. We show that no convex surrogate loss is calibrated with respect to the adversarial 0-1 loss when restricted to the class of linear models. We further introduce a class of nonconvex losses and offer necessary and sufficient conditions for losses in this class to be calibrated.

Keywords: surrogate loss, classification calibration, adversarial robustness

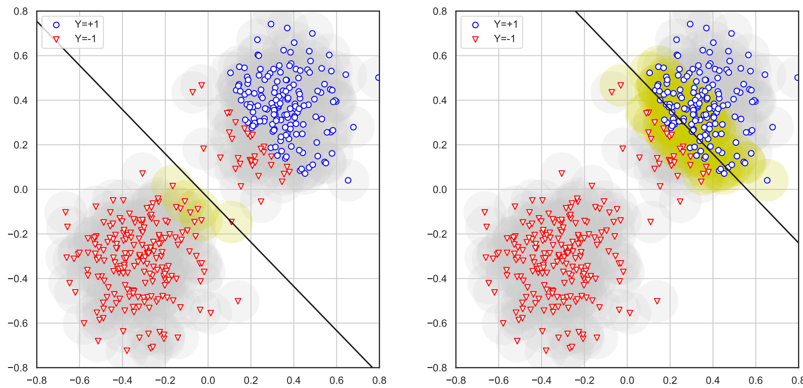
1. Introduction

In conventional machine learning, training and testing instances are assumed to follow the same probability distribution. In *adversarially robust* machine learning, test instances may be perturbed by an adversary before being presented to the predictor. Recent work has shown that seemingly insignificant adversarial perturbations can lead to significant performance degradations of otherwise highly accurate classifiers (Goodfellow et al., 2015). This has led to the development of a number of methods for learning predictors with decreased sensitivity to adversarial perturbations (Xu et al., 2009; Xu and Mannor, 2012; Goodfellow et al., 2015; Cisse et al., 2017; Wong and Kolter, 2018; Raghunathan et al., 2018a; Tsuzuku et al., 2018).

Adversarially robust classification is typically formulated as empirical risk minimization with an *adversarial 0-1 loss*, which is the maximum of the usual 0-1 loss over a set of possible perturbations of the test instance. This minimax optimization problem is nonconvex, and recent work, reviewed in Section 4, has proposed several convex surrogate losses. However, it is still unknown whether minimizing these convex surrogates leads to minimization of the adversarial 0-1 loss.

In this work, we examine the question of which surrogate losses are calibrated with respect to (wrt) the adversarial 0-1 loss. A surrogate loss is said to be *calibrated* wrt a target loss if minimiza-

* This work was performed while the first author was a visitor at University of Michigan.



(*a*) Ramp loss ($\beta = 0.5$) (*b*) Hinge loss ($\beta = 0.5$)

Figure 1: The best linear classifier under each loss. The shift parameter β for a surrogate loss is defined in Section 8. The ℓ_2-balls associated to each instance indicate adversarial perturbations with radii 0.1. The yellow balls indicate instances vulnerable to perturbations, in that they are within 0.1 of the decision boundary. In this example, 1.2% of instances are vulnerable under the ramp loss, while 24.8% of instances are vulnerable under the hinge loss.

tion of the excess surrogate risk (over a specified class of decision functions) implies minimization of the excess target risk. Employing the calibration function perspective of Steinwart (2007), we show that no convex surrogate loss is calibrated wrt the adversarial 0-1 loss when restricted to the class of linear models (Section 6). Intuitively, this is because convex losses prefer predictions close to the decision boundary on average when $\mathbb{P}(Y = +1|X) \approx \frac{1}{2}$, while predictions that are too close to the decision boundary should be penalized in adversarially robust classification. We also provide necessary and sufficient conditions for a certain class of nonconvex losses to be calibrated wrt the adversarial 0-1 loss (Section 7), and provide excess risk bounds that quantify the relationship between the excess surrogate and target risks. These calibrated losses attain robustness by penalizing predictions that are too close to the decision boundary. To our knowledge, this is the first work to formally analyze the adversarial 0-1 loss by calibration analysis. Our analysis depends on the fact that the adversarially robust 0-1 loss equals the horizontally shifted (non-robust) 0-1 loss when restricted to linear models (Proposition 1). In summary, we argue against the use of convex losses in adversarially robust classification (with linear models), and calibrated nonconvex losses serve as good alternatives.

Our results demonstrate that adversarial robustness requires different surrogates than other notions of robustness. For example, symmetric losses such as the sigmoid and ramp losses are robust to label noise (Ghosh et al., 2015), but not calibrated wrt the adversarial 0-1 loss. Figure 1 illustrates the results of learning a linear classifier with respect a *shifted* ramp loss, which is calibrated wrt the adversarial 0-1 loss, and a shifted hinge loss, which is not (these losses are discussed in detail later). While the hinge loss yields a classifier with smaller misclassification rate wrt the conventional 0-1 loss, this classifier is quite sensitive to small perturbations of the test instances. The classifier learned by the ramp loss, on the other hand, makes fewer errors when subjected to adversarial perturbations.

The rest of this paper is organized as follows. Section 3 formalizes notation and the problem. Related work on robust learning and calibration analysis is reviewed in Section 4. Technical details of calibration analysis are reviewed in Section 5. Section 6 describes the nonexistence of convex calibrated surrogate losses, while Section 7 presents general calibration conditions for a certain class of nonconvex losses. Section 8 applies our theory to several convex and nonconvex losses, and presents excess risk bounds for the calibrated nonconvex losses. Section 9 shows simulation results to verify that calibrated losses achieve target excess risk close to zero under the robust 0-1 loss. Conclusions are stated in Section 10.

2. Notation

Let $\|x\|_p$ for a vector $x \in \mathbb{R}^d$ be the ℓ_p-norm, namely, $\|x\|_p = \sqrt[p]{\sum_{i=1}^d |x_i|^p}$. Let $B_p^d(r) \stackrel{\text{def}}{=} \{v \in \mathbb{R}^d \mid \|v\|_p \le r\}$ be the d-dimensional ℓ_p-ball with radius r. The set $\{1, \dots, n\}$ is denoted by $[n]$. The indicator function corresponding to an event A is denoted by $\mathbb{1}_{\{A\}}$. We define the infimum over the empty set as $+\infty$. Denote $h \equiv c$ for a function $h : S \to \mathbb{R}$ and $c \in \mathbb{R}$ if $h(x) = c$ for all $x \in \mathbf{dom}(h)$, where $\mathbf{dom}(h)$ denotes the domain of a function h, and $h \not\equiv c$ otherwise. For a function $h : S \to \mathbb{R}$, we write $h^{**} : S \to \mathbb{R}$ for the Fenchel-Legendre biconjugate of h, characterized by $\mathrm{epi}(h^{**}) = \overline{\mathrm{co}}\,\mathrm{epi}(h)$, where $\overline{\mathrm{co}}\,S$ is the closure of the convex hull of the set S, and $\mathrm{epi}(h)$ is the epigraph of the function h: $\mathrm{epi}(h) \stackrel{\text{def}}{=} \{(x, t) \mid x \in S, h(x) \le t\}$. A function $h : S \to \mathbb{R}$ is said to be *quasiconcave* if for all $x_1, x_2 \in S$ and $\lambda \in [0, 1]$, $h(\lambda x_1 + (1 - \lambda)x_2) \ge \min\{h(x_1), h(x_2)\}$.

Let $\mathcal{X} \stackrel{\text{def}}{=} B_2^d(1)$ be the feature space, $\mathcal{Y} \stackrel{\text{def}}{=} \{\pm 1\}$ be the binary label space, and $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ be a function class. We consider *symmetric* $\mathcal{F}$, that is, $-f \in \mathcal{F}$ for all $f \in \mathcal{F}$. We write $\mathcal{F}_{\text{all}} \subseteq \mathbb{R}^{\mathcal{X}}$ for the space of all measurable functions. Let $\ell : \mathcal{Y} \times \mathcal{X} \times \mathcal{F} \to \mathbb{R}_{\ge 0}$ be a loss function. Then, we write $\mathcal{R}_\ell(f) \stackrel{\text{def}}{=} \mathbb{E}_{(X,Y)}[\ell(Y, X, f)]$ for the ℓ-*risk* of $f \in \mathcal{F}$. If ℓ can be represented by $\ell(y, x, f) = \phi(yf(x))$ with some $\phi : \mathbb{R} \to \mathbb{R}_{\ge 0}$ for any $y \in \mathcal{Y}$, $x \in \mathcal{X}$, and $f \in \mathcal{F}$, ϕ is called a *margin-based* loss function. We define the ϕ-*risk* of $f \in \mathcal{F}$ for a margin-based loss ϕ by

$$\mathcal{R}_\phi(f) \stackrel{\text{def}}{=} \mathbb{E}_{(X,Y)}[\phi(Yf(X))] = \mathbb{E}_X \mathbb{E}_{Y|X}[\phi(Yf(X))], \tag{1}$$

where $\mathbb{E}_X$ and $\mathbb{E}_{Y|X}$ mean the expectation over $\mathbb{P}(X)$ and $\mathbb{P}(Y|X)$, respectively. We can rewrite (1) as $\mathcal{R}_\phi(f) = \mathbb{E}_X[\mathcal{C}_\phi(f(X), \mathbb{P}(Y = +1|X))]$ with $\mathcal{C}_\phi(\alpha, \eta) \stackrel{\text{def}}{=} \eta\phi(\alpha) + (1 - \eta)\phi(-\alpha)$. We call $\mathcal{C}_\phi(\alpha, \eta)$ the *class-conditional* ϕ-*risk* (ϕ-CCR). The minimal ϕ-risk $\mathcal{R}^*_{\phi,\mathcal{F}} \stackrel{\text{def}}{=} \inf_{f \in \mathcal{F}} \mathcal{R}_\phi(f)$ is called the *Bayes* (ϕ, $\mathcal{F}$)*-risk*, and the minimal ϕ-CCR on $\mathcal{F}$ is denoted by $\mathcal{C}^*_{\phi,\mathcal{F}}(\eta) \stackrel{\text{def}}{=} \inf_{\alpha \in \mathcal{A}_\mathcal{F}} \mathcal{C}_\phi(\alpha, \eta)$, where $\mathcal{A}_\mathcal{F} \stackrel{\text{def}}{=} \{\alpha = f(x) \mid f \in \mathcal{F}, x \in \mathcal{X}\}$. We refer to $\mathcal{R}_\phi(f) - \mathcal{R}^*_{\phi,\mathcal{F}}$ as the $(\phi, \mathcal{F})$-excess risk. We occasionally use the abbreviation $\Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha, \eta) \stackrel{\text{def}}{=} \mathcal{C}_\phi(\alpha, \eta) - \mathcal{C}^*_{\phi,\mathcal{F}}(\eta)$.

3. Surrogate Losses for Adversarial Robust Classification

In supervised binary classification, a learner is asked to output a predictor $f : \mathcal{X} \to \mathbb{R}$ that minimizes the classification error $\mathbb{P}\{Yf(X) \le 0\}$, where $\mathbb{P}$ is the unknown underlying distribution. This can be equivalently interpreted as the minimization of the risk $\mathbb{E}_{(X,Y)}[\ell_{01}(Y, X, f)]$ wrt f, where

$$\ell_{01}(y, x, f) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } yf(x) \le 0, \\ 0 & \text{otherwise} \end{cases}$$

is the 0-1 loss. Letting $\phi_{01}(\alpha) \stackrel{\text{def}}{=} \mathbb{1}_{\{\alpha \le 0\}}$, then $\ell_{01}(y, x, f) = \phi_{01}(yf(x))$. On the other hand, an adversarially robust learner is asked to output a predictor f that minimizes the 0-1 loss while being tolerant to small perturbations to input data points. Following existing literature (Xu et al., 2009;

Tsuzuku et al., 2018; Bubeck et al., 2019), we consider ℓ_2-ball perturbations and define the goal as the minimization of $\mathbb{P}\{\exists\Delta_x \in B_2^d(\gamma)$ s.t. $X+\Delta_x \in \mathcal{X}$ and $Yf(X+\Delta_x) \le 0\}$, where Δ_x is a perturbation vector and $\gamma \in (0,1)$ is a pre-defined perturbation budget. Equivalently, the goal of adversarially robust classification is to minimize $\mathbb{E}_{(X,Y)}[\ell_\gamma(Y,X,f)]$ wrt f, where

$$\ell_\gamma(y,x,f) \overset{\text{def}}{=} \begin{cases} 1 & \text{if } \exists\Delta_x \in B_2^d(\gamma) \text{ s.t. } x+\Delta_x \in \mathcal{X} \text{ and } yf(x+\Delta_x) \le 0, \\ 0 & \text{otherwise.} \end{cases}$$

We call this loss function ℓ_γ the *adversarially robust 0-1 loss*, or the *robust 0-1 loss* for short.

The robust 0-1 loss is also a margin-based loss when restricted to the class of linear models $\mathcal{F}_{\text{lin}} \overset{\text{def}}{=} \{x \mapsto \theta^\top x \mid \theta \in \mathbb{R}^d, \|\theta\|_2 = 1\} \subseteq \mathbb{R}^{\mathcal{X}}$. Note that $\mathcal{F}_{\text{lin}}$ is symmetric.

Proposition 1 *For any $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $f \in \mathcal{F}_{\text{lin}}$, we have $\ell_\gamma(y,x,f) = \mathbb{1}_{\{yf(x)\le\gamma\}}$.*

We include the proof in Appendix B for completeness though it is mentioned as a fact by Diakonikolas et al. (2019). Subsequently, when considering $\mathcal{F}_{\text{lin}}$, we work with the loss function $\phi_\gamma(\alpha) \overset{\text{def}}{=} \mathbb{1}_{\{\alpha\le\gamma\}}$ and call ϕ_γ the *γ-robust 0-1 loss*. We will study calibrated surrogates wrt ϕ_γ instead of ℓ_γ, and both are equivalent under the restricted function class $\mathcal{F}_{\text{lin}}$. We can view ϕ_γ as a shifted version of ϕ_{01}.

In many machine learning problems, there are often dichotomies between optimization (learning) and evaluation. For instance, binary classification is evaluated by the 0-1 loss, while common learning methods such as SVM and logistic regression minimize surrogates to the 0-1 loss. This dichotomy arises because minimizing the 0-1 loss directly is known to be NP-hard (Feldman et al., 2012). Much research has investigated surrogates ϕ satisfying

$$\mathcal{R}_\phi(f_i) - \mathcal{R}^*_{\phi,\mathcal{F}} \to 0 \implies \mathcal{R}_\ell(f_i) - \mathcal{R}^*_{\ell,\mathcal{F}} \to 0, \tag{2}$$

for all probability distributions and sequence of $\{f_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}$.

Our learning goal is to minimize the expected γ-robust 0-1 loss on a given function class $\mathcal{F}$:

$$\min_{f\in\mathcal{F}} \mathcal{R}_{\ell_\gamma}(f). \tag{3}$$

In order to solve (3), we aim to characterize surrogate losses ϕ satisfying (2) with $\ell = \ell_\gamma$ and $\mathcal{F} = \mathcal{F}_{\text{lin}}$. By Proposition 1, we have $\mathcal{R}_{\ell_\gamma}(f) = \mathcal{R}_{\phi_\gamma}(f)$ when $\mathcal{F} = \mathcal{F}_{\text{lin}}$.

4. Related Work

From the viewpoint of robust optimization (Ben-Tal et al., 2009; Bertsimas et al., 2011), adversarially robust binary classification can be formulated as

$$\min_{f\in\mathcal{F}} \mathbb{E}_{(X,Y)}\left[\max_{\widetilde{X}\in\mathcal{U}(X)} \ell(Y,\widetilde{X},f)\right], \tag{4}$$

where ℓ is a loss function and $\mathcal{U}(x)$ is a user-specified uncertainty set. Our formulation of adversarially robust classification (3) can be regarded as the special case $\ell = \ell_{01}$ and $\mathcal{U}(x) = x + B_2^d(\gamma)$.

Since the minimax problem (4) is generally nonconvex, it is traditionally tackled by minimizing a convex upper bound. Lanckriet et al. (2002) and Shivaswamy et al. (2006) pick $\mathcal{U}(x) = \{x \sim$

$(\bar{x}, \Sigma_x)\}$ as an uncertainty set, where $x \sim (\bar{x}, \Sigma_x)$ means that x is drawn from a distribution that has prespecified mean $\bar{x}$, covariance Σ_x, and arbitrary higher moments. Lanckriet et al. (2002) and Shivaswamy et al. (2006) convexified (4) and obtained a second-order cone program. Xu et al. (2009) studied the relationship between robustness and regularization, and showed that (4) with the hinge loss and $\mathcal{U}(x) = x + B_2^d(\gamma)$ is equivalent to ℓ_2-regularized SVM. Recently, Wong and Kolter (2018), Madry et al. (2018), Raghunathan et al. (2018a), Raghunathan et al. (2018b), and Khim and Loh (2019) examined (4) with the softmax cross entropy loss and $\mathcal{U}(x) = x + B_\infty^d(\gamma)$ when $\mathcal{F}$ is a set of deep nets, and provided convex upper bounds of the worst-case loss in (4). However, no work except Cranko et al. (2019) studied whether the surrogate objectives minimize the robust 0-1 excess risk. Cranko et al. (2019) showed that no canonical proper loss (Reid and Williamson, 2010) can minimize the robust 0-1 loss. Since canonical proper losses are convex, this result aligns with ours. We show more general results via calibration analysis for $\mathcal{U}(x) = x + B_2^d(\gamma)$.

There are several other approaches to the robust classification such as minimizing the Taylor approximation of the worst-case loss in (4) (Goodfellow et al., 2015; Gu and Rigazio, 2015; Shaham et al., 2018), regularization on the Lipschitz norm of models (Cisse et al., 2017; Hein and Andriushchenko, 2017; Tsuzuku et al., 2018), and injection of random noises to model parameters (Lecuyer et al., 2019; Cohen et al., 2019; Pinot et al., 2019; Salman et al., 2019). It is not known whether these methods imply the minimization of the robust 0-1 excess risk.

Other forms of robustness have also been considered in the literature. A number of existing works considered the worst-case test distribution. This line includes divergence-based methods (Namkoong and Duchi, 2016, 2017; Hu et al., 2018; Sinha et al., 2018), domain adaptation (Mansour et al., 2009; Ben-David et al., 2010; Germain et al., 2013; Kuroki et al., 2019; Zhang et al., 2019b), and methods based on constraints on feature moments (Farnia and Tse, 2016; Fathony et al., 2016).

In addition to adversarial robustness, it is worthwhile to mention outlier and label-noise robustness. It is known that convex losses are vulnerable to outliers, thus truncation making losses nonconvex is useful (Huber, 2011). In the machine learning context, Masnadi-Shirazi and Vasconcelos (2009) and Holland (2019) designed nonconvex losses robust to outliers. On the other hand, label-noise robustness, especially the random classification noise model, has been studied extensively (Angluin and Laird, 1988), where training labels are flipped with a fixed probability. Long and Servedio (2010) showed that there is no convex loss that is robust to label noises. Later, Ghosh et al. (2015), van Rooyen et al. (2015), and Charoenphakdee et al. (2019) discovered a certain class of nonconvex losses is a good alternative for label-noise robustness. In both outlier and label-noise robustness, nonconvex loss functions play an important role as we see in adversarial robustness.

Calibration analysis has been formalized in Lin (2004), Zhang et al. (2004), Bartlett et al. (2006), and Steinwart (2007), and employed to analyze not only binary classification, but also complicated problems such as multi-class classification (Zhang, 2004; Tewari and Bartlett, 2007; Long and Servedio, 2013; Ávila Pires and Szepesvári, 2016; Ramaswamy and Agarwal, 2016), multi-label classification (Gao and Zhou, 2011; Dembczynski et al., 2012), cost-sensitive learning (Scott, 2011, 2012; Ávila Pires et al., 2013), ranking (Duchi et al., 2010; Ravikumar et al., 2011; Ramaswamy et al., 2013), structured prediction (Hazan et al., 2010; Ramaswamy and Agarwal, 2012; Osokin et al., 2017; Blondel, 2019), AUC optimization (Gao and Zhou, 2015), and optimization of non-decomposable metrics (Bao and Sugiyama, 2020). Zhang et al. (2004), Ravikumar et al. (2011), and Gao and Zhou (2015) figured out *ad hoc* derivations of excess risk bounds, while Bartlett et al. (2006), Steinwart (2007), Scott (2012), Ávila Pires et al. (2013), Ávila Pires and Szepesvári (2016), Osokin et al. (2017), and Blondel (2019) used more systematic approaches. As for adversarially robust classification, Zhang et al. (2019a, Theorem 3.1) applied the classical result of calibration analysis on convex losses to upper bound the robust classification risk, resulting in a term requiring numerical approximation in practice.

5. Calibration Analysis

Calibration analysis is a tool to study the relationship between surrogate losses and target losses. This section is devoted to explaining the calibration function introduced in Steinwart (2007) and specializing it to the current paper.[1]

Definition 2 *For a loss $\psi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ and a function class $\mathcal{F}$, we say a loss $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is calibrated wrt (ψ,$\mathcal{F}$), or (ψ,$\mathcal{F}$)-calibrated, if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all $\eta \in [0,1]$ and $\alpha \in \mathcal{A}_{\mathcal{F}}$, we have*

$$\mathcal{C}_\phi(\alpha,\eta) < \mathcal{C}^*_{\phi,\mathcal{F}}(\eta) + \delta \Longrightarrow \mathcal{C}_\psi(\alpha,\eta) < \mathcal{C}^*_{\psi,\mathcal{F}}(\eta) + \varepsilon. \tag{5}$$

If ϕ is (ψ,$\mathcal{F}$)-calibrated, the condition (2) holds for any probability distribution on $\mathcal{X} \times \mathcal{Y}$ (Steinwart, 2007, Theorem 2.8). Thus, consistency of a learner wrt the ϕ-risk implies consistency wrt the ψ-risk.

Next, we introduce the *calibration function* (Steinwart, 2007, Lemma 2.16).

Definition 3 *For a margin-based loss ψ and ϕ, and a function class $\mathcal{F}$, the calibration function of ϕ wrt (ψ, $\mathcal{F}$), or simply calibration function if the context is clear, is defined as*

$$\delta(\varepsilon) = \inf_{\eta\in[0,1]} \inf_{\alpha\in\mathcal{A}_{\mathcal{F}}} \mathcal{C}_\phi(\alpha,\eta) - \mathcal{C}^*_{\phi,\mathcal{F}}(\eta) \quad \textit{s.t. } \mathcal{C}_\psi(\alpha,\eta) - \mathcal{C}^*_{\psi,\mathcal{F}}(\eta) \geq \varepsilon. \tag{6}$$

Note that $\delta(\varepsilon)$ is nondecreasing for $\varepsilon > 0$. The calibration function $\delta(\varepsilon)$ is the maximal δ satisfying the CCR condition (5). Steinwart (2007) established the following two important results.

Proposition 4 (Lemma 2.9 in Steinwart (2007)) *A surrogate loss ϕ is (ψ,$\mathcal{F}$)-calibrated if and only if its calibration function δ satisfies $\delta(\varepsilon) > 0$ for all $\varepsilon > 0$.*

Proposition 5 (Theorem 2.13 in Steinwart (2007)) *Let $\delta : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be the calibration function of ϕ wrt (ψ, $\mathcal{F}$). Define $\check{\delta} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ as $\check{\delta}(\varepsilon) = \delta(\varepsilon)$ if $\varepsilon > 0$ and $\check{\delta}(0) = 0$. Then, for all $f \in \mathcal{F}$, we have*

$$\check{\delta}^{**}\left(\mathcal{R}_\psi(f) - \mathcal{R}^*_{\psi,\mathcal{F}}\right) \leq \mathcal{R}_\phi(f) - \mathcal{R}^*_{\phi,\mathcal{F}}, \tag{7}$$

where $\check{\delta}^{**}$ denotes the Fenchel-Legendre biconjugate of $\check{\delta}$.

The relationship in (7) is called an excess risk transform. The excess risk transform is invertible iff ϕ is (ψ,$\mathcal{F}$)-calibrated (Steinwart, 2007, Remark 2.14). In this case, we obtain the excess risk bound $\mathcal{R}_\psi(f) - \mathcal{R}^*_{\psi,\mathcal{F}} \leq (\check{\delta}^{**})^{-1}(\mathcal{R}_\phi(f) - \mathcal{R}^*_{\phi,\mathcal{F}})$. In the end, the calibration function can be used in two

1. We import toolsets from Steinwart (2007) because of two reasons: (i) Steinwart (2007) formalized calibration analysis that is *dependent* on user-specified function classes, which is useful for our analysis on $\mathcal{F}_{\text{lin}}$. (ii) Steinwart (2007) gave a general form of the calibration function (6), while most of literature focuses on specific target losses.

ways: Proposition 4 enables us to check if a surrogate loss is calibrated, and Proposition 5 gives us a quantitative relationship between the surrogate excess risk and the target excess risk. Such an analysis has been carried out in a number of learning problems as we mention in Section 4.

Next, we review an important result regarding convex surrogates for the non-robust 0-1 loss ϕ_{01}.

Proposition 6 (Theorem 6 in Bartlett et al. (2006)) *Let ϕ be a convex surrogate loss. Then, ϕ is calibrated wrt (ϕ_{01},$\mathcal{F}_{\text{all}}$) if and only if it is differentiable at* 0 *and* $\phi'(0) < 0$.

As a result of Proposition 6, we know that many surrogate losses used in practice such as the hinge loss, logistic loss, and squared loss are calibrated wrt (ϕ_{01},$\mathcal{F}_{\text{all}}$).

Finally, we characterize the calibration function of an arbitrary surrogate loss ϕ wrt ϕ_γ. Its proof is deferred in Appendix B.

Lemma 7 *Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ be a function class such that $\mathcal{A}_{\mathcal{F}} \supseteq [-1, 1]$. For a surrogate loss ϕ, the (ϕ_γ,$\mathcal{F}$)-calibration function is $\delta(\varepsilon) = \inf_{\eta\in[0,1]} \bar{\delta}(\varepsilon, \eta)$, where*

$$\bar{\delta}(\varepsilon, \eta) = \begin{cases} \infty & \text{if } \varepsilon > \max\{\eta, 1-\eta\}, \\ \inf\limits_{|\alpha|\le\gamma} \Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha, \eta) & \text{if } |2\eta-1| < \varepsilon \le \max\{\eta, 1-\eta\}, \\ \inf\limits_{\alpha\in\mathcal{A}_{\mathcal{F}}:(2\eta-1)\alpha\le 0 \text{ or } |\alpha|\le\gamma} \Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha, \eta) & \text{if } \varepsilon \le |2\eta-1|. \end{cases} \tag{8}$$

Lemma 7 is used in the proofs and examples below. Note that $\mathcal{A}_{\mathcal{F}_{\text{lin}}} = [-1, 1]$ and $\mathcal{A}_{\mathcal{F}_{\text{all}}} = \mathbb{R}$.

6. Convex Surrogates are Not (ϕ_γ, $\mathcal{F}_{\text{all}}$)-calibrated

Our first result concerns calibration of convex surrogate losses wrt the γ-robust 0-1 loss.

Theorem 8 *For any margin-based surrogate loss $\phi : \mathbb{R} \to \mathbb{R}_{\ge 0}$ and function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ such that $\mathcal{A}_{\mathcal{F}} \supseteq [-1, 1]$, if ϕ is convex, then ϕ is not calibrated wrt (ϕ_γ,$\mathcal{F}$).*

Corollary 9 *For any margin-based surrogate loss $\phi : \mathbb{R} \to \mathbb{R}_{\ge 0}$, if ϕ is convex, then ϕ is not calibrated wrt (ϕ_γ,$\mathcal{F}_{\text{lin}}$), nor is it calibrated wrt (ϕ_γ,$\mathcal{F}_{\text{all}}$).*

Proof *(Sketch)* Here we focus on function class $\mathcal{F}_{\text{all}}$. In the non-robust setup, Bartlett et al. (2006) showed that a surrogate loss is calibrated wrt (ϕ_{01},$\mathcal{F}_{\text{all}}$) iff $\inf_{(2\eta-1)\alpha\le 0} \mathcal{C}_\phi(\alpha, \eta)$ (the minimum ϕ-risk over 'wrong' predictions) is larger than $\inf_{\alpha\in\mathbb{R}} \mathcal{C}_\phi(\alpha, \eta)$ (the minimum ϕ-risk over all predictions) for $\eta \ne \frac{1}{2}$. This means wrong predictions must be penalized more. In our robust setup, we must penalize not only wrong predictions but also predictions that fall in the γ-margin, i.e.,

$$\inf_{|\alpha|\le\gamma} \mathcal{C}_\phi(\alpha, \eta) > \inf_{\alpha\in\mathbb{R}} \mathcal{C}_\phi(\alpha, \eta), \tag{9}$$

which is an immediate corollary of Proposition 4 and Lemma 7 and stated in part 3 of Lemma 12 in Appendix B. Condition (9) becomes harder to satisfy as a data point gets more uncertain ($\eta \to \frac{1}{2}$). In the limit, we have $\inf_{|\alpha|\le\gamma} \phi(\alpha) + \phi(-\alpha) > \inf_{\alpha\in\mathbb{R}} \phi(\alpha) + \phi(-\alpha)$, meaning that the even part of ϕ "should take larger values in $|\alpha| \le \gamma$ than in the rest of α." However, $\phi(\alpha) + \phi(-\alpha)$ attains the

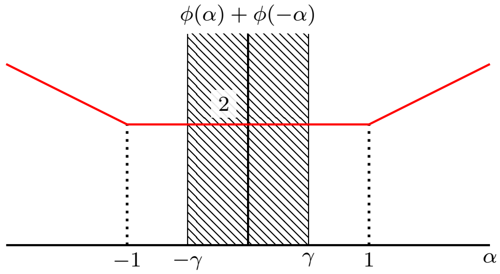


Figure 2: Illustration of $\phi(\alpha)+\phi(-\alpha) = 2\mathcal{C}_\phi\left(\alpha, \frac{1}{2}\right)$, where ϕ is the hinge loss and $\gamma = 0.5$. $\phi(\alpha)+\phi(-\alpha)$ has the same minimizers in both $|\alpha| \leq \gamma$ and $|\alpha| \leq 1$.

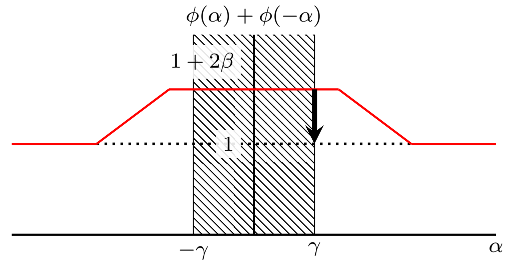


Figure 3: Illustration of $\phi(\alpha)+\phi(-\alpha) = 2\mathcal{C}_\phi\left(\alpha, \frac{1}{2}\right)$, where ϕ is the ramp loss with $\beta = 0.3$ and $\gamma = 0.5$. The condition $\phi(\gamma)+\phi(-\gamma) > \phi(1)+\phi(-1) = 1$ reflects the idea that predictions fall into the shaded area ($|\alpha| \leq \gamma$) must be penalized more than the others.

infimum at $\alpha = 0$ because $\phi(\alpha)+\phi(-\alpha)$ is convex and even as long as ϕ is convex. Therefore, the condition (9) would never be satisfied by convex surrogate ϕ. This idea is illustrated in Figure 2. ■

Hence, many popular surrogate losses such as the hinge, logistic, and squared losses are not calibrated wrt (ϕ_γ,$\mathcal{F}_{\text{all}}$). We defer all proofs to Appendix B.

Note that convex losses can be calibrated wrt (ϕ_γ,$\mathcal{F}_{\text{all}}$) under restricted distributions while we are primarily interested in calibrated losses under all distributions (see Definition 2). Indeed, $\mathcal{C}_\phi(\alpha, \eta)$ would not be minimized in $|\alpha| \leq \gamma$ unless η is close enough to $\frac{1}{2}$. In other words, convex losses may be calibrated wrt (ϕ_γ,$\mathcal{F}_{\text{all}}$) under low-noise conditions (Mammen and Tsybakov, 1999).

7. Calibration Conditions for Nonconvex Surrogates

As seen in Section 6, convex surrogate losses that are calibrated wrt (ϕ_γ,$\mathcal{F}_{\text{all}}$) do not exist. This motivates a search for nonconvex surrogate losses. Nonconvex surrogates are used for outlier robustness (Collobert et al., 2006; Masnadi-Shirazi and Vasconcelos, 2009; Holland, 2019) or label-noise robustness (Ghosh et al., 2015; van Rooyen et al., 2015; Charoenphakdee et al., 2019). Bounded monotone surrogates such as the ramp loss and the sigmoid loss are simple and common choices for those purposes. In this section, we also look for good surrogates from bounded monotone losses.

First, we introduce an important notion that constrains our search space of loss functions.

Definition 10 *We say a margin-based loss function $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is* quasiconcave even *if $\phi(\alpha) + \phi(-\alpha)$ is quasiconcave. Such ϕ is called a* quasiconcave even loss.

The name comes from the fact that any function $h(x)$ may be uniquely expressed as the sum of its even part $\frac{h(x)+h(-x)}{2}$ and odd part $\frac{h(x)-h(-x)}{2}$. This fact is also utilized to study the relationship between loss functions and sufficiency (Patrini et al., 2016).

Next, we state our main positive result. Its proof is included in Appendix B.

Theorem 11 *Let $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a surrogate loss. Assume that ϕ is bounded, nonincreasing, and quasiconcave even. Let $B \stackrel{\text{def}}{=} \phi(1) + \phi(-1)$ and assume $\phi(-1) > \phi(1)$. Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ be a function class such that $\mathcal{A}_\mathcal{F} \supseteq [-1, 1]$. Then,*

1. *ϕ is (ϕ_{01},$\mathcal{F}$)-calibrated.*

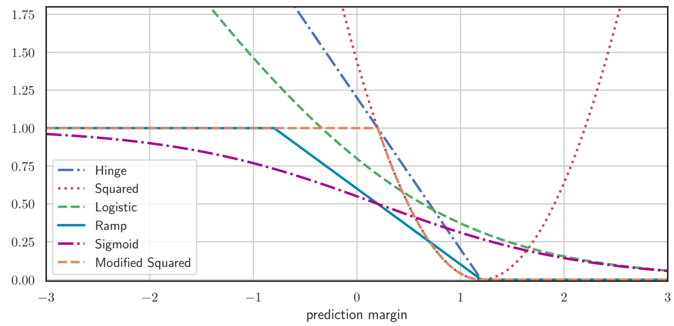


Figure 4: Surrogate losses. They are different from the traditional ones by horizontal translation of $+\beta$ ($\beta = 0.2$ here).

2. *ϕ is (ϕ_γ,$\mathcal{F}$)-calibrated if and only if $\phi(\gamma) + \phi(-\gamma) > B$.*

Proof *(Sketch of 2)* As in the proof sketch of Theorem 8, (9) is needed for (ϕ_γ,$\mathcal{F}$)-calibration, and $\phi(\alpha) + \phi(-\alpha)$ "should take larger values in $|\alpha| \leq \gamma$ than in the rest of α." Quasiconcavity of $\phi(\alpha) + \phi(-\alpha)$ naturally implies this property with a non-strict inequality, and the condition $\phi(\gamma) + \phi(-\gamma) > B$ ensures the strict inequality. Figure 3 illustrates it with the ramp loss. ■

To the best our knowledge, this is the first characterization of losses calibrated to ϕ_γ. This result is especially interesting when $\mathcal{F} = \mathcal{F}_{\text{lin}}$, ensuring that a quasiconcave even surrogate ϕ such that $\phi(\gamma) + \phi(-\gamma) > B$ is (ϕ_γ,$\mathcal{F}_{\text{lin}}$)-calibrated.

We remark that $\phi(\gamma) + \phi(-\gamma) \geq B$ always holds when ϕ is bounded, nonincreasing, and quasiconcave even (see part 4 of Lemma 13 in Appendix B). The strict inequality $\phi(\gamma)+\phi(-\gamma) > B$ is necessary and sufficient for (ϕ_γ,$\mathcal{F}_{\text{lin}}$)-calibration.

We additionally remark that the ramp loss and the sigmoid loss are (ϕ_{01},$\mathcal{F}_{\text{all}}$)-calibrated (Bartlett et al., 2006; Charoenphakdee et al., 2019). Note that these two losses are bounded, nonincreasing, and quasiconcave even, hence (ϕ_{01},$\mathcal{F}_{\text{lin}}$)-calibrated.

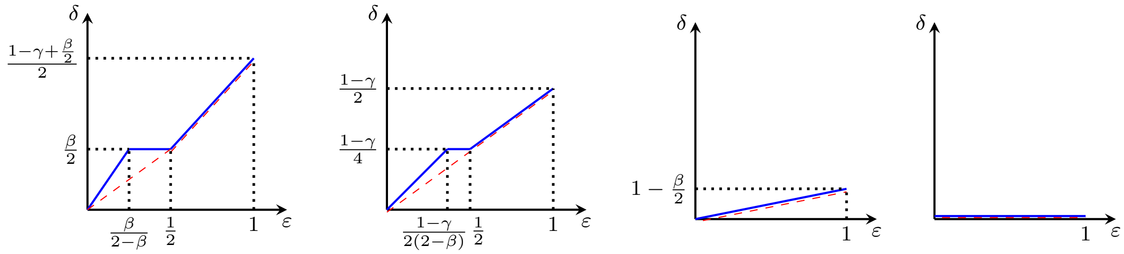
8. Examples

Several examples of loss functions are shown in Figure 4. For each base surrogate ϕ, we consider the shifted surrogate $\phi_\beta(\alpha) \overset{\text{def}}{=} \phi(\alpha - \beta)$ with the horizontal shift parameter β. The ramp, sigmoid, modified squared losses are examples of nonconvex and quasiconcave even losses when $\beta \geq 0$, while the hinge, logistic, and squared losses are examples of convex losses. We show (ϕ_γ,$\mathcal{F}_{\text{lin}}$)-calibration functions in this subsection.[2] As a result, we will see that the ramp, sigmoid, and modified squared losses are calibrated with appropriate shift parameters. Detailed derivations of the calibration functions and the proofs of quasiconcavity are deferred to Appendix C.

8.1. Ramp Loss

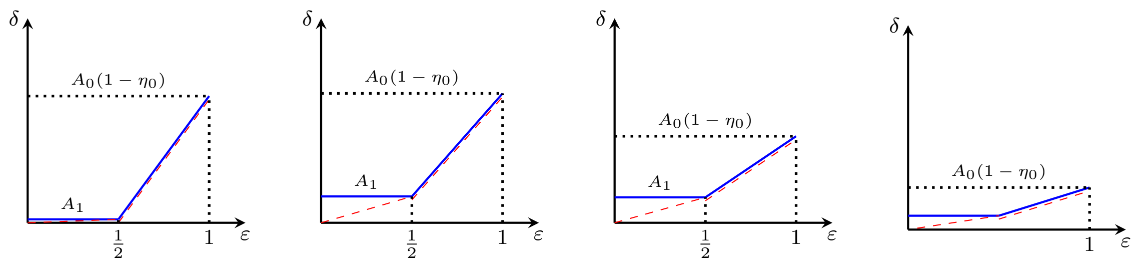
The ramp loss is $\phi(\alpha) = \min\left\{1, \max\left\{0, \frac{1-\alpha}{2}\right\}\right\}$. We consider the shifted ramp loss: $\phi_\beta(\alpha) = \phi(\alpha - \beta) = \min\left\{1, \max\left\{0, \frac{1-\alpha+\beta}{2}\right\}\right\}$. The ($\phi_\gamma$,$\mathcal{F}_{\text{lin}}$)-calibration function and its Fenchel-Legendre biconjugate of the ramp loss are plotted in Figure 5. We can see that the ramp loss

2. We only rely on the fact that $\mathcal{F}_{\text{lin}} \supseteq [-1, 1]$. The results can be extrapolated to $\mathcal{F}$ such that $\mathcal{F} \supseteq [-1, 1]$.



(a) $0 \le \beta < 1-\gamma$ (b) $1-\gamma \le \beta < 1+\gamma$ (c) $1+\gamma \le \beta < 2$ (d) $2 \le \beta$

Figure 5: The calibration function of the ramp loss. The dashed line is $\check{\delta}^{**}$.



(a) $\beta = 0$ (b) $\beta = 1.0$ (c) $\beta = 2.0$ (d) $\beta = 3.0$

Figure 6: The calibration function of the sigmoid loss. $A_0 \overset{\text{def}}{=} \phi_\beta(\gamma) - \phi_\beta(-\gamma) - \phi_\beta(1) + \phi_\beta(-1)$, $A_1 \overset{\text{def}}{=} (\phi_\beta(\gamma) + \phi_\beta(-\gamma) - \phi_\beta(1) - \phi_\beta(-1))/2$, and $\eta_0 \overset{\text{def}}{=} (\phi_\beta(-1) - \phi_\beta(-\gamma))/A_0$. The dashed line is $\check{\delta}^{**}$.

is calibrated wrt (ϕ_γ,$\mathcal{F}_{\text{lin}}$) when $0 < \beta < 2$. Since the ramp loss is quasi-concave even when $\beta \ge 0$, we also observe that the ramp loss is not calibrated when $\beta = 0$ because it is symmetric loss (Charoenphakdee et al., 2019), that is, $\phi_0(\alpha) + \phi_0(-\alpha) = 1$ for all $\alpha \in \mathbb{R}$, which does not satisfy the condition $\phi_0(\gamma) + \phi_0(-\gamma) > B = 1$ in Theorem 2.

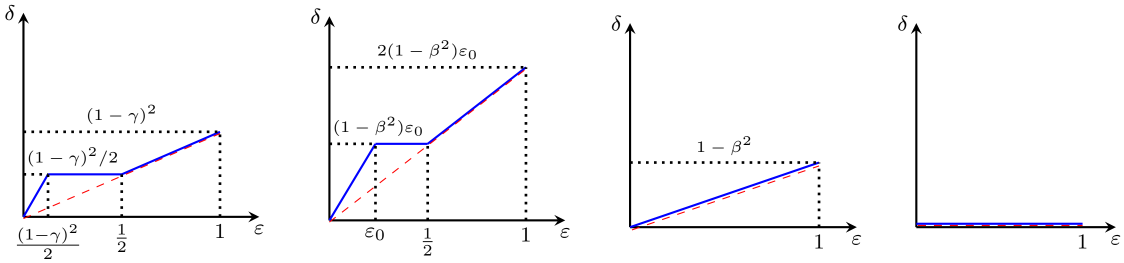
8.2. Sigmoid Loss

The sigmoid loss is $\phi(\alpha) = \frac{1}{1+e^\alpha}$. We consider the shifted sigmoid loss: $\phi_\beta(\alpha) = \frac{1}{1+e^{\alpha-\beta}}$ for $\beta > 0$. The (ϕ_γ,$\mathcal{F}_{\text{lin}}$)-calibration function is plotted in Figure 6. Thus, the sigmoid loss is (ϕ_γ,$\mathcal{F}_{\text{lin}}$)-calibrated when $A_1 > 0$, which is equivalent to $\beta > 0$. Again, we observe that the sigmoid loss with $\beta = 0$ is not calibrated in the same way as the ramp loss because it is symmetric.

8.3. Modified Squared Loss

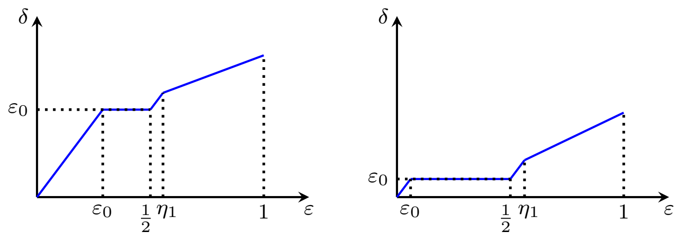
We make a bounded monotone surrogate $\phi(\alpha) = \text{clip}_{[0,1]}(\max\{0, 1-\alpha\}^2)$ by modifying the squared loss, where $\text{clip}_{[a,b]}(\cdot)$ clips values outside the interval $[a, b]$, and consider the shifted version $\phi_\beta(\alpha) \overset{\text{def}}{=} \phi(\alpha - \beta)$. The ($\phi_\gamma$,$\mathcal{F}_{\text{lin}}$)-calibration function and its Fenchel-Legendre biconjugate are plotted in Figure 7. We can deduce that the modified squared loss is calibrated wrt (ϕ_γ,$\mathcal{F}_{\text{lin}}$) for all $0 \le \beta < 1$. In contrast to the proceeding examples, the modified squared loss is not symmetric.

Moreover, the modified squared loss is (ϕ_γ,$\mathcal{F}_{\text{lin}}$)-calibrated even if ϕ_β for $\beta < 0$ is not a quasi-concave even loss. We plot two examples in Figure 8. As seen in the proof sketch of Theorem 11, it is crucial that $\phi_\beta(\alpha) + \phi_\beta(-\alpha)$ takes higher values in $|\alpha| \le \gamma$ than in $|\alpha| > \gamma$. The modified squared loss with $-1 + \frac{1}{\sqrt{2}} < \beta < 0$ satisfies this property (see Figure 9).



(a) $\beta = 0$ (b) $0 < \beta < \gamma$ (c) $\gamma \leq \beta < 1$ (d) $1 \leq \beta$

Figure 7: The calibration function of the modified squared loss. The dashed line is $\check{\delta}^{**}$. $\varepsilon_0 \overset{\text{def}}{=} \frac{(1-\gamma)(1-\gamma+2\beta)}{2(1-\beta^2)}$.



(a) $\beta = -0.1, \gamma = 0.2$ (b) $\beta = -0.2, \gamma = 0.2$

Figure 8: The calibration function of the modified squared loss when $\beta < 0$. $\varepsilon_0 \overset{\text{def}}{=} \beta^2 + 2\beta + \frac{1}{2}$ and $\eta_1 \overset{\text{def}}{=} (2 + 2\beta + \gamma)/4(1+\gamma)$.

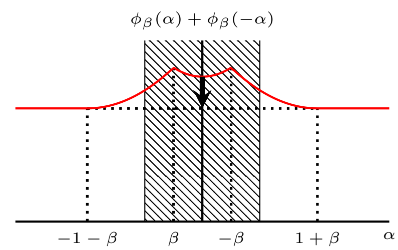


Figure 9: Illustration of $\phi_\beta(\alpha) + \phi_\beta(-\alpha)$ for the modified squared loss when $-1 + \frac{1}{\sqrt{2}} < \beta < 0$. Here, $\beta = -0.2$ and $\gamma = 0.4$.

8.4. Hinge Loss and Squared Loss

Here we consider the shifted hinge loss $\phi_\beta(\alpha) = \max\{0, 1 - \alpha + \beta\}$, and the shifted squared loss $\phi_\beta(\alpha) = (1 - \alpha + \beta)^2$ as examples of convex losses. Their $(\phi_\gamma,\mathcal{F}_{\text{lin}})$-calibration functions are plotted in Figures 10 and 11, respectively, which tell us that the hinge and squared losses are not $(\phi_\gamma,\mathcal{F}_{\text{lin}})$-calibrated. This result aligns with Theorem 8.

9. Simulation

Learning Curve on Synthetic Data. We generate positive and negative data from $\mathcal{N}([2\ 2]^\top, I_2)$ and $\mathcal{N}(-[2\ 2]^\top, I_2)$, respectively, and normalize with the maximum ℓ_2-norm among all data points. This ensures that data points lie in the ℓ_2 unit ball. We generate 800 training and 200 test points.

Linear models $f(x) = \theta^\top x + \theta_0$ are used, where θ and θ_0 are learnable parameters. As surrogate losses, we use the ramp, sigmoid, logistic, and hinge losses, with shift parameter $\beta = 0.2$. Batch gradient descent with the fixed step size 0.1 is used in optimization, and 1,000 steps are run for each trial. After every parameter update, the parameters are normalized to ensure $\|[\theta\ \theta_0]^\top\|_2 = 1$.

The robust 0-1 loss with $\gamma = 0.2$ is used as the target loss. To compute the excess risk, the Bayes risk for each surrogate loss and the robust 0-1 loss is numerically computed. The detail of numerical approximation of the Bayes risks is explained in Appendix D. The surrogate and target excess risks are shown in Figure 12. 20 trials are run for each data realization.

As you can see from Figure 12, optimization trajectories of calibrated surrogates (ramp and sigmoid) have target excess risks close to zero, while those of convex surrogates (logistic and hinge)

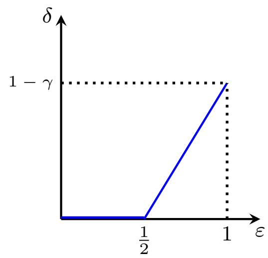
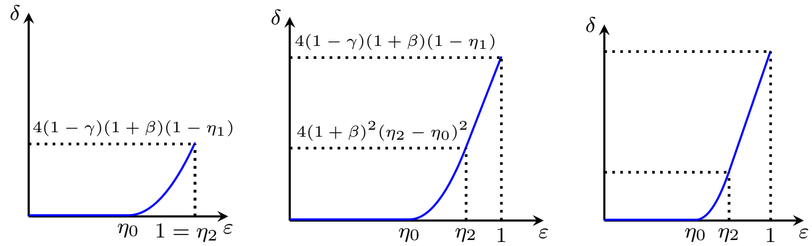


Figure 10: The calibration function of the hinge loss.



(a) $\beta = 0$ (b) $\beta = 0.5$ (c) $\beta = 1.0$

Figure 11: The calibration function of the squared loss. $\eta_0 \stackrel{\text{def}}{=} (1+\gamma+\beta)/2(1+\beta)$, $\eta_2 \stackrel{\text{def}}{=} (2+\beta)/2(1+\beta)$, and $\eta_1 \stackrel{\text{def}}{=} (\eta_0+\eta_2)/2$.

Table 1: The simulation results of the γ-adversarially robust 0-1 loss with $\gamma = 0.1$ and $\beta = 0.5$. 50 trials are conducted for each pair of a method and dataset. Standard errors (multiplied by 10^4) are shown in parentheses. Bold-faces indicate outperforming methods, chosen by one-sided t-test with the significant level 5%.

	Ramp	Sigmoid	Hinge	Logistic
0 vs 1	0.034 (3)	**0.017 (2)**	0.087 (12)	0.321 (19)
0 vs 2	**0.111 (7)**	0.133 (10)	**0.109 (8)**	0.281 (19)
0 vs 3	**0.107 (7)**	0.126 (8)	0.120 (9)	0.307 (18)
0 vs 4	**0.069 (6)**	0.093 (12)	0.072 (7)	0.269 (21)
0 vs 5	**0.233 (21)**	0.340 (25)	**0.233 (21)**	0.269 (16)
0 vs 6	**0.129 (8)**	0.167 (13)	**0.127 (8)**	0.287 (22)
0 vs 7	**0.067 (6)**	0.073 (6)	0.090 (9)	0.302 (18)
0 vs 8	**0.096 (7)**	0.123 (12)	0.100 (9)	0.263 (20)
0 vs 9	**0.082 (6)**	0.101 (8)	0.092 (8)	0.279 (22)

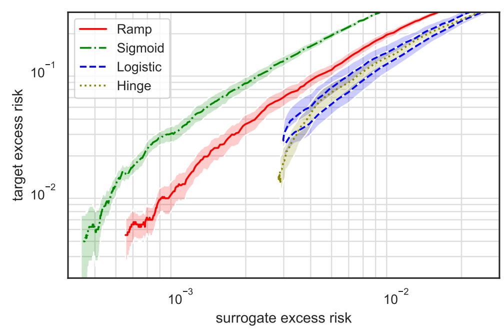


Figure 12: 20 trials of optimization trajectories are shown with standard errors. The horizontal (vertical, resp.) axis shows surrogate excess risk (excess risk of the robust 0-1 loss, resp.) on test data.

fail. This observation supports our theoretical findings in Theorems 8 and 11. Different values of β were tried for the hinge and logistic losses, but the conclusions are not affected.

Benchmark Data. We compare the ramp, sigmoid, hinge, and logistic losses on MNIST. The results are shown in Table 1, where we see that nonconvex losses, especially the ramp loss, outperform convex losses in terms of the robust 0-1 loss. Details and full results appear in Appendix D.

10. Conclusion

Calibration analysis was leveraged to analyze the adversarially robust 0-1 loss. We found that no convex surrogate loss is calibrated wrt the adversarially robust 0-1 loss. We also established necessary and sufficient conditions for a certain class of nonconvex surrogate losses to be calibrated wrt the adversarially robust 0-1 loss, which includes shifted versions of the ramp and sigmoid losses. An important open problem is to extend our calibration results to nonlinear classifier models.

Acknowledgments

HB was supported by JST ACT-I Grant Number JPMJPR18UI. CS was supported in part by NSF Grant Number 1838179. MS was supported by JST CREST Grant Number JPMJCR18A2.

References

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Didier Aussel, JN Corvellec, and Marc Lassonde. Subdifferential characterization of quasiconvexity and convexity. *Journal of Convex Analysis*, 1(2):195–201, 1994.

Bernardo Ávila Pires and Csaba Szepesvári. Multiclass classification calibration functions. *arXiv preprint arXiv:1609.06385*, 2016.

Bernardo Ávila Pires, Csaba Szepesvári, and Mohammad Ghavamzadeh. Cost-sensitive multiclass classification risk bounds. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1391–1399, 2013.

Han Bao and Masashi Sugiyama. Calibrated surrogate maximization of linear-fractional utility in binary classification. In *Proceedings of the 23th International Conference on Artificial Intelligence and Statistics*, 2020.

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.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*, volume 28. Princeton University Press, 2009.

Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

Mathieu Blondel. Structured prediction with projection oracles. In *Advances in Neural Information Processing Systems 32*, pages 12145–12156, 2019.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Sebastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In *Proceedings of the 36th International Conference on Machine Learning*, pages 831–840, 2019.

Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. On symmetric losses for learning from corrupted labels. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning*, pages 854–863, 2017.

Frank H Clarke. *Optimization and Nonsmooth Analysis*, volume 5. SIAM, 1990.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1310–1320, 2019.

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, 2006.

Zac Cranko, Aditya Menon, Richard Nock, Cheng Soon Ong, Zhan Shi, and Christian Walder. Monge blunts Bayes: Hardness results for adversarial training. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1406–1415, 2019.

Krzysztof Dembczynski, Wojciech Kotłowski, and Eyke Hüllermeier. Consistent multilabel ranking through univariate loss minimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1347–1354, 2012.

Ilias Diakonikolas, Daniel Kane, and Pasin Manurangsi. Nearly tight bounds for robust proper learning of halfspaces with a margin. In *Advances in Neural Information Processing Systems 32*, pages 10473–10484, 2019.

John C Duchi, Lester W Mackey, and Michael I Jordan. On the consistency of ranking algorithms. In *Proceedings of the 27th International Conference on Machine Learning*, pages 327–334, 2010.

Farzan Farnia and David Tse. A minimax approach to supervised learning. In *Advances in Neural Information Processing Systems 29*, pages 4240–4248, 2016.

Rizal Fathony, Anqi Liu, Kaiser Asif, and Brian Ziebart. Adversarial multiclass classification: A risk minimization perspective. In *Advances in Neural Information Processing Systems 29*, pages 559–567, 2016.

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

Wei Gao and Zhi-hua Zhou. On the consistency of multi-label learning. In *Proceedings of 24th Annual Conference on Learning*, 2011.

Wei Gao and Zhi-Hua Zhou. On the consistency of AUC pairwise optimization. In *Proceedings of 24th International Joint Conference on Artificial Intelligence*, 2015.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *Proceedings of the 30th International Conference on Machine Learning*, pages 738–746, 2013.

Aritra Ghosh, Naresh Manwani, and PS Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations Workshop*, 2015.

Tamir Hazan, Joseph Keshet, and David A McAllester. Direct loss minimization for structured prediction. In *Advances in Neural Information Processing Systems 23*, pages 1594–1602, 2010.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems 30*, pages 2266–2276, 2017.

Matthew Holland. Classification using margin pursuit. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 712–720, 2019.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *Proceedings of the 35th International Conference on Machine Learning*, pages 2034–2042, 2018.

Peter J Huber. *Robust Statistics*. Springer, 2011.

Justin Khim and Po-Ling Loh. Adversarial risk bounds via function transformation. In *Advances in Neural Information Processing Systems 32*, 2019.

Seiichi Kuroki, Nontawat Charoenphakdee, Han Bao, Junya Honda, Issei Sato, and Masashi Sugiyama. Unsupervised domain adaptation based on source-guided discrepancy. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, volume 33, pages 4122–4129, 2019.

Gert RG Lanckriet, Laurent El Ghaoui, Chiranjib Bhattacharyya, and Michael I Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3(Dec):555–582, 2002.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE, 2019.

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

Phil Long and Rocco Servedio. Consistency versus realizable H-consistency for multiclass classification. In *Proceedings of the 30th International Conference on Machine Learning*, pages 801–809, 2013.

Philip M Long and Rocco A Servedio. Random classification noise defeats all convex potential boosters. *Machine Learning*, 78(3):287–304, 2010.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Enno Mammen and Alexandre B Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of 22th Annual Conference on Learning*, 2009.

Hamed Masnadi-Shirazi and Nuno Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *Advances in Neural Information Processing Systems 22*, pages 1049–1056, 2009.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems 29*, pages 2208–2216, 2016.

Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems 30*, pages 2971–2980, 2017.

Anton Osokin, Francis Bach, and Simon Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Proceedings Systems 30*, pages 302–313, 2017.

Giorgio Patrini, Frank Nielsen, Richard Nock, and Marcello Carioni. Loss factorization, weakly supervised learning and label noise robustness. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 708–717, 2016.

Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cedric Gouy-Pailler, and Jamal Atif. Theoretical evidence for adversarial robustness through randomization. In *Advances in Neural Information Processing Systems 32*, pages 11838–11848, 2019.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *Proceedings of the 6th International Conference on Learning Representations*, 2018a.

Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems 31*, pages 10877–10887, 2018b.

Harish G Ramaswamy and Shivani Agarwal. Classification calibration dimension for general multiclass losses. In *Advances in Neural Information Proceedings Systems 25*, pages 2078–2086, 2012.

Harish G Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *The Journal of Machine Learning Research*, 17(1):397–441, 2016.

Harish G Ramaswamy, Shivani Agarwal, and Ambuj Tewari. Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. In *Advances in Neural Information Processing Systems 26*, pages 1475–1483, 2013.

Pradeep Ravikumar, Ambuj Tewari, and Eunho Yang. On NDCG consistency of listwise ranking methods. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 618–626, 2011.

Mark D Reid and Robert C Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11(Sep):2387–2422, 2010.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems 32*, pages 11289–11300, 2019.

Clayton Scott. Surrogate losses and regret bounds for cost-sensitive classification with example-dependent costs. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 153–160, 2011.

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

Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307:195–204, 2018.

Pannagadatta K Shivaswamy, Chiranjib Bhattacharyya, and Alexander J Smola. Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 7(Jul):1283–1314, 2006.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

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

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

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems 31*, pages 6541–6550, 2018.

Brendan van Rooyen, Aditya Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems 28*, pages 10–18, 2015.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5283–5292, 2018.

Huan Xu and Shie Mannor. Robustness and generalization. *Machine Learning*, 86(3):391–423, 2012.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(Jul):1485–1510, 2009.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7472–7482, 2019a.

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

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

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7404–7413, 2019b.

Appendix A. Convex and Quasiconvex Analysis

This section summarizes basic tools for convex and quasiconvex analysis.

Quasiconvex function: A function $h : S \to \mathbb{R}$ on a (finite-dimensional) vector space S is said to be *quasiconvex* if for all $x, y \in S$ and $\lambda \in [0,1]$, $h(\lambda x + (1-\lambda)y) \le \max\{h(x), h(y)\}$. A function h is said quasiconcave if $-h$ is quasiconvex: For all $x, y \in S$ and $\lambda \in [0,1]$, $h(\lambda x + (1-\lambda)y) \ge \min\{h(x), h(y)\}$. Intuitively, quasiconvexity relaxes convexity in that a function still preserves 'unimodality' though it loses definite curvature. There is an equivalent definition (here we only show for quasiconcavity): h is quasiconcave if every superlevel set $\{x \mid h(x) \ge t\}$ for $t \in \mathbb{R}$ is a convex set (Boyd and Vandenberghe, 2004).

Subderivative: In order to analyze convexity and quasiconvexity, subderivative is a useful tool. We adopt the Clarke definition of subderivative (Clarke, 1990; Aussel et al., 1994). Let S^* be the dual space of S and $\langle \cdot, \cdot \rangle$ be the dual pairing.[3] The (Clarke) subderivative of a lower semicontinuous function h is the operator $\partial h : S \to S^*$ defined for each $x \in S$ such that

$$\partial h(x) \overset{\text{def}}{=} \{x_* \in S^* \mid \langle x_*, x \rangle \le h^\circ(x; v) \quad \forall v \in S\},$$

where $h^\circ(x; v)$ is the Rockafellar directional derivative (see Clarke (1990) and Aussel et al. (1994)) for the formal definition). When h is locally Lipschitz at $x \in S$, Clarke (1990) states that this is equivalent to $\partial h(x) = \mathrm{co}\{\lim \nabla f(x_i) \mid x_i \to x, x_i \notin \Upsilon \cup \Omega_h\}$, where co is the convex hull, Υ is any set of measure zero, and Ω_h is the set of points where h is non-differentiable. In the case $S = \mathbb{R}$, this simply reduces to $\partial h(x) = [\partial_+ h(x), \partial_- h(x)]$, where $\partial_+ h$ and $\partial_- h$ are the right-/left-derivatives of h, respectively.

Properties of subderivative: Several basic properties of subderivatives are shown in Clarke (1990, Section 2.3) such as $\partial(th)(x) = t\partial h(x) \overset{\text{def}}{=} \{tx_* \mid x_* \in \partial h(x)\}$ (scalar multiples), $\partial\left(\sum h_i\right)(x) \subseteq \sum \partial h_i(x) \overset{\text{def}}{=} \{\sum x_{i,*} \mid x_{i,*} \in \partial h_i(x)\}$ (finite sums), and $0 \ni \partial h(x)$ if h attains a local extrema at x. When h is locally Lipschitz, it clearly holds that $\partial h(x) = \{h'(x)\}$ if h is differentiable at x.

Operator monotonicity: Convex smooth functions have monotonically nondecreasing derivatives. This can be extended to non-smooth functions via subderivatives. Let $h : S \to \mathbb{R}$ be a lower semicontinuous function. Then h is convex if and only if $\partial h : S \to S^*$ is a *monotone* operator (Aussel et al., 1994), that is, $\langle y_* - x_*, y - x \rangle \ge 0$ for all $x, y \in \mathbf{dom}(h)$ and $x_* \in \partial h(x), y_* \in \partial h(y)$. In addition, h is quasiconvex if and only if ∂h is a *quasimonotone* operator (Aussel et al., 1994), that is, $\langle x_*, y - x \rangle > 0 \Longrightarrow \langle y_*, y - x \rangle \ge 0$ for all $x, y \in \mathbf{dom}(h)$ and $x_* \in \partial h(x), y_* \in \partial h(y)$.

3. For two vector spaces U and V over the same field F and a bilinear map $\langle \cdot, \cdot, \rangle : U \times V \to F$, we say a triple $(U, V, \langle \cdot, \cdot, \rangle)$ is a dual pair if there exists $v \in V$ such that $\langle u, v \rangle \ne 0$ for all $u \in U$ and there exists $u \in U$ such that $\langle u, v \rangle \ne 0$ for all $v \in V$. Here, V is called a dual space of V, and $\langle \cdot, \cdot \rangle$ is called a dual pairing.

Appendix B. Deferred Proofs

B.1. Proof of Proposition 1

Proof Fix $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $f \in \mathcal{F}_{\text{lin}}$ associated with parameter $\theta \in \mathbb{R}^d$.

When $y = +1$, we divide into three cases depending on the value of $yf(x) = \theta^\top x$. If $\theta^\top x \le 0$, then $\Delta_x = 0$ simply gives $\theta^\top(x + \Delta_x) \le 0$. If $0 < \theta^\top x \le \gamma$, fix $\Delta_x = -\gamma\theta \in B_2^d(\gamma)$. Then, $\theta^\top(x + \Delta_x) = \theta^\top x - \gamma \le 0$. If $\theta^\top x > \gamma$, we observe $\theta^\top \Delta_x$ is minimized by $\Delta_x = -\frac{\gamma}{\|\theta\|_2}\theta \in B_2^d(\gamma)$. Then, $\theta^\top(x + \Delta_x) > \gamma + \theta^\top \Delta_x \ge \gamma - \gamma = 0$. In all cases, $\ell_\gamma(+1, x, f) = \mathbb{1}_{\{f(x) \le \gamma\}}$.

When $y = -1$, we divide the cases as well. If $\theta^\top x > 0$, then $\Delta_x = 0$ simply gives $\theta^\top(x + \Delta_x) > 0$. If $-\gamma \le \theta^\top x \le 0$, fix $\Delta_x = \gamma\theta \in B_2^d(\gamma)$. Then, $\theta^\top(x + \Delta_x) = \theta^\top x + \gamma \ge 0$. If $\theta^\top x < -\gamma$, we observe $\theta^\top \Delta_x$ is maximized by $\Delta_x = \frac{\gamma}{\|\theta\|_2}\theta \in B_2^d(\gamma)$. Then, $\theta^\top(x + \Delta_x) < -\gamma + \theta^\top \Delta_x \le -\gamma + \gamma = 0$. In all cases, $\ell_\gamma(-1, x, f) = \mathbb{1}_{\{f(x) \ge -\gamma\}} = \mathbb{1}_{\{-f(x) \le \gamma\}}$. ■

B.2. Proof of Lemma 7

Proof We first simplify the constraint in the calibration function (6). The ϕ_γ-CCR for $\alpha \in \mathcal{A}_\mathcal{F}$ is

$$\mathcal{C}_{\phi_\gamma}(\alpha, \eta) = \eta \mathbb{1}_{\{\alpha \le \gamma\}} + (1 - \eta)\mathbb{1}_{\{\alpha \ge -\gamma\}} = \begin{cases} 1 & \text{if } |\alpha| \le \gamma, \\ 1 - \eta & \text{if } \gamma < \alpha, \\ \eta & \text{if } \alpha < -\gamma, \end{cases}$$

and the minimal $(\phi_\gamma,\mathcal{F})$-CCR is $\mathcal{C}^*_{\phi_\gamma,\mathcal{F}}(\eta) = \min\{\eta, 1 - \eta\}$. If $\gamma < |\alpha|$, a well-known algebra in the binary classification case shows that $\mathcal{C}_{\phi_\gamma}(\alpha, \eta) - \mathcal{C}^*_{\phi_\gamma,\mathcal{F}}(\eta) = |2\eta - 1| \cdot \mathbb{1}_{\{(2\eta-1)\alpha \le 0\}}$ (see, e.g., Bartlett et al. (2006, Proof of Theorem 3)). If $|\alpha| \le \gamma$, it follows that $\mathcal{C}_{\phi_\gamma}(\alpha, \eta) - \mathcal{C}^*_{\phi_\gamma,\mathcal{F}}(\eta) = 1 - \min\{\eta, 1 - \eta\} = \max\{\eta, 1 - \eta\}$. Hence,

$$\Delta\mathcal{C}_{\phi_\gamma,\mathcal{F}}(\alpha, \eta) = \begin{cases} \max\{\eta, 1 - \eta\} & \text{if } |\alpha| \le \gamma, \\ |2\eta - 1| \cdot \mathbb{1}_{\{(2\eta-1)\alpha \le 0\}} & \text{if } \gamma < |\alpha|. \end{cases}$$

Next, we simplify the inner infimum on α, $\inf_{\alpha\in\mathbb{R}}\{\Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha, \eta) \mid \Delta\mathcal{C}_{\phi_\gamma,\mathcal{F}}(\alpha, \eta) \ge \varepsilon\} = \bar{\delta}(\varepsilon, \eta)$ in (6), for a fixed $\eta \in [0, 1]$. If $\varepsilon > \max\{\eta, 1-\eta\}$, no $\alpha \in \mathcal{A}_\mathcal{F}$ achieves $\Delta\mathcal{C}_{\phi_\gamma,\mathcal{F}}(\alpha, \eta) \ge \varepsilon$, meaning that $\bar{\delta}(\varepsilon, \eta) = \infty$. If $|2\eta - 1| < \varepsilon \le \max\{\eta, 1 - \eta\}$, $\Delta\mathcal{C}_{\phi_\gamma,\mathcal{F}}(\alpha, \eta) \ge \varepsilon$ is achieved when $|\alpha| \le \gamma$. Hence, $\bar{\delta}(\varepsilon, \eta) = \inf_\alpha\{\Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha, \eta) \mid |\alpha| \le \gamma\}$. Note that $|2\eta - 1| \le \max\{\eta, 1 - \eta\} = \frac{1+|2\eta-1|}{2}$ for all $\eta \in [0, 1]$. If $\varepsilon \le |2\eta - 1|$, $\Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha, \eta) \ge \varepsilon$ is achieved if either $|\alpha| \le \gamma$ or $(2\eta - 1)\alpha \le 0$ holds. Hence, $\bar{\delta}(\varepsilon, \eta) = \inf_\alpha\{\Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha, \eta) \mid |\alpha| \le \gamma \text{ or } (2\eta-1)\alpha \le 0\}$. These verify the statement of this lemma. ■

B.3. Useful Lemmas

The following lemmas are useful in the remaining proofs. Their proofs appear in Sections B.6 and B.7.

Lemma 12 *Let $\phi : \mathbb{R} \to \mathbb{R}_{\ge 0}$ be a margin-based loss function and $\mathcal{F} \subseteq \mathbb{R}^\mathcal{X}$ be a symmetric function class such that $\mathcal{A}_\mathcal{F} \supseteq [-1, 1]$.*

1. *For all* $\alpha \in \mathbb{R}$*,* $\mathcal{C}_\phi(\alpha,\eta)$ *and* $\Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha,\eta)$ *are symmetric about* $\eta = \frac{1}{2}$*, i.e.,* $\mathcal{C}_\phi(\alpha,\eta) = \mathcal{C}_\phi(-\alpha,1-\eta)$ *and* $\Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha,\eta) = \Delta\mathcal{C}_{\phi,\mathcal{F}}(-\alpha,1-\eta)$ *for all* $\eta \in [0,1]$*.*

2. *When* $\eta = \frac{1}{2}$*, we have*

$$\begin{aligned}\inf_{|\alpha|\le\gamma} \Delta\mathcal{C}_{\phi,\mathcal{F}}\left(\alpha,\frac{1}{2}\right) &= \inf_{0\le\alpha\le\gamma} \Delta\mathcal{C}_{\phi,\mathcal{F}}\left(\alpha,\frac{1}{2}\right) \\ &= \inf_{0\le\alpha\le\gamma} \mathcal{C}_\phi\left(\alpha,\frac{1}{2}\right) - \inf_{\alpha\in\mathcal{A}_\mathcal{F}:\alpha\ge 0} \mathcal{C}_\phi\left(\alpha,\frac{1}{2}\right).\end{aligned}$$

3. *A surrogate loss* ϕ *is calibrated wrt (*ϕ_γ*,*$\mathcal{F}$*) if and only if*

$$\inf_{|\alpha|\le\gamma} \mathcal{C}_\phi\left(\alpha,\frac{1}{2}\right) > \inf_{\alpha\in\mathcal{A}_\mathcal{F}} \mathcal{C}_\phi\left(\alpha,\frac{1}{2}\right), \text{ and}$$
$$\inf_{\alpha\in\mathcal{A}_\mathcal{F}:\alpha\le\gamma} \mathcal{C}_\phi(\alpha,\eta) > \inf_{\alpha\in\mathcal{A}_\mathcal{F}} \mathcal{C}_\phi(\alpha,\eta),$$

   for all $\eta \in \left(\frac{1}{2},1\right]$*.*

4. *A surrogate loss* ϕ *is calibrated wrt (*ϕ_{01}*,*$\mathcal{F}$*) if and only if*

$$\inf_{\alpha\in\mathcal{A}_\mathcal{F}:\alpha\le 0} \mathcal{C}_\phi(\alpha,\eta) > \inf_{\alpha\in\mathcal{A}_\mathcal{F}} \mathcal{C}_\phi(\alpha,\eta),$$

   for all $\eta \in \left(\frac{1}{2},1\right]$*.*

Note that part 4 of Lemma 12 can be regarded as a generalization of classification calibration (Bartlett et al., 2006, Definition 1) when $\mathcal{F} \neq \mathcal{F}_{\text{all}}$.

Lemma 13 *Let* $\phi : \mathbb{R} \to \mathbb{R}_{\ge 0}$ *be a margin-based loss function. If* ϕ *is nonincreasing, bounded,* $\phi \not\equiv 0$*, and quasiconcave even, then*

1. *the class-conditional* ϕ*-risk* $\mathcal{C}_\phi(\alpha,\eta)$ *is quasiconcave in* $\alpha \in \mathbb{R}$ *for all* $\eta \in [0,1]$*.*
2. *for all* $\eta \in \left(\frac{1}{2},1\right]$*,* $\mathcal{C}_\phi(\alpha,\eta)$ *is nonincreasing in* α *when* $\alpha \ge 0$*.*
3. *for all* $\eta \in \left(\frac{1}{2},1\right]$*,* $\mathcal{C}_\phi(-1,\eta) > \mathcal{C}_\phi(1,\eta)$*.*
4. $\phi(\alpha) + \phi(-\alpha)$ *is nonincreasing in* α *when* $\alpha \ge 0$*.*
5. *for* $l,u \in [-1,1]$ *(*$l \le u$*),* $\inf_{\alpha\in[l,u]} \mathcal{C}_\phi(\alpha,\eta) = \min\{\mathcal{C}_\phi(l,\eta), \mathcal{C}_\phi(u,\eta)\}$ *for all* $\eta \in [0,1]$*.*

B.4. Proof of Theorem 8

Proof Part 3 of Lemma 12 states that ϕ is calibrated wrt (ϕ_γ,$\mathcal{F}$) if and only if

$$\inf_{0\le\alpha\le\gamma} \mathcal{C}_\phi\left(\alpha,\frac{1}{2}\right) > \inf_{\alpha\in\mathcal{A}_\mathcal{F}:\alpha\ge 0} \mathcal{C}_\phi\left(\alpha,\frac{1}{2}\right) \quad \text{and}$$
$$\inf_{\alpha\in\mathcal{A}_\mathcal{F}:\alpha\le\gamma} \mathcal{C}_\phi(\alpha,\eta) > \inf_{\alpha\in\mathcal{A}_\mathcal{F}:\alpha\ge 0} \mathcal{C}_\phi(\alpha,\eta) \quad \text{for any } \eta \in \left(\frac{1}{2},1\right].$$

In order to show ϕ is not calibrated wrt (ϕ_γ,$\mathcal{F}$), it is sufficient to show that

$$\inf_{0\leq\alpha\leq\gamma} \mathcal{C}_\phi\left(\alpha, \frac{1}{2}\right) = \inf_{\alpha\in\mathcal{A}_\mathcal{F}:\alpha\geq 0} \mathcal{C}_\phi\left(\alpha, \frac{1}{2}\right),$$

which is equivalent to

$$\inf_{0\leq\alpha\leq\gamma} \phi(\alpha) + \phi(-\alpha) = \inf_{\alpha\in\mathcal{A}_\mathcal{F}:\alpha\geq 0} \phi(\alpha) + \phi(-\alpha).$$

Since $\bar{\phi}(\alpha) \overset{\text{def}}{=} \phi(\alpha) + \phi(-\alpha)$ is a convex even function, we have $\bar{\phi}(0) \leq \bar{\phi}(\alpha)$ for all $\alpha \in \mathcal{A}_\mathcal{F}$. Too see this, assume that there exists $\alpha_* \in \mathcal{A}_\mathcal{F}$ such that $\alpha_* \neq 0$ and $\bar{\phi}(0) > \bar{\phi}(\alpha_*)$. Then, we also have $\bar{\phi}(-\alpha_*) < \bar{\phi}(0)$ since $\bar{\phi}$ is even. It follows that $\frac{1}{2}\{\bar{\phi}(-\alpha_*) + \bar{\phi}(\alpha_*)\} < \bar{\phi}(0)$. However, we have $\frac{1}{2}\{\bar{\phi}(-\alpha_*) + \bar{\phi}(\alpha_*)\} \geq \bar{\phi}\left(\frac{-\alpha_*+\alpha_*}{2}\right) = \bar{\phi}(0)$ because of convexity of $\bar{\phi}$. Hence, we see $\bar{\phi}(0) \leq \bar{\phi}(\alpha)$ for all $\alpha \in \mathcal{A}_\mathcal{F}$. This means that $\inf_{0\leq\alpha\leq\gamma} \bar{\phi}(\alpha) = \inf_{\alpha\in\mathcal{A}_\mathcal{F}:0\leq\alpha} \bar{\phi}(\alpha) = \bar{\phi}(0)$. ■

B.5. Proof of Theorem 11

Proof of part 1 By part 4 of Lemma 12, (ϕ_{01},$\mathcal{F}$)-calibration is equivalent to

$$\inf_{-1\leq\alpha\leq 0} \mathcal{C}_\phi(\alpha,\eta) > \inf_{\alpha\in[-1,1]} \mathcal{C}_\phi(\alpha,\eta) \quad \text{for all } \eta \in \left(\frac{1}{2}, 1\right]. \tag{10}$$

First, we observe

$$\begin{aligned}
2\phi(0) = \phi(0) + \phi(0) &\geq \inf_{0\leq\alpha\leq 1} \phi(\alpha) + \phi(-\alpha) \\
&= \phi(1) + \phi(-1) && \text{(quasiconcavity of } \phi(\alpha) + \phi(-\alpha)) \\
&= B.
\end{aligned}$$

Next, fix η such that $\frac{1}{2} < \eta \leq 1$. We observe with part 5 of Lemma 13 that

$$\begin{aligned}
\inf_{-1\leq\alpha\leq 0} \mathcal{C}_\phi(\alpha,\eta) &= \min\{\mathcal{C}_\phi(-1,\eta), \mathcal{C}_\phi(0,\eta)\} && \text{(part 5 of Lemma 13)} \\
&= \min\{\eta\overline{B} + (1-\eta)\underline{B}, \phi(0)\},
\end{aligned}$$

and

$$\begin{aligned}
\inf_{-1\leq\alpha\leq 1} \mathcal{C}_\phi(\alpha,\eta) &= \min\{\mathcal{C}_\phi(-1,\eta), \mathcal{C}_\phi(1,\eta)\} && \text{(part 5 of Lemma 13)} \\
&= \mathcal{C}_\phi(1,\eta) && \text{(part 3 of Lemma 13)} \\
&= \eta\underline{B} + (1-\eta)\overline{B},
\end{aligned}$$

where $\underline{B} \overset{\text{def}}{=} \phi(1)$ and $\overline{B} = \phi(-1)$. Here,

$$\begin{aligned}
\mathcal{C}_\phi(-1,\eta) - \mathcal{C}_\phi(1,\eta) &= (\overline{B} - \underline{B})(2\eta - 1) > 0, \\
\mathcal{C}_\phi(0,\eta) - \mathcal{C}_\phi(1,\eta) &= \phi(0) - \overline{B} + \eta(\overline{B} - \underline{B}) \\
&\geq \frac{\overline{B} + \underline{B}}{2} - \overline{B} + \eta(\overline{B} - \underline{B}) && (2\phi(0) \geq B) \\
&> \frac{\overline{B} + \underline{B}}{2} - \overline{B} + \frac{\overline{B} - \underline{B}}{2} && (\phi(-1) > \phi(1) \text{ and } \eta > \tfrac{1}{2}) \\
&= 0.
\end{aligned}$$

Then, we have for all $\eta \in \left(\frac{1}{2}, 1\right]$,

$$\begin{aligned}\inf_{1\leq\alpha\leq 0} \mathcal{C}_\phi(\alpha,\eta) - \inf_{-1\leq\alpha\leq 1} \mathcal{C}_\phi(\alpha,\eta) &= \min\{\mathcal{C}_\phi(-1,\eta) - \mathcal{C}_\phi(1,\eta), \mathcal{C}_\phi(0,\eta) - \mathcal{C}_\phi(1,\eta)\} \\ &> 0.\end{aligned}$$

This verifies the condition (10). ■

Proof of part 2 ϕ is calibrated wrt (ϕ_γ,$\mathcal{F}$) if and only if

$$\begin{aligned} &\text{(i)} \quad \inf_{|\alpha|\leq\gamma} \mathcal{C}_\phi\left(\alpha, \frac{1}{2}\right) > \inf_{-1\leq\alpha\leq 1} \mathcal{C}_\phi\left(\alpha, \frac{1}{2}\right), \quad \text{and} \\ &\text{(ii)} \quad \inf_{-1\leq\alpha\leq\gamma} \mathcal{C}_\phi(\alpha,\eta) > \inf_{-1\leq\alpha\leq 1} \mathcal{C}_\phi(\alpha,\eta) \quad \text{for all } \eta \in \left(\frac{1}{2}, 1\right] \end{aligned} \tag{11}$$

by part 3 of Lemma 12. Now we show $\phi(\gamma) + \phi(-\gamma) > B$ assuming (i) and (ii).

$$\begin{aligned} \phi(\gamma) + \phi(-\gamma) &= \inf_{0\leq\alpha\leq\gamma} \phi(\alpha) + \phi(-\alpha) && \text{(part 4 of Lemma 13)} \\ &> \inf_{-1\leq\alpha\leq 1} \phi(\alpha) + \phi(-\alpha) && \text{((i) is used)} \\ &= \inf_{0\leq\alpha\leq 1} \phi(\alpha) + \phi(-\alpha) && (\phi(\alpha) + \phi(-\alpha) \text{ is even}) \\ &= \phi(1) + \phi(-1) && \text{(part 4 of Lemma 13)} \\ &= B. \end{aligned}$$

Conversely, assume $\phi(\gamma) + \phi(-\gamma) > B$. We will show (i) and (ii) in (11). Since $\phi(\alpha) + \phi(-\alpha)$ is nonincreasing in $\alpha \geq 0$ (part 4 of Lemma 13), we have

$$\begin{aligned} \inf_{|\alpha|\leq\gamma} \phi(\alpha) + \phi(-\alpha) &= \inf_{0\leq\alpha\leq\gamma} \phi(\alpha) + \phi(-\alpha) && (\phi(\alpha) + \phi(-\alpha) \text{ is even}) \\ &= \phi(\gamma) + \phi(-\gamma) && \text{(part 4 of Lemma 13)} \\ &> B \\ &= \phi(1) + \phi(-1) \\ &= \inf_{0\leq\alpha\leq 1} \phi(\alpha) + \phi(-\alpha), && \text{(part 4 of Lemma 13)} \\ &= \inf_{-1\leq\alpha\leq 1} \phi(\alpha) + \phi(-\alpha), && (\phi(\alpha) + \phi(-\alpha) \text{ is even}) \end{aligned}$$

which is equivalent to (i). For (ii), fix η such that $\frac{1}{2} < \eta \leq 1$. We first observe with parts 3 and 5 of Lemma 13 that

$$\begin{aligned} \inf_{-1\leq\alpha\leq\gamma} \mathcal{C}_\phi(\alpha,\eta) &= \min\left\{\mathcal{C}_\phi(-1,\eta), \mathcal{C}_\phi(\gamma,\eta)\right\}, \\ \inf_{-1\leq\alpha\leq 1} \mathcal{C}_\phi(\alpha,\eta) &= \min\left\{\mathcal{C}_\phi(-1,\eta), \mathcal{C}_\phi(1,\eta)\right\} = \mathcal{C}_\phi(1,\eta). \end{aligned}$$

Here, we have

$$\begin{aligned} \mathcal{C}_\phi(1,\eta) &= (\underline{B} - \overline{B})\eta + \overline{B}, \\ \mathcal{C}_\phi(\gamma,\eta) &= (\phi(\gamma) - \phi(-\gamma))\eta + \phi(-\gamma), \end{aligned}$$

where $\underline{B} \stackrel{\text{def}}{=} \phi(1)$ and $\overline{B} \stackrel{\text{def}}{=} \phi(-1)$. Observing that

$$\overline{B} - \underline{B} + \phi(\gamma) - \phi(-\gamma) \geq \overline{B} - \underline{B} + \phi(1) - \phi(-1) = 0, \qquad (\phi \text{ is nonincreasing})$$

we have for all $\eta \in \left(\frac{1}{2}, 1\right]$,

$$\begin{aligned}
\mathcal{C}_\phi(\gamma,\eta) - \mathcal{C}_\phi(1,\eta) &= (\phi(\gamma) - \phi(-\gamma) + \overline{B} - \underline{B})\eta + (\phi(-\gamma) - \overline{B}) \\
&\geq (\phi(\gamma) - \phi(-\gamma) + \overline{B} - \underline{B})\frac{1}{2} + \phi(-\gamma) - \overline{B} \\
&= \frac{\phi(\gamma) + \phi(-\gamma) - B}{2} \\
&> 0,
\end{aligned}$$

where the first inequality holds since $(\phi(\gamma) - \phi(-\gamma) + \overline{B} - \underline{B}) > 0$ and $\eta > \frac{1}{2}$, and the second inequality holds because of the assumption $\phi(\gamma) + \phi(-\gamma) > B$. In addition, we have $\mathcal{C}_\phi(-1,\eta) > \mathcal{C}_\phi(1,\eta)$ for $\eta > \frac{1}{2}$ by part 3 of Lemma 13. Therefore,

$$\begin{aligned}
\inf_{-1 \leq \alpha \leq \gamma} \mathcal{C}_\phi(\alpha,\eta) - \inf_{-1 \leq \alpha \leq 1} \mathcal{C}_\phi(\alpha,\eta) &= \min\{\mathcal{C}_\phi(-1,\eta) - \mathcal{C}_\phi(1,\eta), \mathcal{C}_\phi(\gamma,\eta) - \mathcal{C}_\phi(1,\eta)\} \\
&> 0
\end{aligned}$$

holds for all η such that $\frac{1}{2} < \eta \leq 1$, and this verifies (ii). ■

B.6. Proof of Lemma 12

Proof Parts 1 and 2 are obvious from the definition of the class-conditional ϕ-risk.
Part 3: Let $\delta : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be the $(\phi_\gamma,\mathcal{F})$-calibration function of ϕ, and $\bar{\delta} : \mathbb{R}_{\geq 0} \times [0,1] \to \mathbb{R}_{\geq 0}$ be the inner infimum of δ in (8):

$$\bar{\delta}(\varepsilon,\eta) = \begin{cases} \inf\limits_{|\alpha| \leq \gamma} \Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha,\eta) & \text{if } |2\eta - 1| < \varepsilon \leq \max\{\eta, 1-\eta\}, \\ \inf\limits_{\alpha \in \mathcal{A}_\mathcal{F} : |\alpha| \leq \gamma \text{ or } (2\eta-1)\alpha \leq 0} \Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha,\eta) & \text{if } \varepsilon \leq |2\eta - 1|, \end{cases}$$

and $\delta(\varepsilon) = \inf_{\eta \in [0,1]} \bar{\delta}(\varepsilon,\eta)$. Then, by Proposition 4, ϕ is $(\phi_\gamma,\mathcal{F})$-calibrated if and only if $\delta(\varepsilon) > 0$ for all $\varepsilon > 0$. If $\bar{\delta}(\varepsilon,\eta)$ is lower semicontinuous in η, this is equivalent to $\bar{\delta}(\varepsilon,\eta) > 0$ for all $\varepsilon > 0$ and $\eta \in [0,1]$. Using part 1 of Lemma 12 and symmetry of $\mathcal{F}$, since we have for $\eta \leq \frac{1}{2}$,

$$\begin{aligned}
\inf_{\alpha \in \mathcal{A}_\mathcal{F} : |\alpha| \leq \gamma \text{ or } (2\eta-1)\alpha \leq 0} \mathcal{C}_\phi(\alpha,\eta) &= \inf_{\alpha \in \mathcal{A}_\mathcal{F} : \alpha \geq -\gamma} \mathcal{C}_\phi(\alpha,\eta) \\
&= \inf_{\alpha \in \mathcal{A}_\mathcal{F} : \alpha \geq -\gamma} \mathcal{C}_\phi(-\alpha, 1-\eta) && \text{(part 1 of Lemma 12)} \\
&= \inf_{\alpha \in \mathcal{A}_\mathcal{F} : \alpha \leq \gamma} \mathcal{C}_\phi(\alpha, 1-\eta), && \text{(replace } -\alpha \text{ with } \alpha\text{)}
\end{aligned}$$

and for $\eta \geq \frac{1}{2}$,

$$\inf_{\alpha \in \mathcal{A}_\mathcal{F} : |\alpha| \leq \gamma \text{ or } (2\eta-1)\alpha \leq 0} \mathcal{C}_\phi(\alpha,\eta) = \inf_{\alpha \in \mathcal{A}_\mathcal{F} : \alpha \leq \gamma} \mathcal{C}_\phi(\alpha,\eta),$$

$\inf_{\alpha\in\mathcal{A}_{\mathcal{F}}:\alpha\leq\gamma} \Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha,\eta) > 0$ for all $\eta \geq \frac{1}{2}$ implies $\inf_{\alpha\in\mathcal{A}_{\mathcal{F}}:|\alpha|\leq\gamma \text{ or } (2\eta-1)\alpha\leq 0} \Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha,\eta) > 0$ for all $\eta \in [0,1]$. Hence,

$$\inf_{\alpha\in\mathcal{A}_{\mathcal{F}}:|\alpha|\leq\gamma \text{ or } (2\eta-1)\alpha\leq 0} \Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha,\eta) > 0$$

for $\varepsilon > 0$ and $\eta \in [0,1]$ such that $\varepsilon \leq |2\eta - 1|$ if and only if

$$\inf_{\alpha\in\mathcal{A}_{\mathcal{F}}:\alpha\leq\gamma} \Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha,\eta) > 0$$

for $\varepsilon > 0$ and $\eta \in [\frac{1}{2}, 1]$ such that $\varepsilon \leq 2\eta - 1$.

Therefore, $\bar{\delta}(\varepsilon,\eta) > 0$ for all $\varepsilon > 0$ and $\eta \in [0,1]$ if and only if

$$\begin{cases} \inf\limits_{|\alpha|\leq\gamma} \mathcal{C}_\phi(\alpha,\eta) > \inf\limits_{\alpha\in\mathcal{A}_{\mathcal{F}}} \mathcal{C}_\phi(\alpha,\eta) & \text{for all } \eta \geq \frac{1}{2} \text{ such that } 2\eta - 1 < \varepsilon \leq \eta, \\ \inf\limits_{\alpha\in\mathcal{A}_{\mathcal{F}}:\alpha\leq\gamma} \mathcal{C}_\phi(\alpha,\eta) > \inf\limits_{\alpha\in\mathcal{A}_{\mathcal{F}}} \mathcal{C}_\phi(\alpha,\eta) & \text{for all } \eta \geq \frac{1}{2} \text{ such that } \varepsilon \leq 2\eta - 1, \end{cases}$$

for all $\varepsilon > 0$, which is equivalent to

$$\begin{cases} \inf\limits_{|\alpha|\leq\gamma} \mathcal{C}_\phi(\alpha,\eta) > \inf\limits_{\alpha\in\mathcal{A}_{\mathcal{F}}} \mathcal{C}_\phi(\alpha,\eta) & \text{for all } \eta \geq \frac{1}{2} \text{ such that } \varepsilon \leq \eta < \frac{1+\varepsilon}{2}, \\ \inf\limits_{\alpha\in\mathcal{A}_{\mathcal{F}}:\alpha\leq\gamma} \mathcal{C}_\phi(\alpha,\eta) > \inf\limits_{\alpha\in\mathcal{A}_{\mathcal{F}}} \mathcal{C}_\phi(\alpha,\eta) & \text{for all } \eta \geq \frac{1}{2} \text{ such that } \frac{1+\varepsilon}{2} \leq \eta \leq 1, \end{cases}$$

for all $\varepsilon > 0$.

We immediately observe that

$$\left\{\eta \geq \frac{1}{2} \,\middle|\, \varepsilon \leq \eta < \frac{1+\varepsilon}{2}, \varepsilon > 0\right\} = \left\{\frac{1}{2} \leq \eta \leq 1\right\}, \quad \text{and}$$
$$\left\{\eta \geq \frac{1}{2} \,\middle|\, \frac{1+\varepsilon}{2} \leq \eta \leq 1, \varepsilon > 0\right\} = \left\{\frac{1}{2} < \eta \leq 1\right\}.$$

Therefore, we reduce the above conditions as

$$\begin{cases} \inf\limits_{|\alpha|\leq\gamma} \mathcal{C}_\phi(\alpha,\eta) > \inf\limits_{\alpha\in\mathcal{A}_{\mathcal{F}}} \mathcal{C}_\phi(\alpha,\eta) & \text{if } \frac{1}{2} \leq \eta \leq 1, \\ \inf\limits_{\alpha\in\mathcal{A}_{\mathcal{F}}:\alpha\leq\gamma} \mathcal{C}_\phi(\alpha,\eta) > \inf\limits_{\alpha\in\mathcal{A}_{\mathcal{F}}} \mathcal{C}_\phi(\alpha,\eta) & \text{if } \frac{1}{2} < \eta \leq 1. \end{cases}$$

Note that $\inf_{|\alpha|\leq\gamma} \mathcal{C}_\phi(\alpha,\eta) \geq \inf_{\alpha\in\mathcal{A}_{\mathcal{F}}:\alpha\leq\gamma} \mathcal{C}_\phi(\alpha,\eta)$ for all η. Since the first case is included in the second case except when $\eta = \frac{1}{2}$, this is equivalent to

$$\inf_{|\alpha|\leq\gamma} \mathcal{C}_\phi\left(\alpha, \tfrac{1}{2}\right) > \inf_{\alpha\in\mathcal{A}_{\mathcal{F}}} \mathcal{C}_\phi\left(\alpha, \tfrac{1}{2}\right), \text{ and } \inf_{\alpha\in\mathcal{A}_{\mathcal{F}}:\alpha\leq\gamma} \mathcal{C}_\phi(\alpha,\eta) > \inf_{\alpha\in\mathcal{A}_{\mathcal{F}}} \mathcal{C}_\phi(\alpha,\eta) \text{ for } \eta \in \left(\tfrac{1}{2}, 1\right].$$

Finally, we check lower semicontinuity of $\bar{\delta}(\varepsilon,\eta)$ in η. Fix a fixed α, $\mathcal{C}_\phi(\alpha,\eta)$ is lower semicontinuous in η since $\mathcal{C}_\phi(\alpha,\eta)$ is linear in η. Because pointwise infimum preserves lower semicontinuity, $\inf_{\alpha\in\mathcal{A}_{\mathcal{F}}} \mathcal{C}_\phi(\alpha,\eta)$, $\inf_{|\alpha|\leq\gamma} \mathcal{C}_\phi(\alpha,\eta)$, and $\inf_{\alpha\in\mathcal{A}_{\mathcal{F}}:|\alpha|\leq\gamma \text{ or } (2\eta-1)\alpha\leq 0} \mathcal{C}_\phi(\alpha,\eta)$ are lower semicontinuous in η. Hence, $\bar{\delta}(\varepsilon,\eta)$ is lower semicontinuous in η. This concludes the proof of part 3.

Part 4: We follow the same direction as part 3. If we take $\gamma \to 0$,

$$\bar{\delta}(\varepsilon, \eta) = \inf_{\alpha \in \mathcal{A}_{\mathcal{F}}:(2\eta-1)\alpha \le 0} \Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha, \eta) \quad \text{such that } \varepsilon \le |2\eta - 1|.$$

Hence, by Proposition 4 and lower semicontinuity of $\bar{\delta}(\varepsilon, \eta)$ in η (proven in part 3), ϕ is (ϕ_{01},$\mathcal{F}$)-calibrated if and only if

$$\inf_{\alpha \in \mathcal{A}_{\mathcal{F}}:(2\eta-1)\alpha \le 0} \Delta\mathcal{C}_{\phi,\mathcal{F}}(\alpha, \eta) > 0$$

for all $\varepsilon > 0$ and $\eta \in [0, 1]$ such that $\varepsilon \le |2\eta - 1|$. In the same way as part 3 of Lemma 12, this is equivalent to

$$\inf_{\alpha \in \mathcal{A}_{\mathcal{F}}:\alpha \le 0} \mathcal{C}_{\phi}(\alpha, \eta) > \inf_{\alpha \in \mathcal{A}_{\mathcal{F}}} \mathcal{C}_{\phi}(\alpha, \eta) \quad \text{for all } \eta \ge \frac{1}{2} \text{ such that } \frac{1+\varepsilon}{2} \le \eta \le 1,$$

for all $\varepsilon > 0$, by using part 1 of Lemma 12 and symmetry of $\mathcal{F}$. In the same way as part 3 of Lemma 12, simple observations on ranges ε and η reduce the above joint conditions on ε and η to η alone:

$$\inf_{\alpha \in \mathcal{A}_{\mathcal{F}}:\alpha \le 0} \mathcal{C}_{\phi}(\alpha, \eta) > \inf_{\alpha \in \mathcal{A}_{\mathcal{F}}} \mathcal{C}_{\phi}(\alpha, \eta) \quad \text{for all } \eta \text{ such that } \frac{1}{2} < \eta \le 1.$$

This is the lemma statement. ■

B.7. Proof of Lemma 13

Denote $\bar{\phi}(\alpha) \stackrel{\text{def}}{=} \phi(\alpha) + \phi(-\alpha)$. $\bar{\phi}$ is quasiconcave and even. To prove part 1, we use the following lemmas.

Lemma 14 *A function $h : \mathbb{R} \to \mathbb{R}$ is quasiconcave if and only if $\min\{x_{1,*}(x_1 - x_2), x_{2,*}(x_2 - x_1)\} \le 0$ for all $x_1, x_2 \in \mathbf{dom}(h)$, and $x_{1,*} \in \partial h(x_1)$ and $x_{2,*} \in \partial h(x_2)$.*

Proof If h is quasimonotone, Theorem 4.1 in Aussel et al. (1994) implies that $-\partial h$ is a quasimonotone operator, i.e.,

$$x_{1,*}(x_2 - x_1) < 0 \Longrightarrow x_{2,*}(x_2 - x_1) \le 0$$
$$\text{for all } x_1, x_2 \in \mathbf{dom}(h) \text{ and } x_{1,*} \in \partial h(x_1) \text{ and } x_{2,*} \in \partial h(x_2).$$

This is clearly equivalent to $\min\{x_{1,*}(x_1 - x_2), x_{2,*}(x_2 - x_1)\} \le 0$. ■

Lemma 15 *Any element in $\partial\bar{\phi}(\alpha_0)$ can be represented by $\alpha_*^+ - \alpha_*^-$ for some $a_*^+ \in \partial\phi(\alpha_0)$ and $a_*^- \in \partial\phi(-\alpha_0)$. For any $\eta \in \left[\frac{1}{2}, 1\right]$, $\alpha_*^+ \in \partial\phi(\alpha_0)$, and $\alpha_*^- \in \partial\phi(-\alpha_0)$, if $\alpha_*^+ - \alpha_*^- \in \partial\bar{\phi}(\alpha_0)$, then $\eta\alpha_*^+ - (1-\eta)\alpha_*^- \in \partial\mathcal{C}_{\phi}(\alpha_0, \eta)$.*

Proof By calculus of subderivative, we have $\partial\bar{\phi}(\alpha_0) \subseteq \partial\phi(\alpha_0) - \partial\phi(-\alpha_0)$. The first statement follows from this fact. In order to prove the second statement, note that left-/right-derivatives of ϕ exists because ϕ is nonincreasing. We first observe that

$$\begin{aligned}
&\text{(i)} \quad \partial_-\phi(\alpha_0) \le \alpha_*^+ \le \partial_+\phi(\alpha_0),\\
&\text{(ii)} \quad \partial_-\phi(-\alpha_0) \le \alpha_*^- \le \partial_+\phi(-\alpha_0), \text{ and}\\
&\text{(iii)} \quad \partial_-\phi(\alpha_0) - \partial_-\phi(-\alpha_0) \le \alpha_*^+ - \alpha_*^- \le \partial_+\phi(\alpha_0) - \partial_+\phi(-\alpha_0).
\end{aligned}$$

We have (iii) because

$$\begin{aligned}
\alpha_*^+ - \alpha_*^- \in \partial\bar{\phi}(\alpha_0) &= [\partial_-\bar{\phi}(\alpha_0), \partial_+\bar{\phi}(\alpha_0)]\\
&= [\partial_-\phi(\alpha_0) - \partial_-\phi(-\alpha_0), \partial_+\phi(\alpha_0) - \partial_+\phi(-\alpha_0)].
\end{aligned}$$

Then, for $\eta \in \left[\frac{1}{2}, 1\right]$, $(\eta - \frac{1}{2}) \times$ (i) $+ (\eta - \frac{1}{2}) \times$ (ii) $+ \frac{1}{2} \times$ (iii) gives

$$\eta\partial_-\phi(\alpha_0) - (1-\eta)\partial_-\phi(-\alpha_0) \le \eta\alpha_*^+ - (1-\eta)\alpha_*^- \le \eta\partial_+\phi(\alpha_0) - (1-\eta)\partial_+\phi(-\alpha_0),$$

which is equivalent to

$$\begin{aligned}
\eta\alpha_*^+ - (1-\eta)\alpha_*^- &\in [\eta\partial_-\phi(\alpha_0) - (1-\eta)\partial_-\phi(-\alpha_0), \eta\partial_+\phi(\alpha_0) - (1-\eta)\partial_+\phi(-\alpha_0)]\\
&= [\partial_-\mathcal{C}_\phi(\alpha_0, \eta), \partial_+\mathcal{C}_\phi(\alpha_0, \eta)]\\
&= \partial\mathcal{C}_\phi(\alpha_0, \eta).
\end{aligned}$$

■

Now we proceed with the proof of Lemma 13.

Proof *(of Lemma 13)*

Part 1: Fix $\eta \in \left[\frac{1}{2}, 1\right]$. Since $\bar{\phi}$ is quasiconcave, by Lemma 14, $\min\{\alpha_{1,*}(\alpha_1 - \alpha_2), \alpha_{2,*}(\alpha_2 - \alpha_1)\} \le 0$ for any $\alpha_1, \alpha_2 \in [-1, 1]$ and $\alpha_{1,*} \in \partial\bar{\phi}(\alpha_1)$ and $\alpha_{2,*} \in \partial\bar{\phi}(\alpha_2)$. Let us fix $\alpha_1, \alpha_2 \in \mathbb{R}$ such that $\alpha_1 \ge \alpha_2$, which can be assumed without loss of generality. Since $\partial\mathcal{C}_\phi(\alpha, \eta) \subseteq \eta\partial\phi(\alpha) - (1-\eta)\partial\phi(-\alpha)$ (the subdifferentiation is taken on α) and $\partial\bar{\phi}(\alpha) \subseteq \partial\phi(\alpha) - \partial\phi(-\alpha)$, we can pick $\alpha_{1,*}^+ \in \partial\phi(\alpha_1)$ and $\alpha_{1,*}^- \in \partial\phi(-\alpha_1)$ such that $\alpha_{1,*}^+ - \alpha_{1,*}^- \in \partial\bar{\phi}(\alpha_1)$ and $\eta\alpha_{1,*}^+ - (1-\eta)\alpha_{1,*}^- \in \partial\mathcal{C}_\phi(\alpha_1, \eta)$ by Lemma 15; in the same way, we can pick $\alpha_{2,*}^+ \in \partial\phi(\alpha_2)$ and $\alpha_{2,*}^- \in \partial\phi(-\alpha_2)$ such that $\alpha_{2,*}^+ - \alpha_{2,*}^- \in \partial\bar{\phi}(\alpha_2)$ and $\eta\alpha_{2,*}^+ - (1-\eta)\alpha_{2,*}^- \in \partial\mathcal{C}_\phi(\alpha_2, \eta)$.

Here, let us divide $\min\{(\alpha_{1,*}^+ - \alpha_{1,*}^-)(\alpha_1 - \alpha_2), (\alpha_{2,*}^+ - \alpha_{2,*}^-)(\alpha_2 - \alpha_1)\} \le 0$, the necessary condition of quasiconcavity of $\bar{\phi}$, into two cases. Consider the case $(\alpha_{1,*}^+ - \alpha_{1,*}^-)(\alpha_1 - \alpha_2) \le 0$ first. In this case,

$$\begin{aligned}
\{\eta\alpha_{1,*}^+ - (1-\eta)\alpha_{1,*}^-\}(\alpha_1 - \alpha_2) &= \eta\alpha_{1,*}^+(\alpha_1 - \alpha_2) - (1-\eta)\alpha_{1,*}^-(\alpha_1 - \alpha_2)\\
&\le \eta\alpha_{1,*}^+(\alpha_1 - \alpha_2) - (1-\eta)\alpha_{1,*}^+(\alpha_1 - \alpha_2)\\
&= \underbrace{(2\eta - 1)(\alpha_1 - \alpha_2)}_{\ge 0}\alpha_{1,*}^+\\
&\le 0,
\end{aligned}$$

where the last inequality holds because ϕ is nonincreasing (hence $\alpha_{1,*}^+ \le 0$). Note again that $\eta\alpha_{1,*}^+ - (1-\eta)\alpha_{1,*}^- \in \partial\mathcal{C}_\phi(\alpha_1, \eta)$.

In another case $(\alpha_{2,*}^+ - \alpha_{2,*}^-)(\alpha_2 - \alpha_1) \le 0$, we can show $\{\eta\alpha_{2,*}^+ - (1-\eta)\alpha_{2,*}^-\}(\alpha_2 - \alpha_1) \le 0$ in the same way, and note that $\eta\alpha_{2,*}^+ - (1-\eta)\alpha_{2,*}^- \in \partial\mathcal{C}_\phi(\alpha_2, \eta)$. Since we take α_1 and α_2 arbitrarily, now we have $\min\{\alpha_{1,*}^\eta(\alpha_1 - \alpha_2), \alpha_{2,*}^\eta(\alpha_2 - \alpha_1)\} \le 0$ for all $\alpha_{1,*}^\eta \in \partial\mathcal{C}_\phi(\alpha_1, \eta)$ and $\alpha_{2,*}^\eta \in \partial\mathcal{C}_\phi(\alpha_2, \eta)$. This is the sufficient condition of quasiconcavity of $\mathcal{C}_\phi(\alpha, \eta)$ in $\alpha \in \mathbb{R}$ by Lemma 14. Therefore, we confirm quasiconcavity of $\mathcal{C}_\phi(\alpha, \eta)$ in α when $\eta \ge \frac{1}{2}$ given quasiconcavity of $\bar{\phi}$.

Finally, if $\eta \in \left[0, \frac{1}{2}\right)$, we know by part 1 of Lemma 12 that $\mathcal{C}_\phi(\alpha, \eta) = \mathcal{C}_\phi(-\alpha, 1-\eta)$ for $\alpha \in \mathbb{R}$. Then, quasiconcavity of $\mathcal{C}_\phi(\alpha, \eta)$ in $\alpha \in \mathbb{R}$ follows since $\mathcal{C}_\phi(-\alpha, 1-\eta)$ is quasiconcave in $-\alpha \in \mathbb{R}$.

Part 2: Fix a $\eta \in \left(\frac{1}{2}, 1\right]$ and $\alpha_1, \alpha_2 \ge 0$ such that $\alpha_1 < \alpha_2$. By the fact that ϕ is nonincreasing, we have

$$\begin{aligned}\phi(\alpha_1) - \phi(-\alpha_1) - \phi(\alpha_2) + \phi(-\alpha_2) &= (\phi(\alpha_1) - \phi(\alpha_2)) + (\phi(-\alpha_2) - \phi(-\alpha_1)) \\ &\ge 0.\end{aligned}$$

Then,

$$\begin{aligned}\mathcal{C}_\phi(\alpha_1, \eta) - \mathcal{C}_\phi(\alpha_2, \eta) &= (\phi(\alpha_1) - \phi(-\alpha_1) - \phi(\alpha_2) + \phi(-\alpha_2))\eta + \phi(-\alpha_1) - \phi(-\alpha_2) \\ &\ge (\phi(\alpha_1) - \phi(-\alpha_1) - \phi(\alpha_2) + \phi(-\alpha_2))\frac{1}{2} + \phi(-\alpha_1) - \phi(-\alpha_2) \\ &= \frac{\phi(\alpha_1) + \phi(-\alpha_1) - \phi(\alpha_2) - \phi(-\alpha_2)}{2} \\ &\ge 0,\end{aligned}$$

where the last inequality holds because $\phi(\alpha) + \phi(-\alpha)$ is nonincreasing when $\alpha \ge 0$ by part 4. Therefore, $\mathcal{C}_\phi(\alpha, \eta)$ is nonincreasing in $\alpha \ge 0$.

Part 3: Fix a $\eta \in (\frac{1}{2}, 1]$. Then,

$$\begin{aligned}\mathcal{C}_\phi(-1, \eta) - \mathcal{C}_\phi(1, \eta) &= (2\eta - 1)(\phi(-1) - \phi(1)) \\ &> 0,\end{aligned}$$

where the inequality holds due to $\eta > \frac{1}{2}$ and $\phi \not\equiv 0$ and ϕ is non-increasing.

Part 4: $\bar{\phi}$ is an even function, so it is symmetric in $\alpha = 0$. Since ϕ is quasiconcave even, i.e., $\bar{\phi}$ is quasiconcave. Every quasiconcave function is nondecreasing, or nonincreasing, or there is global maxima in its domain (Boyd and Vandenberghe, 2004). If $\bar{\phi}$ is either nondecreasing or nonincreasing in $\alpha \in [-1, 1]$, it is a constant function in $\alpha \in [-1, 1]$ and clearly nonincreasing in $\alpha \ge 0$. If $\bar{\phi}$ has global maxima, i.e., there is a point $\alpha_* \in [-1, 1]$ such that $\bar{\phi}$ is nondecreasing for $\alpha \le \alpha_*$ and nonincreasing for $\alpha \ge \alpha_*$, it is still nonincreasing in $\alpha \ge 0$. This is clear when $\alpha_* \le 0$. When $\alpha_* > 0$, $\bar{\phi}$ may only be a constant function in $\alpha \in [0, \alpha_*]$ otherwise we have a point $\widetilde{\alpha} \in [0, \alpha_*)$ such that $\bar{\phi}(\widetilde{\alpha}) < \bar{\phi}(\alpha_*)$; hence $\bar{\phi}(\alpha_*) = \bar{\phi}(-\alpha_*)$ (write this value as $\bar{\phi}_*$) by the symmetry and $\bar{\phi}_0 \stackrel{\text{def}}{=} \bar{\phi}(\widetilde{\alpha}) < \bar{\phi}_*$, which means there is no convex superlevel sets for $\bar{\phi}$ within the range $(\bar{\phi}_0, \bar{\phi}_*)$. For example, pick $t \in (\bar{\phi}_0, \bar{\phi}_*)$ and consider t-superlevel set of $\bar{\phi}$. If t-superlevel set is convex, it must contain every point in $[-\alpha_*, \alpha_*]$ since $t < \bar{\phi}_* = \bar{\phi}(-\alpha_*) = \bar{\phi}(\alpha_*)$. However, t-superlevel set would not contain $\widetilde{\alpha} \in [-\alpha_*, \alpha_*]$ since $t > \bar{\phi}_0 = \bar{\phi}(\widetilde{\alpha})$. This contradicts with quasiconcavity of $\bar{\phi}$. In either case, $\bar{\phi}$ is nonincreasing in $\alpha \ge 0$.

Part 5: Fix $\eta \in [0, 1]$. This is an immediate consequence of quasiconcavity of $\mathcal{C}_\phi(\alpha, \eta)$ (part 1). Boyd and Vandenberghe (2004) states that there are three cases for a quasiconcave function. If

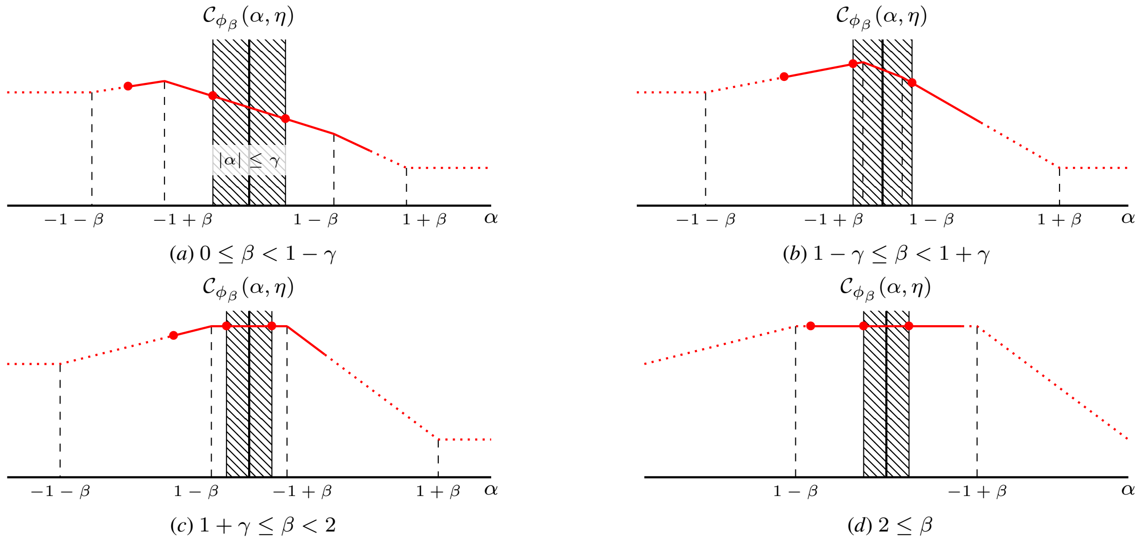


Figure 13: The class-conditional risk for the ramp loss.

$\mathcal{C}_\phi(\alpha, \eta)$ is nondecreasing or nonincreasing, the statement is clear. If there is a point $\alpha_* \in [l, u]$ such that $\mathcal{C}_\phi(\alpha, \eta)$ is nondecreasing for $\alpha \le \alpha_*$ and nonincreasing for $\alpha \ge \alpha_*$, the statement is clear again. In all cases, we have $\inf_{\alpha \in [l,u]} \mathcal{C}_\phi(\alpha, \eta) = \min\{\mathcal{C}_\phi(l, \eta), \mathcal{C}_\phi(u, \eta)\}$. ■

Appendix C. Derivation of Calibration Functions

C.1. Ramp Loss

The ramp loss is $\phi(\alpha) = \min\left\{1, \max\left\{0, \frac{1-\alpha}{2}\right\}\right\}$. We consider the shifted ramp loss: $\phi_\beta(\alpha) = \phi(\alpha - \beta)$:

$$\phi_\beta(\alpha) = \begin{cases} 1 & \text{if } \alpha \le -1+\beta, \\ \frac{1-\alpha+\beta}{2} & \text{if } -1+\beta < \alpha \le 1+\beta, \\ 0 & \text{if } 1+\beta < \alpha. \end{cases}$$

C.1.1. CALIBRATION FUNCTION

We analyze ϕ_β-CCR $\mathcal{C}_{\phi_\beta}(\alpha, \eta) = \eta\phi_\beta(\alpha) + (1-\eta)\phi_\beta(-\alpha)$, and restrict $\eta > \frac{1}{2}$ by virtue of the symmetry of $\mathcal{C}_{\phi_\beta}$ (part 1 in Lemma 12). $\mathcal{C}_{\phi_\beta}(\alpha, \eta)$ is plotted in Figure 13. By part 5 of Lemma 13, it is easy to see

$$\mathcal{C}^*_{\phi_\beta, \mathcal{F}_{\text{lin}}}(\eta) = \min\{\mathcal{C}_{\phi_\beta}(-1, \eta), \mathcal{C}_{\phi_\beta}(1, \eta)\} = \mathcal{C}_{\phi_\beta}(1, \eta).$$

Subsequently, we divide into cases depending on the relationship among $\mathcal{C}_{\phi_\beta}(-1, \eta)$, $\mathcal{C}_{\phi_\beta}(\gamma, \eta)$, and $\mathcal{C}_{\phi_\beta}(-\gamma, \eta)$.

(A) When $0 \le \beta < 1-\gamma$:

$$\begin{aligned}
\mathcal{C}_{\phi_\beta}(1,\eta) &= \frac{\beta}{2}\eta + (1-\eta),\\
\mathcal{C}_{\phi_\beta}(-1,\eta) &= \eta + \frac{\beta}{2}(1-\eta),\\
\mathcal{C}_{\phi_\beta}(\gamma,\eta) &= \frac{1-\gamma+\beta}{2}\eta + \frac{1+\gamma+\beta}{2}(1-\eta),\\
\mathcal{C}_{\phi_\beta}(-\gamma,\eta) &= \frac{1+\gamma+\beta}{2}\eta + \frac{1-\gamma+\beta}{2}(1-\eta),
\end{aligned}$$

from which it follows that $\mathcal{C}_{\phi_\beta}(-\gamma,\eta) - \mathcal{C}_{\phi_\beta}(\gamma,\eta) = \frac{\gamma}{2}(2\eta-1) > 0$, that is, $\mathcal{C}_{\phi_\beta}(-\gamma,\eta) > \mathcal{C}_{\phi_\beta}(\gamma,\eta)$ for all $\eta > \frac{1}{2}$. In addition, since

$$\mathcal{C}_{\phi_\beta}(\gamma,\eta) - \mathcal{C}_{\phi_\beta}(-1,\eta) = -\left(1+\gamma-\frac{\beta}{2}\right)(\eta-\eta_0) \quad \text{where} \quad \eta_0 \overset{\text{def}}{=} \frac{1+\gamma}{2\left(1+\gamma-\frac{\beta}{2}\right)},$$

we have $\mathcal{C}_{\phi_\beta}(\gamma,\eta) > \mathcal{C}_{\phi_\beta}(-1,\eta)$ if $\eta < \eta_0$ and $\mathcal{C}_{\phi_\beta}(\gamma,\eta) \le \mathcal{C}_{\phi_\beta}(-1,\eta)$ if $\eta \ge \eta_0$.

- If $\frac{1}{2} < \eta < \eta_0$: By part 5 in Lemma 13, it follows that

$$\inf_{|\alpha|\le\gamma} \mathcal{C}_{\phi_\beta}(\alpha,\eta) = \mathcal{C}_{\phi_\beta}(\gamma,\eta) \quad \text{and} \quad \inf_{|\alpha|\le 1 : |\alpha|\le\gamma \text{ or } (2\eta-1)\alpha\le 0} \mathcal{C}_{\phi_\beta}(\alpha,\eta) = \mathcal{C}_{\phi_\beta}(-1,\eta).$$

  Thus, by Lemma 7,

$$\bar{\delta}(\varepsilon,\eta) = \begin{cases} \infty & \text{if } \eta < \varepsilon,\\ \mathcal{C}_{\phi_\beta}(\gamma,\eta) - \mathcal{C}^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta) = \left(1-\gamma-\frac{\beta}{2}\right)\left(\eta-\frac{1}{2}\right) + \frac{\beta}{2} & \text{if } \varepsilon \le \eta < \frac{1+\varepsilon}{2},\\ \mathcal{C}_{\phi_\beta}(-1,\eta) - \mathcal{C}^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta) = (2-\beta)\left(\eta-\frac{1}{2}\right) & \text{if } \frac{1+\varepsilon}{2} \le \eta. \end{cases}$$

  Hence we obtain

$$\inf_{\eta\in\left(\frac{1}{2},\eta_0\right]} \bar{\delta}(\varepsilon,\eta) = \begin{cases} \left(1-\frac{\beta}{2}\right)\varepsilon & \text{if } 0 < \varepsilon \le \frac{\beta}{2-\beta},\\ \frac{\beta}{2} & \text{if } \frac{\beta}{2-\beta} < \varepsilon \le \frac{1}{2},\\ \left(1-\gamma-\frac{\beta}{2}\right)\left(\varepsilon-\frac{1}{2}\right) + \frac{\beta}{2} & \text{if } \frac{1}{2} < \varepsilon. \end{cases}$$

  Note that $0 \le \frac{\beta}{2-\beta} \le \frac{1-\gamma}{2} < \frac{1}{2}$.

- If $\eta_0 \le \eta \le 1$: By part 5 in Lemma 13, it follows that

$$\inf_{|\alpha|\le\gamma} \mathcal{C}_{\phi_\beta}(\alpha,\eta) = \inf_{|\alpha|\le 1 : |\alpha|\le\gamma \text{ or } (2\eta-1)\alpha\le 0} \mathcal{C}_{\phi_\beta}(\alpha,\eta) = \mathcal{C}_{\phi_\beta}(\gamma,\eta).$$

  Thus, by Lemma 7,

$$\bar{\delta}(\varepsilon,\eta) = \begin{cases} \infty & \text{if } \eta < \varepsilon,\\ \mathcal{C}_{\phi_\beta}(\gamma,\eta) - \mathcal{C}^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta) = \left(1-\gamma-\frac{\beta}{2}\right)\left(\eta-\frac{1}{2}\right) + \frac{\beta}{2} & \text{if } \varepsilon \le \eta. \end{cases}$$

  Hence we obtain

$$\inf_{\eta\in(\eta_0,1]} \bar{\delta}(\varepsilon,\eta) = \begin{cases} \frac{\beta}{2} & \text{if } 0 < \varepsilon \le \frac{1}{2},\\ \left(1-\gamma-\frac{\beta}{2}\right)\left(\varepsilon-\frac{1}{2}\right) + \frac{\beta}{2} & \text{if } \frac{1}{2} < \varepsilon. \end{cases}$$

Combining the above, we obtain the ϕ_γ-calibration function from Lemma 7:

$$\delta(\varepsilon) = \begin{cases} \left(1 - \frac{\beta}{2}\right)\varepsilon & \text{if } 0 < \varepsilon \le \frac{\beta}{2-\beta}, \\ \frac{\beta}{2} & \text{if } \frac{\beta}{2-\beta} < \varepsilon \le \frac{1}{2}, \\ \left(1 - \gamma - \frac{\beta}{2}\right)\left(\varepsilon - \frac{1}{2}\right) + \frac{\beta}{2} & \text{if } \frac{1}{2} < \varepsilon. \end{cases}$$

(B) When $1 - \gamma \le \beta < 1 + \gamma$**:**

$$\begin{aligned} \mathcal{C}_{\phi_\beta}(1, \eta) &= \frac{\beta}{2}\eta + (1 - \eta), \\ \mathcal{C}_{\phi_\beta}(-1, \eta) &= \eta + \frac{\beta}{2}(1 - \eta), \\ \mathcal{C}_{\phi_\beta}(\gamma, \eta) &= \frac{1 - \gamma + \beta}{2}\eta + (1 - \eta), \\ \mathcal{C}_{\phi_\beta}(-\gamma, \eta) &= \eta + \frac{1 - \gamma + \beta}{2}(1 - \eta), \end{aligned}$$

from which it follows that $\mathcal{C}_{\phi_\beta}(-\gamma, \eta) - \mathcal{C}_{\phi_\beta}(\gamma, \eta) = \frac{1+\gamma-\beta}{2}(2\eta - 1) > 0$, that is, $\mathcal{C}_{\phi_\beta}(-\gamma, \eta) > \mathcal{C}_{\phi_\beta}(\gamma, \eta)$ for all $\eta > \frac{1}{2}$. In addition, since

$$\mathcal{C}_{\phi_\beta}(\gamma, \eta) - \mathcal{C}_{\phi_\beta}(-1, \eta) = -\frac{3 + \gamma - 2\beta}{2}(\eta - \eta_0), \quad \left(\eta_0 \overset{\text{def}}{=} \frac{2}{3 + \gamma - 2\beta}\right)$$

we have $\mathcal{C}_{\phi_\beta}(\gamma, \eta) > \mathcal{C}_{\phi_\beta}(-1, \eta)$ if $\eta < \eta_0$ and $\mathcal{C}_{\phi_\beta}(\gamma, \eta) \le \mathcal{C}_{\phi_\beta}(-1, \eta)$ if $\eta \ge \eta_0$.

- If $\frac{1}{2} < \eta < \eta_0$: By part 5 in Lemma 13, it follows that

  $$\inf_{|\alpha| \le \gamma} \mathcal{C}_{\phi_\beta}(\alpha, \eta) = \mathcal{C}_{\phi_\beta}(\gamma, \eta) \quad \text{and} \quad \inf_{|\alpha| \le 1 : |\alpha| \le \gamma \text{ or } (2\eta - 1)\alpha \le 0} \mathcal{C}_{\phi_\beta}(\alpha, \eta) = \mathcal{C}_{\phi_\beta}(-1, \eta).$$

  Thus, by Lemma 7,

  $$\bar{\delta}(\varepsilon, \eta) = \begin{cases} \infty & \text{if } \eta < \varepsilon, \\ \mathcal{C}_{\phi_\beta}(\gamma, \eta) - \mathcal{C}^*_{\phi_\beta, \mathcal{F}_{\text{lin}}}(\eta) = \frac{1-\gamma}{2}\eta & \text{if } \varepsilon \le \eta < \frac{1+\varepsilon}{2}, \\ \mathcal{C}_{\phi_\beta}(-1, \eta) - \mathcal{C}^*_{\phi_\beta, \mathcal{F}_{\text{lin}}}(\eta) = (2 - \beta)\left(\eta - \frac{1}{2}\right) & \text{if } \frac{1+\varepsilon}{2} \le \eta. \end{cases}$$

  Hence we obtain

  $$\inf_{\eta \in \left(\frac{1}{2}, \eta_0\right]} \bar{\delta}(\varepsilon, \eta) = \begin{cases} \left(1 - \frac{\beta}{2}\right)\varepsilon & \text{if } 0 < \varepsilon \le \frac{1-\gamma}{2(2-\beta)}, \\ \frac{1-\gamma}{4} & \text{if } \frac{1-\gamma}{2(2-\beta)} < \varepsilon \le \frac{1}{2}, \\ \frac{1-\gamma}{2}\varepsilon & \text{if } \frac{1}{2} < \varepsilon, \end{cases}$$

- If $\eta_0 \le \eta \le \frac{1}{2}$: By part 5 in Lemma 13, it follows that

  $$\inf_{|\alpha| \le \gamma} \mathcal{C}_{\phi_\beta}(\alpha, \eta) = \inf_{|\alpha| \le 1 : |\alpha| \le \gamma \text{ or } (2\eta - 1)\alpha \le 0} \mathcal{C}_{\phi_\beta}(\alpha, \eta) = \mathcal{C}_{\phi_\beta}(\gamma, \eta).$$

Thus, by Lemma 7,

$$\bar{\delta}(\varepsilon,\eta)=\begin{cases}\infty & \text{if } \eta<\varepsilon,\\ \mathcal{C}_{\phi_\beta}(\gamma,\eta)-\mathcal{C}^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta)=\frac{1-\gamma}{2}\eta & \text{if } \varepsilon\le\eta.\end{cases}$$

Hence we obtain

$$\inf_{\eta\in(\eta_0,1]}\bar{\delta}(\varepsilon,\eta)=\begin{cases}\varepsilon_0' & \text{if } 0<\varepsilon\le\varepsilon_0',\\ \frac{1-\gamma}{2}\varepsilon & \text{if } \varepsilon_0'<\varepsilon,\end{cases}$$

where $\varepsilon_0'\overset{\text{def}}{=}\frac{1-\gamma}{2}\eta_0(\ge\frac{1-\gamma}{4})$.

Combining the above, we obtain the ϕ_γ-calibration function from Lemma 7:

$$\delta(\varepsilon)=\begin{cases}\left(1-\frac{\beta}{2}\right)\varepsilon & \text{if } 0<\varepsilon\le\frac{1-\gamma}{2(2-\beta)},\\ \frac{1-\gamma}{4} & \text{if } \frac{1-\gamma}{2(2-\beta)}<\varepsilon\le\frac{1}{2},\\ \frac{1-\gamma}{2}\varepsilon & \text{if } \frac{1}{2}<\varepsilon.\end{cases}$$

Note that $\frac{1-\gamma}{2(2-\beta)}\le\frac{1-\gamma}{2(1+\gamma)}<\frac{1}{2}$ when $1-\gamma\le\beta<1+\gamma$. This means the second case would not degenerate.

(C) When $1+\gamma\le\beta<2$: It is easy to see

$$\inf_{|\alpha|\le\gamma}\mathcal{C}_{\phi_\beta}(\alpha,\eta)=1,$$

$$\inf_{|\alpha|\le1:|\alpha|\le\gamma\text{ or }(2\eta-1)\alpha\le0}\mathcal{C}_{\phi_\beta}(\alpha,\eta)=\mathcal{C}_{\phi_\beta}(-1,\eta)=\eta+\frac{\beta}{2}(1-\eta),$$

$$\mathcal{C}^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta)=\mathcal{C}_{\phi_\beta}(1,\eta)=\frac{\beta}{2}\eta+(1-\eta).$$

Hence, by part 5 in Lemma 13, it follows that

$$\bar{\delta}(\varepsilon,\eta)=\begin{cases}\infty & \text{if } \eta<\varepsilon,\\ 1-\mathcal{C}_{\phi_\beta}(1,\eta)=\left(1-\frac{\beta}{2}\right)\eta & \text{if } \varepsilon\le\eta<\frac{1+\varepsilon}{2},\\ \mathcal{C}_{\phi_\beta}(-1,\eta)-\mathcal{C}_{\phi_\beta}(1,\eta)=(2-\beta)\left(\eta-\frac{1}{2}\right) & \text{if } \frac{1+\varepsilon}{2}\le\eta.\end{cases}$$

Thus, by Lemma 7, $\delta(\varepsilon)=\inf_{\eta\in\left(\frac{1}{2},1\right]}\bar{\delta}(\varepsilon,\eta)=\left(1-\frac{\beta}{2}\right)\varepsilon$.

(D) When $2\le\beta$: In this case, $\mathcal{C}_{\phi_\beta}(\alpha,\eta)=1$ for all $\eta\in[0,1]$ and $\alpha\in[-1,1]$. Hence, $\Delta\mathcal{C}_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\alpha,\eta)=0$ and $\delta(\varepsilon)=0$.

To sum up, the (ϕ_γ,$\mathcal{F}_{\text{lin}}$)-calibration function and its Fenchel-Legendre biconjugate of the ramp loss is as follows:

- If $0\le\beta<1-\gamma$,

$$\delta(\varepsilon)=\begin{cases}\left(1-\frac{\beta}{2}\right)\varepsilon & \text{if } 0<\varepsilon\le\frac{\beta}{2-\beta},\\ \frac{\beta}{2} & \text{if } \frac{\beta}{2-\beta}<\varepsilon\le\frac{1}{2},\\ \left(1-\gamma-\frac{\beta}{2}\right)\left(\varepsilon-\frac{1}{2}\right)+\frac{\beta}{2} & \text{if } \frac{1}{2}<\varepsilon,\end{cases}$$

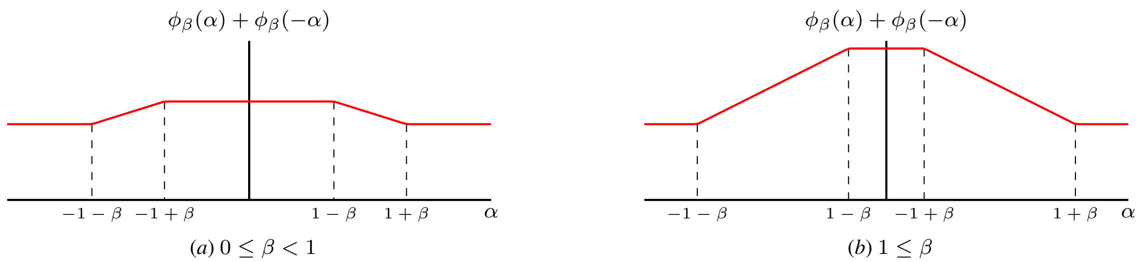


Figure 14: The even part of the ramp loss.

and

$$\delta^{**}(\varepsilon) = \begin{cases} \beta\varepsilon & \text{if } 0 < \varepsilon \leq \frac{1}{2}, \\ \left(1-\gamma-\frac{\beta}{2}\right)\left(\varepsilon-\frac{1}{2}\right)+\frac{\beta}{2} & \text{if } \frac{1}{2} < \varepsilon. \end{cases}$$

- If $1-\gamma \leq \beta < 1+\gamma$,

$$\delta(\varepsilon) = \begin{cases} \left(1-\frac{\beta}{2}\right)\varepsilon & \text{if } 0 < \varepsilon \leq \frac{1-\gamma}{2(2-\beta)}, \\ \frac{1-\gamma}{4} & \text{if } \frac{1-\gamma}{2(2-\beta)} < \varepsilon \leq \frac{1}{2}, \\ \frac{1-\gamma}{2}\varepsilon & \text{if } \frac{1}{2} < \varepsilon, \end{cases} \quad \text{and} \quad \delta^{**}(\varepsilon) = \left(1-\frac{\gamma}{2}\right)\varepsilon.$$

- If $1+\gamma \leq \beta < 2$, $\delta(\varepsilon) = \delta^{**}(\varepsilon) = \left(1-\frac{\beta}{2}\right)\varepsilon$.

- If $2 \leq \beta$, $\delta(\varepsilon) = \delta^{**}(\varepsilon) = 0$.

We can see that the ramp loss is calibrated wrt (ϕ_γ,$\mathcal{F}_{\text{lin}}$) when $0 < \beta < 2$.

C.1.2. QUASICONCAVITY OF EVEN PART

We confirm that $\phi_\beta(\alpha)+\phi_\beta(-\alpha)$ is quasiconcave when $\beta \geq 0$. In each case, $\phi_\beta(\alpha)+\phi_\beta(-\alpha)$ is plotted in Figure 14.
(A) When $0 \leq \beta < 1$:

$$\phi_\beta(\alpha)+\phi_\beta(-\alpha) = \begin{cases} 1 & \alpha \leq -1-\beta, \\ \frac{3+\alpha+\beta}{2} & -1-\beta \leq \alpha < -1+\beta, \\ 1+\beta & -1+\beta \leq \alpha < 1-\beta, \\ \frac{3-\alpha+\beta}{2} & 1-\beta \leq \alpha < 1+\beta, \\ 1 & 1+\beta \leq \alpha. \end{cases}$$

The t-superlevel set of $\phi_\beta(\alpha)+\phi_\beta(-\alpha)$ (denote S_t) is as follows.

- If $t < 1$, $S_t = \mathbb{R}$.
- If $1 \leq t \leq 1+\beta$, $S_t = \{\alpha \mid |\alpha| \leq 3+\beta-2t\}$.
- If $1+\beta < t$, $S_t = \emptyset$.

In all cases, S_t is convex.
(B) When $1 \leq \beta$:

$$\phi_\beta(\alpha)+\phi_\beta(-\alpha) = \begin{cases} 1 & \alpha \leq -1-\beta, \\ \frac{3+\alpha+\beta}{2} & -1-\beta \leq \alpha < 1-\beta, \\ 2 & 1-\beta \leq \alpha < -1+\beta, \\ \frac{3-\alpha+\beta}{2} & -1+\beta \leq \alpha < 1+\beta, \\ 1 & 1+\beta \leq \alpha. \end{cases}$$

The t-superlevel set of $\phi_\beta(\alpha)+\phi_\beta(-\alpha)$ (denote S_t) is as follows.

- If $t < 1$, $S_t = \mathbb{R}$.
- If $1 \leq t \leq 2$, $S_t = \{\alpha \mid |\alpha| \leq 3+\beta-2t\}$.
- If $2 < t$, $S_t = \emptyset$.

In all cases, S_t is convex.

C.2. Sigmoid Loss

The sigmoid loss is $\phi(\alpha) = \frac{1}{1+e^{\alpha}}$. We consider the shifted sigmoid loss: $\phi_\beta(\alpha) = \frac{1}{1+e^{\alpha-\beta}}$ for $\beta \geq 0$. ϕ_β-CCR is

$$\mathcal{C}_{\phi_\beta}(\alpha,\eta) = \frac{\eta}{1+e^{\alpha-\beta}} + \frac{1-\eta}{1+e^{-\alpha-\beta}}.$$

$\mathcal{C}_{\phi_\beta}$ is plotted in Figure 15.

C.2.1. CALIBRATION FUNCTION

We focus on the case $\eta > \frac{1}{2}$ due to the symmetry of $\mathcal{C}_{\phi_\beta}$. By part 5 of Lemma 13, it is easy to check

$$\mathcal{C}^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta) = \min\{\mathcal{C}_{\phi_\beta}(-1,\eta), \mathcal{C}_{\phi_\beta}(1,\eta)\} = \mathcal{C}_{\phi_\beta}(1,\eta) = \frac{\eta}{1+e^{1-\beta}} + \frac{1-\eta}{1+e^{-1-\beta}}.$$

Since

$$\begin{aligned} \mathcal{C}_{\phi_\beta}(-\gamma,\eta) - \mathcal{C}_{\phi_\beta}(\gamma,\eta) &= \left(\frac{\eta}{1+e^{-\gamma-\beta}} + \frac{1-\eta}{1+e^{\gamma-\beta}}\right) - \left(\frac{\eta}{1+e^{\gamma-\beta}} + \frac{1-\eta}{1+e^{-\gamma-\beta}}\right) \\ &= (2\eta-1)\left(\frac{1}{1+e^{-\gamma-\beta}} - \frac{1}{1+e^{\gamma-\beta}}\right) \\ &> 0, \qquad (\text{since } -\gamma-\beta < \gamma-\beta) \end{aligned}$$

we have $\mathcal{C}_{\phi_\beta}(\gamma,\eta) < \mathcal{C}_{\phi_\beta}(-\gamma,\eta)$ for all $\eta > \frac{1}{2}$. On the other hand, since

$$\begin{aligned} &\mathcal{C}_{\phi_\beta}(-1,\eta) - \mathcal{C}_{\phi_\beta}(\gamma,\eta) \\ &\quad= \left(\frac{1}{1+e^{-1-\beta}} - \frac{1}{1+e^{1-\beta}} - \frac{1}{1+e^{\gamma-\beta}} + \frac{1}{1+e^{-\gamma-\beta}}\right)\eta + \left(\frac{1}{1+e^{1-\beta}} - \frac{1}{1+e^{-\gamma-\beta}}\right) \\ &\quad> \frac{1}{2}\left(\frac{1}{1+e^{-1-\beta}} + \frac{1}{1+e^{1-\beta}} + \frac{1}{1+e^{\gamma-\beta}} + \frac{1}{1+e^{-\gamma-\beta}}\right) \\ &\quad> 0, \end{aligned}$$

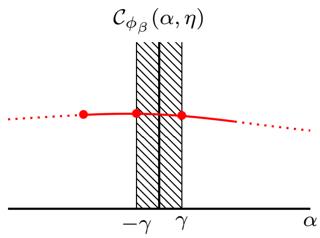


Figure 15: The class-conditional risk of the sigmoid loss.

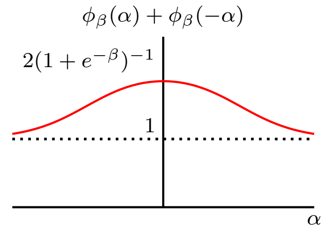


Figure 16: The even part of the sigmoid loss.

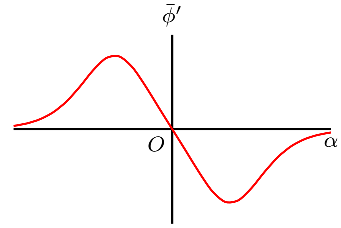


Figure 17: The derivative of the even part of the sigmoid loss.

we have $\mathcal{C}_{\phi_\beta}(\gamma, \eta) < \mathcal{C}_{\phi_\beta}(-1, \eta)$ for all $\eta \in [0, 1]$. By part 5 in Lemma 13, it follows that

$$\inf_{|\alpha| \le 1 : |\alpha| \le \gamma \text{ or } (2\eta - 1)\alpha \le 0} \mathcal{C}_{\phi_\beta}(\alpha, \eta) = \inf_{|\alpha| \le \gamma} \mathcal{C}_{\phi_\beta}(\alpha, \eta) = \mathcal{C}_{\phi_\beta}(\gamma, \eta).$$

Thus, by Lemma 7, $\bar{\delta}(\varepsilon, \eta) = A_0(\eta - \eta_0)$ if $\varepsilon \le \eta$, where

$$A_0 \overset{\text{def}}{=} \phi_\beta(\gamma) - \phi_\beta(-\gamma) - \phi_\beta(1) + \phi_\beta(-1), \qquad \eta_0 \overset{\text{def}}{=} \frac{\phi_\beta(-1) - \phi_\beta(-\gamma)}{A_0},$$

and $\bar{\delta}(\varepsilon, \eta) = \infty$ if $\varepsilon > \eta$. Note that $A_0 > 0$, $\eta_0 \le \frac{1}{2}$, and $\eta_0 = \frac{1}{2} \Leftrightarrow \beta = 0$. Hence we obtain

$$\delta(\varepsilon) = \inf_{\eta \in \left[\frac{1}{2}, 1\right]} \bar{\delta}(\varepsilon, \eta) = \begin{cases} A_1 & \text{if } 0 < \varepsilon \le \frac{1}{2}, \\ A_0(\varepsilon - \eta_0) & \text{if } \frac{1}{2} < \varepsilon, \end{cases}$$

where $A_1 \overset{\text{def}}{=} A_0\left(\frac{1}{2} - \eta_0\right) = (\phi_\beta(\gamma) + \phi_\beta(-\gamma) - \phi_\beta(1) - \phi_\beta(-1))/2$.

Thus, the sigmoid loss is calibrated wrt (ϕ_γ,$\mathcal{F}_{\text{lin}}$) when $A_1 > 0$, which is equivalent to $\beta > 0$.

Let $\check{\delta} : [0, 1] \to \mathbb{R}_{\ge 0}$ be a function such that $\check{\delta}(0) = \delta(0)$ and $\check{\delta}(\varepsilon) = \delta(\varepsilon)$ for all $\varepsilon > 0$. Then, the Fenchel-Legendre biconjugate of $\check{\delta}$ is

$$\check{\delta}^{**}(\varepsilon) = \begin{cases} 2A_1\varepsilon & \text{if } 0 \le \varepsilon \le \frac{1}{2}, \\ A_0(\varepsilon - \eta_0) & \text{if } \frac{1}{2} < \varepsilon. \end{cases}$$

C.2.2. QUASICONCAVITY OF EVEN PART

We confirm that $\phi_\beta(\alpha) + \phi_\beta(-\alpha)$ is quasiconcave when $\beta \ge 0$. $\phi_\beta(\alpha) + \phi_\beta(-\alpha)$ is plotted in Figure 16, and

$$\phi_\beta(\alpha) + \phi_\beta(-\alpha) = \frac{1}{1 + e^{\alpha - \beta}} + \frac{1}{1 + e^{-\alpha - \beta}} \quad \left(\overset{\text{def}}{=} \bar{\phi}(\alpha)\right).$$

Here, we use quasimonotonicity of $-\bar{\phi}'$ to show its quasiconcavity (see Appendix A). The derivative of $\bar{\phi}$ is

$$\bar{\phi}'(\alpha) = 4\cosh^{-2}\left(\frac{\alpha + \beta}{2}\right) - 4\cosh^{-2}\left(\frac{\alpha - \beta}{2}\right),$$

which is plotted in Figure 17. Take $\alpha_1, \alpha_2 \in \mathbb{R}$ such that $\alpha_1 < \alpha_2$ without loss of generality. Assume that $(-\bar{\phi}'(\alpha_1))(\alpha_2 - \alpha_1) > 0$, which implies $\bar{\phi}'(\alpha_1) < 0 \Leftrightarrow \alpha_1 > 0$. Hence, $\alpha_2 > 0$ and $\bar{\phi}'(\alpha_2) < 0$. We obtain $(-\bar{\phi}'(\alpha_2))(\alpha_2 - \alpha_1) > 0$ straightforwardly, which implies quasimonotonicity of $-\bar{\phi}'$. Therefore, $\bar{\phi}$ is quasiconcave.

C.3. Modified Squared Loss

We design a bounded and nonincreasing surrogate loss by modifying the squared loss, which we call modified squared loss here:

$$\phi(\alpha) = \begin{cases} 1 & \text{if } \alpha \leq 0, \\ (1-\alpha)^2 & \text{if } 0 < \alpha \leq 1, \\ 0 & \text{if } 1 < \alpha, \end{cases}$$

and consider the shifted version $\phi_\beta(\alpha) \stackrel{\text{def}}{=} \phi(\alpha - \beta)$:

$$\phi_\beta(\alpha) = \begin{cases} 1 & \text{if } \alpha \leq \beta, \\ (1-\alpha+\beta)^2 & \text{if } \beta < \alpha \leq 1+\beta, \\ 0 & \text{if } 1+\beta < \alpha. \end{cases}$$

C.3.1. CALIBRATION FUNCTION

Now we consider ϕ_β-CCR $\mathcal{C}_{\phi_\beta}(\alpha,\eta) = \eta\phi(\alpha) + (1-\eta)\phi(-\alpha)$ and focus on the case $\eta > \frac{1}{2}$ due to the symmetry of $\mathcal{C}_{\phi_\beta}$. $\mathcal{C}_{\phi_\beta}$ is plotted in Figure 18. By part 5 of Lemma 13, it is easy to see $\mathcal{C}^*_{\phi_\beta,\mathcal{F}_{\mathrm{lin}}}(\eta) = \min\{\mathcal{C}_{\phi_\beta}(-1,\eta), \mathcal{C}_{\phi_\beta}(1,\eta)\} = \mathcal{C}_{\phi_\beta}(1,\eta)$. We divide into three cases depending on the relationship among $\mathcal{C}_{\phi_\beta}(-1,\eta)$, $\mathcal{C}_{\phi_\beta}(-\gamma,\eta)$, and $\mathcal{C}_{\phi_\beta}(\gamma,\eta)$,

(A) When $0 \leq \beta < \gamma$**:** Since

$$\begin{aligned} \mathcal{C}_{\phi_\beta}(-\gamma,\eta) - \mathcal{C}_{\phi_\beta}(\gamma,\eta) &= \left\{\eta\cdot 1 + (1-\eta)(1-\gamma+\beta)^2\right\} - \left\{\eta(1-\gamma+\beta)^2 + (1-\eta)\cdot 1\right\} \\ &= (2\eta-1)(\gamma-\beta)\left\{2-(\gamma-\beta)\right\} \\ &\geq 0, \end{aligned}$$

we have $\mathcal{C}_{\phi_\beta}(\gamma,\eta) < \mathcal{C}_{\phi_\beta}(-\gamma,\eta)$ for all $\eta > \frac{1}{2}$. On the other hand, since

$$\mathcal{C}_{\phi_\beta}(\gamma,\eta) - \mathcal{C}_{\phi_\beta}(-1,\eta) = -\{(2-\gamma+\beta)(\gamma-\beta) + (1-\beta)^2\}(\eta-\eta_0)$$
$$\text{where} \quad \eta_0 \stackrel{\text{def}}{=} \frac{1-\beta^2}{(2-\gamma+\beta)(\gamma-\beta)+(1-\beta^2)},$$

and $\frac{1}{2} < \eta_0 \leq 1$, we have $\mathcal{C}_{\phi_\beta}(\gamma,\eta) \geq \mathcal{C}_{\phi_\beta}(-1,\eta)$ if $\frac{1}{2} < \eta \leq \eta_0$ and $\mathcal{C}_{\phi_\beta}(\gamma,\eta) < \mathcal{C}_{\phi_\beta}(-1,\eta)$ if $\eta > \eta_0$.

- <u>If $\frac{1}{2} < \eta \leq \eta_0$</u>: By part 5 in Lemma 13,

$$\inf_{|\alpha|\leq 1: |\alpha|\leq\gamma \text{ or } (2\eta-1)\alpha\leq 0} \mathcal{C}_{\phi_\beta}(\alpha,\eta) = \mathcal{C}_{\phi_\beta}(-1,\eta) \quad \text{and} \quad \inf_{|\alpha|\leq\gamma} \mathcal{C}_{\phi_\beta}(\alpha,\eta) = \mathcal{C}_{\phi_\beta}(\gamma,\eta).$$

  Thus, by Lemma 7,

$$\bar{\delta}(\varepsilon,\eta) = \begin{cases} \infty & \text{if } \eta < \varepsilon, \\ \mathcal{C}_{\phi_\beta}(\gamma,\eta) - \mathcal{C}^*_{\phi_\beta,\mathcal{F}_{\mathrm{lin}}}(\eta) = (1-\gamma)(1-\gamma+2\beta)\eta & \text{if } \varepsilon \leq \eta < \frac{1+\varepsilon}{2}, \\ \mathcal{C}_{\phi_\beta}(-1,\eta) - \mathcal{C}^*_{\phi_\beta,\mathcal{F}_{\mathrm{lin}}}(\eta) = (1-\beta^2)(2\eta-1) & \text{if } \frac{1+\varepsilon}{2} \leq \eta. \end{cases}$$

Hence we obtain

$$\inf_{\eta\in(\frac{1}{2},\eta_0]} \bar{\delta}(\varepsilon,\eta) = \begin{cases} (1-\beta^2)\varepsilon & \text{if } 0<\varepsilon\le\varepsilon_0, \\ \frac{(1-\gamma+2\beta)(1-\gamma)}{2} & \text{if } \varepsilon_0<\varepsilon\le\frac{1}{2}, \\ (1-\gamma)(1-\gamma+2\beta)\varepsilon & \text{if } \frac{1}{2}<\varepsilon, \end{cases}$$

where $\varepsilon_0 \stackrel{\text{def}}{=} \frac{(1-\gamma)(1-\gamma+2\beta)}{2(1-\beta^2)}$.

- If $\eta_0<\eta\le 1$: By part 5 in Lemma 13, it follows that

$$\inf_{|\alpha|\le 1:|\alpha|\le\gamma \text{ or } (2\eta-1)\alpha\le 0} \mathcal{C}_{\phi_\beta}(\alpha,\eta) = \inf_{|\alpha|\le\gamma} \mathcal{C}_{\phi_\beta}(\alpha,\eta) = \mathcal{C}_{\phi_\beta}(\gamma,\eta).$$

Thus, by Lemma 7, $\bar{\delta}(\varepsilon,\eta) = \mathcal{C}_{\phi_\beta}(\gamma,\eta) - \mathcal{C}^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta) = (1-\gamma)(1-\gamma+2\beta)\eta$ if $\varepsilon\le\eta$ and $\bar{\delta}(\varepsilon,\eta)=\infty$ if $\eta<\varepsilon$. Hence we obtain

$$\inf_{\eta\in(\eta_0,1]} \bar{\delta}(\varepsilon,\eta) = \begin{cases} \frac{(1-\gamma)(1-\gamma+2\beta)}{2} & \text{if } 0<\varepsilon\le\frac{1}{2}, \\ (1-\gamma)(1-\gamma+2\beta)\varepsilon & \text{if } \frac{1}{2}<\varepsilon. \end{cases}$$

Combining the above, we obtain the ϕ_γ-calibration function from Lemma 7:

$$\delta(\varepsilon) = \begin{cases} (1-\beta^2)\varepsilon & \text{if } 0<\varepsilon\le\varepsilon_0, \\ \frac{(1-\gamma+2\beta)(1-\gamma)}{2} & \text{if } \varepsilon_0<\varepsilon\le\frac{1}{2}, \\ (1-\gamma)(1-\gamma+2\beta)\varepsilon & \text{if } \frac{1}{2}<\varepsilon. \end{cases}$$

Note that $\varepsilon_0\le\frac{1}{2}$, which means the second case is not vacuous.
(B) When $\gamma\le\beta<1$: It is easy to see

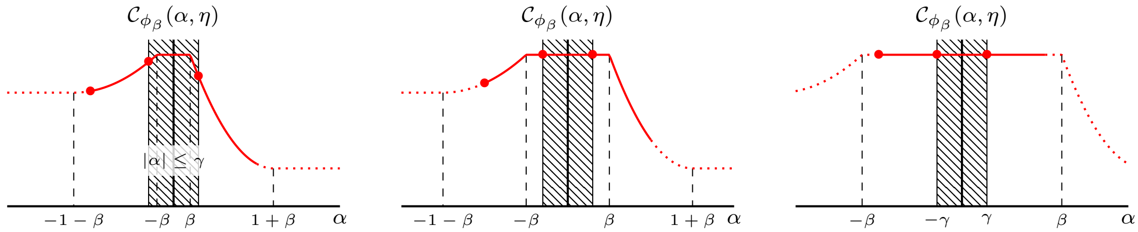
$$\inf_{|\alpha|\le\gamma} \mathcal{C}_{\phi_\beta}(\alpha,\eta) = 1,$$
$$\inf_{|\alpha|\le 1:|\alpha|\le\gamma \text{ or } (2\eta-1)\alpha\le 0} \mathcal{C}_{\phi_\beta}(\alpha,\eta) = \mathcal{C}_{\phi_\beta}(-1,\eta).$$

Hence, by part 5 in Lemma 13, it follows that

$$\bar{\delta}(\varepsilon,\eta) = \begin{cases} \infty & \text{if } \eta<\varepsilon, \\ 1-\mathcal{C}_{\phi_\beta}(1,\eta) = (1-\beta^2)\eta & \text{if } \varepsilon\le\eta<\frac{1+\varepsilon}{2}, \\ \mathcal{C}_{\phi_\beta}(-1,\eta)-\mathcal{C}_{\phi_\beta}(1,\eta) = (1-\beta^2)(2\eta-1) & \text{if } \frac{1+\varepsilon}{2}\le\eta. \end{cases}$$

Thus, by Lemma 7, $\delta(\varepsilon) = \inf_{\eta\in(\frac{1}{2},1]} \bar{\delta}(\varepsilon,\eta) = (1-\beta^2)\varepsilon$.
(C) When $1\le\beta$: In this case, $\mathcal{C}_{\phi_\beta}(\alpha,\eta)=1$ for all $\alpha\in[-1,1]$. Hence, $\Delta\mathcal{C}_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\alpha,\eta)=0$ and $\delta(\varepsilon)=0$.

To sum up, the (ϕ_γ,$\mathcal{F}_{\text{lin}}$)-calibration function and its Fenchel-Legendre biconjugate of the modified squared loss are as follows:



(a) $0 \le \beta < \gamma$ (b) $\gamma \le \beta < 1$ (c) $1 \le \beta$

Figure 18: The class-conditional risk of the modified squared loss.

- If $0 \le \beta < \gamma$,

$$\delta(\varepsilon) = \begin{cases} (1-\beta^2)\varepsilon & \text{if } 0 < \varepsilon \le \varepsilon_0, \\ \frac{(1-\gamma+2\beta)(1-\gamma)}{2} & \text{if } \varepsilon_0 < \varepsilon \le \frac{1}{2}, \\ (1-\gamma)(1-\gamma+2\beta)\varepsilon & \text{if } \frac{1}{2} < \varepsilon, \end{cases} \quad \text{and} \quad \delta^{**}(\varepsilon) = (1-\gamma)(1-\gamma+2\beta)\varepsilon,$$

  where $\varepsilon_0 \stackrel{\text{def}}{=} \frac{(1-\gamma)(1-\gamma+2\beta)}{2(1-\beta^2)}$.

- If $\gamma \le \beta < 1$, $\delta(\varepsilon) = \delta^{**}(\varepsilon) = (1-\beta^2)\varepsilon$.

- If $1 \le \beta$, $\delta(\varepsilon) = \delta^{**}(\varepsilon) = 0$.

We deduce that the modified squared loss is calibrated wrt (ϕ_γ,$\mathcal{F}_{\text{lin}}$) if $0 \le \beta < 1$.

C.3.2. QUASICONCAVITY OF EVEN PART

We confirm that $\phi_\beta(\alpha) + \phi_\beta(-\alpha)$ is quasiconcave when $\beta \ge 0$.

$$\phi_\beta(\alpha) + \phi_\beta(-\alpha) = \begin{cases} 1 & \alpha < -1-\beta, \\ (1+\alpha+\beta)^2 + 1 & -1-\beta \le \alpha < -\beta, \\ 2 & -\beta \le \alpha < \beta, \\ (1-\alpha+\beta)^2 + 1 & \beta \le \alpha < 1+\beta, \\ 1 & 1+\beta \le \alpha. \end{cases}$$

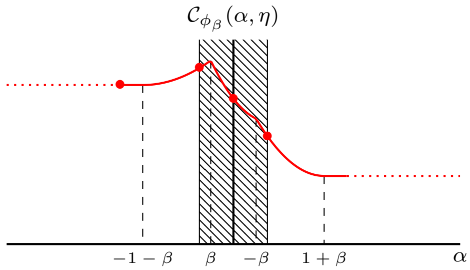
Its t-superlevel set S_t is as follows.

- If $t < 1$, $S_t = \mathbb{R}$.

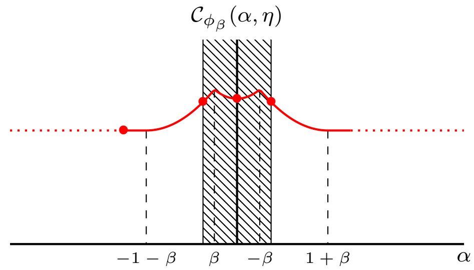
- If $1 \le t \le 2$, $S_t = \{\alpha \mid |\alpha| \le 1 + \beta - \sqrt{t-1}\}$.

- If $2 < t$, $S_t = \emptyset$.

In all cases, S_t is convex. Thus, $\phi_\beta(\alpha) + \phi_\beta(-\alpha)$ is quasiconcave.



(a) $\eta = 0.7$



(b) $\eta = 0.5$

Figure 19: The class-conditional risk of the modified squared loss when $\gamma < \frac{1}{4}$ and $-1 + \frac{1}{\sqrt{2}} < \beta < 0$.

C.3.3. WHEN $\beta < 0$

In this case, the modified squared loss is no longer quasiconcave even (see Figure 19 (b)). However, ϕ_β is still (ϕ_γ,$\mathcal{F}_{\text{lin}}$)-calibrated under some γ and $\beta < 0$. Here, we show an example.

Assume that $\gamma < \frac{1}{4}$ and $-1 + \frac{1}{\sqrt{2}} < \beta < 0$. We focus on $\eta > \frac{1}{2}$ due to the symmetry of $\mathcal{C}_{\phi_\beta}$. In these β and γ, we still have $\eta_0 > \frac{1}{2}$, because

$$\begin{aligned}
\eta_0 = \frac{1-\beta^2}{(2-\gamma+\beta)(\gamma-\beta)+(1-\beta^2)} &> \frac{1}{2} \\
\iff 2(1-\beta^2) &> (2-\gamma+\beta)(\gamma-\beta)+(1-\beta^2) \\
\iff \gamma^2 - 2(1+\beta)\gamma + (2\beta+1) &> 0 \\
\iff \gamma < 1+2\beta, \ \underbrace{1<\gamma}_{\text{always false}} & \\
\iff \gamma < 1+2\beta, &
\end{aligned}$$

and $1 + 2\beta > \sqrt{2}\left(1 - \frac{1}{\sqrt{2}}\right) > \frac{1}{4} > \gamma$ always holds when $\gamma < \frac{1}{4}$. Then, we can confirm in the same way as the case (A) that

- $\mathcal{C}_{\phi_\beta}(-\gamma,\eta) > \mathcal{C}_{\phi_\beta}(\gamma,\eta)$ for all $\eta > \frac{1}{2}$.
- $\mathcal{C}_{\phi_\beta}(\gamma,\eta) \geq \mathcal{C}_{\phi_\beta}(-1,\eta)$ if $\frac{1}{2} < \eta \leq \eta_0$, and $\mathcal{C}_{\phi_\beta}(\gamma,\eta) < \mathcal{C}_{\phi_\beta}(-1,\eta)$ if $\eta_0 < \eta$.

In addition, we see that

$$\begin{aligned}
\mathcal{C}_{\phi_\beta}(-1,\eta) - \mathcal{C}_{\phi_\beta}(0,\eta) &= \eta - (1+\beta)^2, \\
\mathcal{C}_{\phi_\beta}(0,\eta) - \mathcal{C}_{\phi_\beta}(1,\eta) &= (1+\beta)^2 - (1-\eta) > \frac{1}{2} - (1-\eta) > 0, \\
\mathcal{C}_{\phi_\beta}(\gamma,\eta) - \mathcal{C}_{\phi_\beta}(1,\eta) &= \underbrace{(1-4\gamma(1+\beta))}_{>1-4\gamma>0}\eta + (1+\gamma+\beta)^2 - 1 \\
&> \left(\frac{1}{2} - 2\gamma(1+\beta)\right) + (1+\gamma+\beta)^2 - 1 \\
&= \gamma^2 + \underbrace{\beta^2 + 2\beta + \frac{1}{2}}_{=(\beta+1)^2 - \frac{1}{2} > 0} \\
&> 0, \\
\mathcal{C}_{\phi_\beta}(0,\eta) - \mathcal{C}_{\phi_\beta}(\gamma,\eta) &= 4\gamma(\beta+1)\left(\eta - \eta_1\right),
\end{aligned}$$

where $\eta_1 \overset{\text{def}}{=} \frac{2+2\beta+\gamma}{4(1+\beta)} \in \left(\frac{1}{2}, 1\right]$. Then, we have

- $\mathcal{C}_{\phi_\beta}(-1,\eta) > \mathcal{C}_{\phi_\beta}(0,\eta)$ for $\eta > (1+\beta)^2$, and $\mathcal{C}_{\phi_\beta}(-1,\eta) \le \mathcal{C}_{\phi_\beta}(0,\eta)$ for $\frac{1}{2} < \eta \le (1+\beta)^2$.
- $\mathcal{C}_{\phi_\beta}(0,\eta) > \mathcal{C}_{\phi_\beta}(1,\eta)$ for all $\eta > \frac{1}{2}$.
- $\mathcal{C}_{\phi_\beta}(\gamma,\eta) > \mathcal{C}_{\phi_\beta}(1,\eta)$ for all $\eta > \frac{1}{2}$.
- $\mathcal{C}_{\phi_\beta}(\gamma,\eta) \ge \mathcal{C}_{\phi_\beta}(0,\eta)$ if $\frac{1}{2} < \eta \le \eta_1$, and $\mathcal{C}_{\phi_\beta}(\gamma,\eta) < \mathcal{C}_{\phi_\beta}(0,\eta)$ if $\eta_1 < \eta$.

Figure 19 and the above comparisons give us

$$\begin{aligned}
\mathcal{C}^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta) = \inf_{\alpha\in[-1,1]} \mathcal{C}_{\phi_\beta}(\alpha,\eta) &= \mathcal{C}_{\phi_\beta}(1,\eta), \\
\inf_{|\alpha|\le\gamma} \mathcal{C}_{\phi_\beta}(\alpha,\eta) &= \min\{\mathcal{C}_{\phi_\beta}(0,\eta), \mathcal{C}_{\phi_\beta}(\gamma,\eta)\}, \\
\inf_{\alpha\in[-1,\gamma]} \mathcal{C}_{\phi_\beta}(\alpha,\eta) &= \min\{\mathcal{C}_{\phi_\beta}(-1,\eta), \mathcal{C}_{\phi_\beta}(0,\eta), \mathcal{C}_{\phi_\beta}(\gamma,\eta)\}.
\end{aligned}$$

By Lemma 7, when $\varepsilon \le \eta < \frac{1+\varepsilon}{2}$,

$$\begin{aligned}
\bar{\delta}(\varepsilon,\eta) &= \inf_{|\alpha|\le\gamma} \mathcal{C}_{\phi_\beta}(\alpha,\eta) - \mathcal{C}^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta) \\
&= \min\left\{\mathcal{C}_{\phi_\beta}(0,\eta) - \mathcal{C}_{\phi_\beta}(1,\eta), \mathcal{C}_{\phi_\beta}(\gamma,\eta) - \mathcal{C}_{\phi_\beta}(1,\eta)\right\} \\
&= \min\left\{\eta + (\beta^2 + 2\beta), (1-4\gamma(1+\beta))\eta + (1+\gamma+\beta)^2 - 1\right\},
\end{aligned}$$

and

$$\inf_{\eta \in \left[\varepsilon, \frac{1+\varepsilon}{2}\right] \cap \left(\frac{1}{2}, 1\right]} \bar{\delta}(\varepsilon, \eta) = \min \left\{ \overbrace{\left[\varepsilon - \frac{1}{2}\right]_+ + \frac{1}{2} + (\beta^2 + 2\beta)}^{\text{(i)}}, \underbrace{(1 - 4\gamma(1+\beta)) \left(\left[\varepsilon - \frac{1}{2}\right]_+ + \frac{1}{2} \right) + (1 + \gamma + \beta)^2 - 1}_{\text{(ii)}} \right\}.$$

When $\frac{1+\varepsilon}{2} \leq \eta$,

$$\begin{aligned}
\bar{\delta}(\varepsilon, \eta) &= \inf_{\alpha \in [-1, \gamma]} \mathcal{C}_{\phi_\beta}(\alpha, \eta) - \mathcal{C}^*_{\phi_\beta, \mathcal{F}_{\text{lin}}}(\eta) \\
&= \min \left\{ \mathcal{C}_{\phi_\beta}(-1, \eta) - \mathcal{C}_{\phi_\beta}(1, \eta), \mathcal{C}_{\phi_\beta}(0, \eta) - \mathcal{C}_{\phi_\beta}(1, \eta), \mathcal{C}_{\phi_\beta}(\gamma, \eta) - \mathcal{C}_{\phi_\beta}(1, \eta) \right\} \\
&= \min \left\{ 2\eta - 1, \eta + (\beta^2 + 2\beta), (1 - 4\gamma(1+\beta))\eta + (1 + \gamma + \beta)^2 - 1 \right\},
\end{aligned}$$

and

$$\begin{aligned}
&\inf_{\eta \in \left[\frac{1+\varepsilon}{2}, 1\right] \cap \left(\frac{1}{2}, 1\right]} \bar{\delta}(\varepsilon, \eta) \\
&= \min \left\{ \underbrace{\varepsilon}_{\text{(iii)}}, \underbrace{\frac{1+\varepsilon}{2} + (\beta^2 + 2\beta)}_{\text{(iv)}}, \underbrace{(1 - 4\gamma(1+\beta)) \frac{1+\varepsilon}{2} + (1 + \gamma + \beta)^2 - 1}_{\text{(v)}} \right\}.
\end{aligned}$$

Note that for any $\gamma \in \left(0, \frac{1}{4}\right)$, $-1 + \frac{1}{\sqrt{2}} < \beta < 0$, and $\varepsilon > 0$, we have (iv) $\geq$ (i) and (v) $\geq$ (ii), which means that (ϕ_γ,$\mathcal{F}_{\text{lin}}$)-calibration function of ϕ_β is

$$\begin{aligned}
\delta(\varepsilon) &= \min \left\{ \inf_{\eta \in \left[\varepsilon, \frac{1+\varepsilon}{2}\right] \cap \left(\frac{1}{2}, 1\right]} \bar{\delta}(\varepsilon, \eta), \inf_{\eta \in \left[\frac{1+\varepsilon}{2}, 1\right] \cap \left(\frac{1}{2}, 1\right]} \bar{\delta}(\varepsilon, \eta) \right\} \\
&= \min\{\text{(i)}, \text{(ii)}, \text{(iii)}\} \\
&= \begin{cases}
\varepsilon & \text{if } 0 < \varepsilon \leq \varepsilon_0, \\
\varepsilon_0 & \text{if } \varepsilon_0 < \varepsilon \leq \frac{1}{2}, \\
\varepsilon + \beta^2 + 2\beta & \text{if } \frac{1}{2} < \varepsilon \leq \eta_1, \\
(1 - 4\gamma(1+\beta))\varepsilon + (1 + \gamma + \beta)^2 - 1 & \text{if } \eta_1 < \varepsilon \leq 1,
\end{cases}
\end{aligned}$$

where $\varepsilon_0 \stackrel{\text{def}}{=} \beta^2 + 2\beta + \frac{1}{2}$. From this result, we see that the modified squared loss is still (ϕ_γ,$\mathcal{F}_{\text{lin}}$)-calibrated when $0 < \gamma < \frac{1}{4}$ and $-1 + \frac{1}{\sqrt{2}} < \beta < 0$, and it is no longer calibrated once β becomes $-1 + \frac{1}{\sqrt{2}}$ (since $\varepsilon_0 = 0$ at $\beta = -1 + \frac{1}{\sqrt{2}}$).

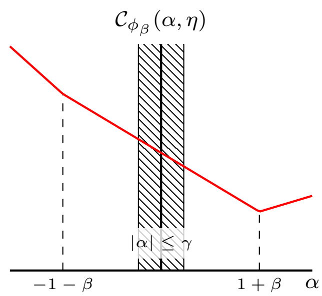


Figure 20: The class-conditional risk of the hinge loss.

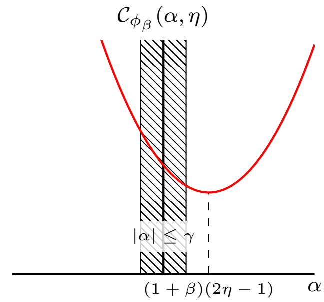


Figure 21: The class-conditional risk of the squared loss.

C.4. Hinge Loss

The ϕ_β-CCR is

$$\mathcal{C}_{\phi_\beta}(\alpha,\eta) = \begin{cases} -\eta\alpha + \eta(1+\beta) & \text{if } \alpha < -(1+\beta), \\ (1-2\eta)\alpha + (1+\beta) & \text{if } -(1+\beta) \le \alpha < 1+\beta, \\ (1-\eta)\alpha + (1-\eta)(1+\beta) & \text{if } 1+\beta < \alpha. \end{cases}$$

We restrict the range of η to $\eta > \frac{1}{2}$ by virtue of part 1 of Lemma 12. Then, $\mathcal{C}^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta) = \mathcal{C}_{\phi_\beta}(1,\eta) = -2\eta + (2+\beta)$. $\mathcal{C}_{\phi_\beta}(\alpha,\eta)$ is plotted in Figure 20 in case of $\eta > \frac{1}{2}$. Then, it follows that

$$\inf_{\substack{|\alpha|\le 1: |\alpha|\le\gamma \text{ or} \\ (2\eta-1)\alpha\le 0}} \mathcal{C}_{\phi_\beta}(\alpha,\eta) = \inf_{|\alpha|\le\gamma} \mathcal{C}_{\phi_\beta}(\alpha,\eta) = \mathcal{C}_{\phi_\beta}(\gamma,\eta) = (1-2\eta)\gamma + (1+\beta).$$

Hence, by Lemma 7,

$$\bar{\delta}(\varepsilon,\eta) = \begin{cases} \infty & \text{if } \varepsilon < \eta, \\ \mathcal{C}_{\phi_\beta}(\gamma,\eta) - \mathcal{C}^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta) = (1-\gamma)(2\eta-1) & \text{if } \eta \le \varepsilon, \end{cases}$$

and

$$\delta(\varepsilon) = \begin{cases} 0 & \text{if } 0 < \varepsilon \le \frac{1}{2}, \\ (1-\gamma)(2\varepsilon-1) & \text{if } \frac{1}{2} < \varepsilon. \end{cases}$$

C.5. Squared Loss

The ϕ_β-CCR is

$$\begin{aligned} \mathcal{C}_{\phi_\beta}(\alpha,\eta) &= \eta(1-\alpha+\beta)^2 + (1-\eta)(1+\alpha+\beta)^2 \\ &= \{\alpha - (1+\beta)(2\eta-1)\}^2 + 4(1+\beta)^2\eta(1-\eta). \end{aligned}$$

We restrict the range of η to $\eta > \frac{1}{2}$ by virtue of part 1 of Lemma 12. $\mathcal{C}_{\phi_\beta}(\alpha,\eta)$ is plotted in Figure 21 in case of $\eta > \frac{1}{2}$. By comparing $\alpha_* \overset{\text{def}}{=} (1+\beta)(2\eta-1)$ and 1, we have

$$\mathcal{C}^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta) = \begin{cases} \mathcal{C}_{\phi_\beta}(1,\eta) & \text{if } \alpha_* < 1, \\ \mathcal{C}_{\phi_\beta}(\alpha_*,\eta) & \text{if } \alpha_* \ge 1. \end{cases}$$

Then, it follows that

$$\inf_{\substack{|\alpha|\le 1:|\alpha|\le\gamma \text{ or} \\ (2\eta-1)\alpha\le 0}} \mathcal{C}_{\phi_\beta}(\alpha,\eta) = \inf_{|\alpha|\le\gamma} \mathcal{C}_{\phi_\beta}(\alpha,\eta) = \begin{cases} 0 & \text{if } \gamma > \alpha_*, \\ \mathcal{C}_{\phi_\beta}(\gamma,\eta) & \text{if } \gamma \le \alpha_*. \end{cases}$$

Hence, by Lemma 7,

$$\begin{aligned} \bar{\delta}(\varepsilon,\eta) &= \begin{cases} \infty & \text{if } \eta < \varepsilon, \\ \mathcal{C}_{\phi_\beta}(\alpha_*,\eta) - \mathcal{C}_{\phi_\beta}(\alpha_*,\eta) & \text{if } \varepsilon \le \eta \text{ and } \alpha_* \le \gamma, \\ \mathcal{C}_{\phi_\beta}(\gamma,\eta) - \mathcal{C}_{\phi_\beta}(\alpha_*,\eta) & \text{if } \varepsilon \le \eta \text{ and } \gamma < \alpha_* \le 1, \\ \mathcal{C}_{\phi_\beta}(\gamma,\eta) - \mathcal{C}_{\phi_\beta}(1,\eta) & \text{if } \varepsilon \le \eta \text{ and } 1 < \alpha_*, \end{cases} \\ &= \begin{cases} \infty & \text{if } \eta < \varepsilon, \\ 0 & \text{if } \varepsilon \le \eta \text{ and } \alpha_* \le \gamma, \\ 4(1+\beta)^2(\eta-\eta_0)^2 & \text{if } \varepsilon \le \eta \text{ and } \gamma < \alpha_* \le 1, \\ 4(1-\gamma)(1+\beta)(\eta-\eta_1) & \text{if } \varepsilon \le \eta \text{ and } 1 < \alpha_*, \end{cases} \\ &= \begin{cases} \infty & \text{if } \eta < \varepsilon, \\ 0 & \text{if } \varepsilon \le \eta \text{ and } \frac{1}{2} < \eta \le \eta_0, \\ 4(1+\beta)^2(\eta-\eta_0)^2 & \text{if } \varepsilon \le \eta \text{ and } \eta_0 < \eta \le \eta_2, \\ 4(1-\gamma)(1+\beta)(\eta-\eta_1) & \text{if } \varepsilon \le \eta \text{ and } \eta_2 < \eta \le 1, \end{cases} \end{aligned}$$

where $\eta_0 \stackrel{\text{def}}{=} \frac{1+\gamma+\beta}{2(1+\beta)}$, $\eta_1 \stackrel{\text{def}}{=} \frac{3+\gamma+2\beta}{4(1+\beta)}$, and $\eta_2 \stackrel{\text{def}}{=} \frac{2+\beta}{2(1+\beta)}$. Hence,

$$\delta(\varepsilon) = \begin{cases} 0 & \text{if } 0 < \varepsilon < \eta_0, \\ 4(1+\beta)^2(\varepsilon-\eta_0)^2 & \text{if } \eta_0 \le \varepsilon < \eta_2, \\ 4(1-\gamma)(1+\beta)(\varepsilon-\eta_1) & \text{if } \eta_2 \le \varepsilon. \end{cases}$$

Appendix D. Simulation Results

D.1. Detail of Numerical Approximation of Bayes Risks

Consider to compute the Bayes (ϕ,$\mathcal{F}_{\text{lin}}$)-risk for a loss ϕ.

$$\begin{aligned} \inf_{f\in\mathcal{F}_{\text{lin}}} \mathcal{R}_\phi(f) &= \mathbb{E}_X\left[\inf_{\alpha\in\mathcal{A}_{\mathcal{F}_{\text{lin}}}} \mathcal{C}_\phi(\alpha, \mathbb{P}(Y=+1|X))\right] \\ &= \mathbb{E}_X\left[\mathcal{C}^*_{\phi,\mathcal{F}_{\text{lin}}}(\mathbb{P}(Y=+1|X))\right]. \end{aligned} \tag{12}$$

We can utilize (12) to numerically approximate the Bayes risk. Let q_+ and q_- be probability density functions of $\mathcal{N}([2\ 2]^\top, I_2)$ and $\mathcal{N}([-2\ \ -2], I_2)$, respectively. Then,

$$\mathbb{P}(Y=+1|X) = \frac{\mathbb{P}(Y=+1)\mathbb{P}(X|Y=+1)}{\mathbb{P}(Y=+1)\mathbb{P}(X|Y=+1)+\mathbb{P}(Y=-1)\mathbb{P}(X|Y=-1)} = \frac{\frac{1}{2}q_+(X)}{\frac{1}{2}q_+(X)+\frac{1}{2}q_-(X)}.$$

For the concrete forms of $\mathcal{C}^*_{\phi,\mathcal{F}_{\text{lin}}}$, we obtain in Appendix C except the logistic loss as follows.

- Robust 0-1 loss: $\mathcal{C}^*_{\phi,\mathcal{F}_{\text{lin}}}(\eta) = \min\{\eta, 1-\eta\}$
- Ramp loss: $\mathcal{C}^*_{\phi,\mathcal{F}_{\text{lin}}}(\eta) = \min\left\{\frac{\beta}{2}\eta + (1-\eta), \eta + \frac{\beta}{2}\eta\right\}$
- Sigmoid loss: $\mathcal{C}^*_{\phi,\mathcal{F}_{\text{lin}}}(\eta) = \min\left\{\frac{\eta}{1+e^{1-\beta}} + \frac{1-\eta}{1+e^{-1-\beta}}, \frac{\eta}{1+e^{-1-\beta}} + \frac{1-\eta}{1+e^{1-\beta}}\right\}$
- Hinge loss: $\mathcal{C}^*_{\phi,\mathcal{F}_{\text{lin}}}(\eta) = 2\min\{\eta, 1-\eta\} + \beta$

For the logistic loss, it is not difficult to see

$$\mathcal{C}^*_{\phi,\mathcal{F}_{\text{lin}}}(\eta) = \begin{cases} \eta\log(1+e^{-\alpha^*+\beta}) + (1-\eta)\log(1+e^{\alpha^*+\beta}) & \text{if } \eta > \frac{1}{2}, \\ \eta\log(1+e^{\alpha^*+\beta}) + (1-\eta)\log(1+e^{-\alpha^*+\beta}) & \text{if } \eta \le \frac{1}{2}, \end{cases}$$

where $\alpha^* = \text{clip}_{[-1,1]}\left(\log\left(\frac{\eta}{1-\eta}\right)\right)$. By plugging these expressions into (12) and performing numerical integration, we can approximate the Bayes risks. In the simulation results, we performed numerical integration for the range $[-10, 10] \times [-10, 10]$ split by 200×200 segments. The partitioning quadrature method was used. The approximated Bayes risks are as follows.

- Robust 0-1 loss: 0.0023474
- Ramp loss: 0.10211
- Sigmoid loss: 0.31110
- Hinge loss: 0.20469
- Logistic loss: 0.37356

D.2. Full Simulation Results of Benchmark Dataset

We show the full simulation results of MNIST in Tables 2 and 3. Simulation details are as follows.

- Dataset: MNIST extracted with two digits (7,000 instances for each digit).
- Preprocessing: Reduced to 2-dimension with the principal component analysis.
- Train-test split: $14{,}000$ instances are randomly split into training and test data with the ratio 4 to 1.
- Model: Linear models $f(x) = \theta^\top x + \theta_0$ (θ and θ_0 are learnable parameters)
- Surrogate loss: The ramp, sigmoid, hinge, and logistic losses with shift $\beta = +0.5$.
- Target loss: the γ-adversarially robust 0-1 loss with $\gamma = 0.1$.
- Optimization: Batch gradient descent with $1{,}000$ iterations.

Table 2: The simulation results of the γ-adversarially robust 0-1 loss with $\gamma = 0.1$ and $\beta = 0.5$. 50 trials are conducted for each pair of a method and dataset. Standard errors (multiplied by 10^4) are shown in parentheses. Bold-faces indicate outperforming methods, chosen by one-sided t-test with the significant level 5%.

	Ramp	Sigmoid	Hinge	Logistic
0 vs 1	0.034 (3)	**0.017 (2)**	0.087 (12)	0.321 (19)
0 vs 2	**0.111 (7)**	0.133 (10)	**0.109 (8)**	0.281 (19)
0 vs 3	**0.107 (7)**	0.126 (8)	0.120 (9)	0.307 (18)
0 vs 4	**0.069 (6)**	0.093 (12)	0.072 (7)	0.269 (21)
0 vs 5	**0.233 (21)**	0.340 (25)	**0.233 (21)**	0.269 (16)
0 vs 6	**0.129 (8)**	0.167 (13)	**0.127 (8)**	0.287 (22)
0 vs 7	**0.067 (6)**	0.073 (6)	0.090 (9)	0.302 (18)
0 vs 8	**0.096 (7)**	0.123 (12)	0.100 (9)	0.263 (20)
0 vs 9	**0.082 (6)**	0.101 (8)	0.092 (8)	0.279 (22)

Table 3: The simulation results of the 0-1 loss with $\beta = 0.5$. 50 trials are conducted for each pair of a method and dataset. Standard errors (multiplied by 10^4) are shown in parentheses. Bold-faces indicate outperforming methods, chosen by one-sided t-test with the significant level 5%.

	Ramp	Sigmoid	Hinge	Logistic
0 vs 1	0.012 (2)	**0.005 (1)**	0.038 (7)	0.228 (18)
0 vs 2	**0.050 (5)**	0.059 (7)	0.058 (7)	0.206 (18)
0 vs 3	**0.047 (4)**	0.054 (6)	0.064 (8)	0.229 (15)
0 vs 4	**0.028 (4)**	**0.029 (4)**	0.032 (6)	0.184 (18)
0 vs 5	**0.117 (11)**	0.185 (20)	**0.117 (11)**	0.193 (15)
0 vs 6	**0.060 (5)**	0.080 (8)	0.063 (6)	0.206 (18)
0 vs 7	**0.027 (3)**	**0.027 (4)**	0.045 (6)	0.214 (18)
0 vs 8	**0.050 (6)**	0.054 (6)	0.054 (7)	0.186 (18)
0 vs 9	**0.040 (4)**	0.044 (5)	0.046 (6)	0.192 (20)