Calibration and Consistency of Adversarial Surrogate Losses

Pranjal Awasthi
Google Research
New York, NY 10011
pranjalawasthi@google.com

Natalie S. Frank
Courant Institute
New York, NY 10012
nf1066@nyu.edu

Anqi Mao
Courant Institute
New York, NY 10012
aqmao@cims.nyu.edu

Mehryar Mohri
Google Research & Courant Institute
New York, NY 10011
mohri@google.com

Yutao Zhong
Courant Institute
New York, NY 10012
yutao@cims.nyu.edu

Abstract

Adversarial robustness is an increasingly critical property of classifiers in applications. The design of robust algorithms relies on surrogate losses since the optimization of the adversarial loss with most hypothesis sets is NP-hard. But, which surrogate losses should be used and when do they benefit from theoretical guarantees? We present an extensive study of this question, including a detailed analysis of the $\mathcal{H}$-*calibration* and $\mathcal{H}$-*consistency* of adversarial surrogate losses. We show that convex loss functions, or the supremum-based convex losses often used in applications, are not $\mathcal{H}$-*calibrated* for common hypothesis sets used in machine learning. We then give a characterization of $\mathcal{H}$-*calibration* and prove that some surrogate losses are indeed $\mathcal{H}$-*calibrated* for the adversarial zero-one loss, with common hypothesis sets. In particular, we fix some calibration results presented in prior work for a family of linear models and significantly generalize the results to the nonlinear hypothesis sets. Next, we show that $\mathcal{H}$-*calibration* is not sufficient to guarantee consistency and prove that, in the absence of any distributional assumption, no continuous surrogate loss is consistent in the adversarial setting. This, in particular, proves that a claim made in prior work is inaccurate. Next, we identify natural conditions under which some surrogate losses that we describe in detail are $\mathcal{H}$-*consistent*. We also report a series of empirical results which show that many $\mathcal{H}$-*calibrated* surrogate losses are indeed not $\mathcal{H}$-*consistent*, and validate our theoretical assumptions. Our adversarial $\mathcal{H}$-consistency results are novel, even for the case where $\mathcal{H}$ is the family of all measurable functions.

1 Introduction

Complex multi-layer neural networks trained on large datasets have achieved a remarkable performance in several applications in recent years, in particular in speech and visual recognition tasks (Sutskever et al., 2014; Krizhevsky et al., 2012). However, these rich models are susceptible to imperceptible perturbations (Szegedy et al., 2013). A complex neural network may, for example, misclassify a traffic sign, as a result of a minor variation, which may be the presence of a small advertisement sticker on the sign. Such misclassifications can have dramatic consequences in practice, for example with self-driving cars. These concerns have motivated the study of *adversarial robustness*, that is the design of classifiers that are robust to small ℓ_p norm input perturbations (Goodfellow et al., 2014; Madry et al., 2017; Tsipras et al., 2018; Carlini and Wagner, 2017). The standard $0/1$ loss is

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

then replaced with a more stringent *adversarial loss*, which requires a predictor to correctly classify an input point $\mathbf{x}$ and also to maintain the same classification for all points at a small ℓ_p distance of $\mathbf{x}$. But, can we devise efficient learning algorithms with theoretical guarantees for the adversarial loss?

Designing such robust algorithms requires resorting to appropriate surrogate losses since optimizing the adversarial loss is NP-hard for most hypothesis sets. A key property for surrogate adversarial losses is their consistency, that is, that exact or near optimal minimizers of the surrogate loss be also exact or near optimal minimizers of the original adversarial loss. The notion of consistency has been extensively studied in the case of the standard $0/1$ loss or the multi-class setting (Zhang, 2004; Bartlett et al., 2006; Tewari and Bartlett, 2007; Steinwart, 2007). However, those results or proof techniques cannot be used to establish or characterize consistency in adversarial settings. This is because the adversarial loss of a predictor f at point $\mathbf{x}$ is inherently not just a function of $f(\mathbf{x})$ but also of its values around a neighborhood of $\mathbf{x}$. As we shall see, the study of consistency is significantly more complex in the adversarial setting, with subtleties that have in fact led to some inaccurate claims made in prior work that we discuss later.

Consistency requires a property of the surrogate and the original losses to hold true for the family of all measurable functions. As argued by Long and Servedio (2013), the notion of $\mathcal{H}$-*consistency*, which requires a similar property for the surrogate and original losses, but with the near or optimal minimizers considered on the restricted hypothesis set $\mathcal{H}$, is a more relevant and desirable property for learning. Long and Servedio (2013) gave examples of surrogate losses that are not $\mathcal{H}$-*consistent* when $\mathcal{H}$ is the class of all measurable functions but that satisfy a *realizable* $\mathcal{H}$-*consistency* condition when $\mathcal{H}$ is the class of linear functions. More recently, Zhang and Agarwal (2020) studied the notion of *improper realizable* $\mathcal{H}$-*consistency* of linear classes where the surrogate ϕ can be optimized over a larger class, such as that of piecewise linear functions. Note that these studies only deal with the standard $0/1$ classification loss. This motivates our main objective: an extensive study of the $\mathcal{H}$-*consistency* of adversarial surrogate losses, which is critical to the design of robust algorithms with guarantees in this setting.

A more convenient notion in the study of $\mathcal{H}$-*consistency* is that of $\mathcal{H}$-*calibration*, which is a related notion that involves conditioning on the input point. $\mathcal{H}$-*calibration* often is a sufficient condition for $\mathcal{H}$-*consistency* in the standard classification settings (Steinwart, 2007). However, the adversarial loss presents new challenges and requires carefully distinguishing among these notions to avoid drawing false conclusions. As an example, the recent COLT 2020 paper of Bao et al. (2020a) presents a study of $\mathcal{H}$-*calibration* for the adversarial loss in the special case where $\mathcal{H}$ is the class of linear functions. However, several comments are due regarding that work. See a detailed discussion in Appendix B.

Our Contributions. We present a more systematic study of the $\mathcal{H}$-calibration and $\mathcal{H}$-consistency including for the case where $\mathcal{H} = \mathcal{H}_{\mathrm{all}}$ of adversarial surrogate losses. In Section 4, we give a detailed analysis of the $\mathcal{H}$-*calibration* properties of several natural surrogate losses. We present a series of new negative results showing that, under some general assumptions, convex loss functions and *supremum-based convex losses*, that are loss functions defined as the supremum over a ball of a convex function, which are those commonly used in applications, are not $\mathcal{H}$-*calibrated* for common hypothesis sets used in machine learning. Next, we give a characterization of calibration and prove that a family of proposed surrogates are $\mathcal{H}$-*calibrated*, with common hypothesis sets. These fix previous calibration results presented for the family of linear models in (Bao et al., 2020a) and significantly generalize the results to the nonlinear hypothesis sets. In Section 5, we study the $\mathcal{H}$-*consistency* of surrogate loss functions. We prove that, in the absence of distributional assumptions, many surrogate losses shown to be $\mathcal{H}$-*calibrated* in Section 4 are in fact not $\mathcal{H}$-*consistent*. This, in particular, proves that a claim presented in a COLT 2020 publication is inaccurate. Next, in contrast, we show that when the minimum of the surrogate loss is achieved within $\mathcal{H}$, under some general conditions, the ρ-margin ramp loss (see, for example, (Mohri et al., 2018)) is $\mathcal{H}$-*consistent* for $\mathcal{H}$ being the linear hypothesis set, or any non-decreasing and continuous g-based hypothesis set, including the ReLU-based hypothesis set. We then give similar $\mathcal{H}$-*consistency* guarantees for supremum-based surrogate losses based on a non-increasing auxiliary function, including the calibrated supremum-based ρ-margin ramp loss when $\mathcal{H}$ is any symmetric hypothesis set, e.g., the multi-layer neural networks. In Section 6, we further report a series of empirical results on simulated data, which show that many $\mathcal{H}$-*calibrated* surrogate losses are indeed not $\mathcal{H}$-*consistent*, and justify our conditions for consistency. Overall, our results imply that the loss functions commonly used in practice for optimizing the adversarial loss are not $\mathcal{H}$-consistent and that minimizing such losses may not lead to a more favorable adversarial loss. This could be in fact the reason why the empirical results reported in the literature have not been

favorable. Instead, we suggest alternative surrogate losses that we prove are $\mathcal{H}$-consistent and that can be useful to the design of effective algorithms.

We give a detailed discussion of related work in Appendix A. We start with basic concepts of calibration and consistency (Section 2) and an introduction of robust classification (Section 3).

2 Preliminaries

We will denote vectors as lowercase bold letters (e.g. $\mathbf{x}$). The d-dimensional l_2-ball with radius r is denoted by $B_2^d(r) := \left\{\mathbf{z} \in \mathbb{R}^d \mid \|\mathbf{z}\|_2 \leq r\right\}$. We denote by $\mathcal{X}$ the set of all possible examples. $\mathcal{X}$ is also sometimes referred to as the input space. The set of all possible labels is denoted by $\mathcal{Y}$. We will limit ourselves to the case of binary classification where $\mathcal{Y} = \{-1, +1\}$. Let $\mathcal{H}$ be a family of functions from $\mathbb{R}^d$ to $\mathbb{R}$. Given a fixed but unknown distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$, the binary classification learning problem is then formulated as follows. The learner seeks to select a predictor $f \in \mathcal{H}$ with small *generalization error* with respect to the distribution $\mathcal{P}$. The *generalization error* of a classifier $f \in \mathcal{H}$ is defined by $\mathcal{R}_{\ell_0}(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[\ell_0(f, \mathbf{x}, y)]$, where $\ell_0(f, \mathbf{x}, y) = \mathbb{1}_{yf(\mathbf{x})\leq 0}$ is the standard $0/1$ loss. More generally, the *ℓ-risk* of a classifier f for a surrogate loss $\ell(f, \mathbf{x}, y)$ is defined by

$$\mathcal{R}_\ell(f) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{P}}[\ell(f, \mathbf{x}, y)]. \tag{1}$$

Moreover, the *minimal (ℓ, $\mathcal{H}$)-risk*, which is also called the *Bayes (ℓ, $\mathcal{H}$)-risk*, is defined by $\mathcal{R}^*_{\ell,\mathcal{H}} = \inf_{f\in\mathcal{H}} \mathcal{R}_\ell(f)$. In the standard classification setting, the goal of a consistency analysis is to determine whether the minimization of a surrogate loss ℓ can lead to that of the binary loss generalization error. Similarly, in adversarially robust classification, the goal of a consistency analysis is to determine if the minimization of a surrogate loss ℓ yields that of the *adversarial generalization error* defined by $\mathcal{R}_{\ell_\gamma}(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[\ell_\gamma(f, \mathbf{x}, y)]$, where

$$\ell_\gamma(f, \mathbf{x}, y) := \sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\leq\gamma} \mathbb{1}_{yf(\mathbf{x}')\leq 0} \tag{2}$$

is the *adversarial* $0/1$ *loss*. This motivates the definition of $\mathcal{H}$-*consistency*.

Definition 1 ($\mathcal{H}$-Consistency)**.** *Given a hypothesis set $\mathcal{H}$, we say that a loss function ℓ_1 is* $\mathcal{H}$-consistent *with respect to a loss function ℓ_2, if the following holds:*

$$\mathcal{R}_{\ell_1}(f_n) - \mathcal{R}^*_{\ell_1,\mathcal{H}} \xrightarrow{n\to+\infty} 0 \implies \mathcal{R}_{\ell_2}(f_n) - \mathcal{R}^*_{\ell_2,\mathcal{H}} \xrightarrow{n\to+\infty} 0, \tag{3}$$

for all probability distributions and sequences of $\{f_n\}_{n\in\mathbb{N}} \subset \mathcal{H}$.

For a distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$ with random variables X and Y, let $\eta_\mathcal{P}: \mathcal{X} \to [0, 1]$ be a measurable function such that, for any $\mathbf{x} \in \mathcal{X}$, $\eta_\mathcal{P}(\mathbf{x}) = \mathcal{P}(Y = 1 \mid X = \mathbf{x})$. By the property of conditional expectation, we can rewrite (1) as $\mathcal{R}_\ell(f) = \mathbb{E}_X[\mathcal{C}_\ell(f, \mathbf{x}, \eta_\mathcal{P}(\mathbf{x}))]$, where $\mathcal{C}_\ell(f, \mathbf{x}, \eta)$ is the *inner ℓ-risk* defined as followed:

$$\forall \mathbf{x} \in \mathcal{X}, \forall \eta \in [0, 1], \quad \mathcal{C}_\ell(f, \mathbf{x}, \eta) := \eta\ell(f, \mathbf{x}, +1) + (1 - \eta)\ell(f, \mathbf{x}, -1). \tag{4}$$

Moreover, the *minimal inner ℓ-risk* on $\mathcal{H}$ is denoted by $\mathcal{C}^*_{\ell,\mathcal{H}}(\mathbf{x}, \eta) := \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f, \mathbf{x}, \eta)$. For a margin-based loss ϕ, the *generic conditional ϕ-risk* is $\bar{\mathcal{C}}_\phi(t, \eta) := \eta\phi(t) + (1 - \eta)\phi(-t)$ for any $\eta \in [0, 1]$ and $t \in \mathbb{R}$ (Bartlett et al., 2006). The notion of *calibration* for the inner risk is often a powerful tool for the analysis of $\mathcal{H}$-consistency (Steinwart, 2007).

Definition 2 ($\mathcal{H}$-Calibration)**.** [Definition 2.7 in (Steinwart, 2007)] *Given a hypothesis set $\mathcal{H}$, we say that a loss function ℓ_1 is* $\mathcal{H}$-calibrated *with respect to a loss function ℓ_2 if, for any $\epsilon > 0$, $\eta \in [0, 1]$, and $\mathbf{x} \in \mathcal{X}$, there exists $\delta > 0$ such that for all $f \in \mathcal{H}$ we have*

$$\mathcal{C}_{\ell_1}(f, \mathbf{x}, \eta) < \mathcal{C}^*_{\ell_1,\mathcal{H}}(\mathbf{x}, \eta) + \delta \implies \mathcal{C}_{\ell_2}(f, \mathbf{x}, \eta) < \mathcal{C}^*_{\ell_2,\mathcal{H}}(\mathbf{x}, \eta) + \epsilon. \tag{5}$$

Steinwart (2007) points out that if ℓ_1 is $\mathcal{H}$-calibrated wrt ℓ_2, then $\mathcal{H}$-consistency, that is condition (3), holds for any probability distribution verifying the additional condition of *minimizability* (Steinwart, 2007, Definition 2.4). Next, we introduce the notions of *calibration function* from (Steinwart, 2007).

Definition 3 (Calibration function)**.** *Given a hypothesis set $\mathcal{H}$, we define the* calibration function $\delta_{\max}$ *for a pair of losses (ℓ_1, ℓ_2) as follows: for all $\mathbf{x} \in \mathcal{X}$, $\eta \in [0, 1]$ and $\epsilon > 0$,*

$$\delta_{\max}(\epsilon, \mathbf{x}, \eta) = \inf_{f\in\mathcal{H}} \left\{\mathcal{C}_{\ell_1}(f, \mathbf{x}, \eta) - \mathcal{C}^*_{\ell_1,\mathcal{H}}(\mathbf{x}, \eta) \mid \mathcal{C}_{\ell_2}(f, \mathbf{x}, \eta) - \mathcal{C}^*_{\ell_2,\mathcal{H}}(\mathbf{x}, \eta) \geq \epsilon\right\}. \tag{6}$$

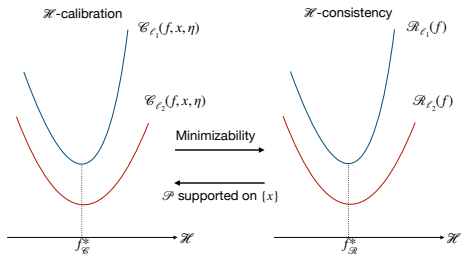

Figure 1: Illustration of $\mathcal{H}$-calibration and $\mathcal{H}$-consistency. Left: $\mathcal{H}$-calibration, for any $\mathbf{x} \in \mathcal{X}$, minimization of $\mathcal{C}_{\ell_1}(f, \mathbf{x})$ can lead to that of $\mathcal{C}_{\ell_2}(f, \mathbf{x})$. Right: $\mathcal{H}$-consistency, minimization of $\mathcal{R}_{\ell_1}(f)$ can lead to that of $\mathcal{R}_{\ell_2}(f)$. $\mathcal{H}$-consistency reduces to $\mathcal{H}$-calibration when the support of underlying distribution $\mathcal{P}$ is the single point set $\{\mathbf{x}\} \subset \mathcal{X}$; Under the minimizability condition, $\mathcal{H}$-calibration would imply $\mathcal{H}$-consistency.

For any $\mathbf{x} \in \mathcal{X}$, $\eta \in [0,1]$ and $\epsilon > 0$, the calibration function gives the maximal δ satisfying the calibration condition (5). The following proposition is an important result from (Steinwart, 2007).

Proposition 4 (Lemma 2.9 in (Steinwart, 2007))**.** *Given a hypothesis set $\mathcal{H}$, loss ℓ_1 is $\mathcal{H}$-calibrated with respect to ℓ_2 if and only if its calibration function $\delta_{\max}$ satisfies $\delta_{\max}(\epsilon, \mathbf{x}, \eta) > 0$ for all $\mathbf{x} \in \mathcal{X}$, $\eta \in [0,1]$ and $\epsilon > 0$.*

Since the concepts of calibration and consistency may not be familiar to readers without an extensive background in this area, we further comment on these notions before presenting our main results. Informally, a loss function is $\mathcal{H}$-consistent if minimizing it results in a classifier whose generalization error is close to the minimal generalization error within $\mathcal{H}$. Similarly, a loss function is $\mathcal{H}$-calibrated if minimizing it results in a classifier whose inner ℓ_2-risk is close to the minimal inner ℓ_2-risk within $\mathcal{H}$ for each $\mathbf{x} \in \mathcal{X}$. $\mathcal{H}$-calibration (5) is a necessary condition for $\mathcal{H}$-consistency (3), but is not always sufficient. As an example, we show in Section 5.1 that $\mathcal{H}$-calibrated surrogate losses proposed in (Bao et al., 2020a) are not $\mathcal{H}$-consistent. For this reason, $\mathcal{H}$-consistency, that is consistency for a particular hypothesis set $\mathcal{H}$, is a difficult problem even in standard non-adversarial scenarios. When $\mathcal{H}$ is the family of all measurable functions, the notions of calibration and consistency with respect to the $0/1$ loss have been widely studied in the literature to analyze the properties of margin-based losses (Zhang, 2004; Bartlett et al., 2006). In this special case, calibration implies consistency. Steinwart (2007) further establishes a sufficient condition called minimizability under which $\mathcal{H}$-calibration (5) implies $\mathcal{H}$-consistency (3). Note that the minimizability condition holds in (Zhang, 2004; Bartlett et al., 2006). However, it does not hold in general in the adversarial scenario and thus analyzing $\mathcal{H}$-consistency becomes much harder. To the best of our knowledge, our work is the first to prove $\mathcal{H}$-consistency results for general hypothesis sets $\mathcal{H}$, including for the case where $\mathcal{H} = \mathcal{H}_{\text{all}}$, in the context of adversarial classification. We conclude this section with an illustration of the connection between the notations of calibration and consistency in Figure 1.

3 Adversarially Robust Classification

In adversarially robust classification, the loss at $(\mathbf{x}, y)$ is measured in terms of the worst loss incurred over an adversarial perturbation of $\mathbf{x}$ within a ball of a certain radius in a norm. For simplicity, we will consider perturbations in the l_2 norm $\|\cdot\|$.[1] We will denote by γ the maximum magnitude of the allowed perturbations. Given $\gamma > 0$, a data point $(\mathbf{x}, y)$, a function $f \in \mathcal{H}$, and a margin-based loss $\phi: \mathbb{R} \to \mathbb{R}_+$, we define the *adversarial loss* of f at $(\mathbf{x}, y)$ as

$$\tilde{\phi}(f, \mathbf{x}, y) = \sup_{\mathbf{x}': \|\mathbf{x}-\mathbf{x}'\| \le \gamma} \phi(y f(\mathbf{x}')). \tag{7}$$

The above naturally motivates *supremum-based* surrogate losses that are commonly used to optimize the adversarial $0/1$ loss (Goodfellow et al., 2014; Madry et al., 2017; Shafahi et al., 2019; Wong et al.,

[1] Our analysis in the paper can be extended directly to other perturbations such as the l_1 ball or l_∞ ball, and in fact for any l_p norm for $p \in [1, \infty]$. In particular, the proofs of our calibration and consistency results for general hypothesis sets (e.g., Theorem 6, Theorem 7, Theorem 10, Theorem 16, Theorem 20, Theorem 23, Theorem 24) do not require the norm being l_2 and work for other norms too.

2020). We say that a surrogate loss $\tilde{\phi}(f, \mathbf{x}, y)$ is *supremum-based* if it is of the form defined in (7). We say that the supremum-based surrogate is convex if the function ϕ in (7) is convex. When ϕ is non-increasing, the following equality holds (Yin et al., 2019):

$$\sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\leq\gamma} \phi(yf(\mathbf{x}')) = \phi\left(\inf_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\leq\gamma} yf(\mathbf{x}')\right). \tag{8}$$

The adversarial $0/1$ loss defined in (2) is a special case of (7), where ϕ is the $0/1$ loss, that is, $\phi(yf(\mathbf{x})) = \ell_0(f, \mathbf{x}, y) = \mathbb{1}_{yf(\mathbf{x})\leq 0}$. Therefore, the adversarial $0/1$ loss has the equivalent form

$$\ell_\gamma(f, \mathbf{x}, y) = \sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\leq\gamma} \mathbb{1}_{yf(\mathbf{x}')\leq 0} = \mathbb{1}_{\inf_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\leq\gamma} yf(\mathbf{x}')\leq 0}. \tag{9}$$

This alternative equivalent form of adversarial $0/1$ loss is more advantageous to analyze than (2) and would be adopted in our proofs. Without loss of generality, let $\mathcal{X} = B_2^d(1)$ and $\gamma \in (0, 1)$. In this paper, we aim to characterize surrogate losses ℓ_1 satisfying $\mathcal{H}$-consistency (3) and $\mathcal{H}$-calibration (5) with $\ell_2 = \ell_\gamma$ and for the hypothesis sets $\mathcal{H}$ which are *regular for adversarial calibration*.

Definition 5 (Regularity for Adversarial Calibration)**.** *We say that a hypothesis set $\mathcal{H}$ is* regular for adversarial calibration *if there exists a* distinguishing $\mathbf{x}$ *in $\mathcal{X}$, that is if there exist $f, g \in \mathcal{H}$ such that* $\inf_{\|\mathbf{x}'-\mathbf{x}\|\leq\gamma} f(\mathbf{x}') > 0$ *and* $\sup_{\|\mathbf{x}'-\mathbf{x}\|\leq\gamma} g(\mathbf{x}') < 0$.

When studying $\mathcal{H}$-calibration of surrogate losses, it suffices to study sets $\mathcal{H}$ that are regular for adversarial calibration not only because all common hypothesis sets admit the property, but also because of the following result. (See Appendix E.1 for the proof.) We say that a hypothesis set $\mathcal{H}$ is *symmetric*, if for any $f \in \mathcal{H}$, $-f$ is also in $\mathcal{H}$.

Theorem 6. *Let $\mathcal{H}$ be a symmetric hypothesis set. If $\mathcal{H}$ is not regular for adversarial calibration, then any surrogate loss ℓ is $\mathcal{H}$-calibrated with respect to ℓ_γ.*

Moreover, we specifically study the following hypothesis sets that are regular for adversarial calibration: *linear models*: $\mathcal{H}_{\text{lin}} = \{\mathbf{x} \to \mathbf{w}\cdot\mathbf{x} \mid \|\mathbf{w}\| = 1\}$, as in (Bao et al., 2020a); *generalized linear models*: $\mathcal{H}_g = \{\mathbf{x} \to g(\mathbf{w}\cdot\mathbf{x}) + b \mid \|\mathbf{w}\| = 1, |b| \leq G\}$ where g is a non-decreasing function; *the family of all measurable functions*: $\mathcal{H}_{\text{all}}$; and *multi-layer neural networks*: $\mathcal{H}_{\text{NN}} = \{\mathbf{x} \to \mathbf{u}\cdot\rho_n(\mathbf{W}_n(\cdots\rho_2(\mathbf{W}_2\rho_1(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2)\cdots) + b_n) \mid \|\mathbf{u}\|_1 \leq \Lambda, \|\mathbf{W}_j\| \leq W, \|\mathbf{b}_j\|_1 \leq B\}$, where ρ_j is an activation function; In the special case of $g = (\cdot)_+ = \max(\cdot, 0)$, we denote the corresponding *ReLU-based hypothesis set* by $\mathcal{H}_{\text{relu}} = \{\mathbf{x} \to (\mathbf{w}\cdot\mathbf{x})_+ + b \mid \|\mathbf{w}\| = 1, |b| \leq G\}$.

4 $\mathcal{H}$-Calibration

Calibration is a condition that often guarantees consistency and is a first step in analyzing surrogate losses. Thus, in this section, we first present a detailed study of the calibration properties of several loss functions. We first give a series of negative results showing that, under general assumptions, convex losses and supremum-based convex losses, which are typically used in practice for adversarial robustness, are not calibrated. We then complement these results with positive ones by identifying a family of losses that are indeed calibrated under certain general conditions.

4.1 Negative results: convex losses

We first study convex losses, which are often used for standard binary classification problems.

Theorem 7. *Assume $\mathcal{H}$ is such that there exists a distinguishing $\mathbf{x}_0 \in \mathcal{X}$ and $f_0 \in \mathcal{H}$ such that $f_0(\mathbf{x}_0) = 0$. If a margin-based loss $\phi: \mathbb{R} \to \mathbb{R}_+$ is convex, then it is not $\mathcal{H}$-calibrated with respect to ℓ_γ.*

In particular, the assumption holds when $\mathcal{H}$ is regular for adversarial calibration and contains 0. By Theorem 7, we obtain the following corollary, which fixes the main negative result of Bao et al. (2020a) and generalizes the result to nonlinear hypothesis sets. Note $\mathcal{H}_{\text{lin}}$, $\mathcal{H}_{\text{NN}}$ and $\mathcal{H}_{\text{all}}$ all satisfy there exists a distinguishing $\mathbf{x}_0 \in \mathcal{X}$ and $f_0 \in \mathcal{H}$ such that $f_0(\mathbf{x}_0) = 0$. When $g(-\gamma) + G > 0$ and $g(\gamma) - G < 0$, $\mathcal{H}_g$ also satisfies this assumption. Verifying this condition on $\mathcal{H}_g$ is straightforward for G sufficiently large.

Corollary 8. *If a margin-based loss $\phi: \mathbb{R} \to \mathbb{R}_+$ is convex, then ϕ is not $\mathcal{H}$-calibrated with respect to ℓ_γ, for $\mathcal{H} = \mathcal{H}_{\text{lin}}$, $\mathcal{H}_g$ with a non-decreasing and continuous function g such that $g(-\gamma) + G > 0$ and $g(\gamma) - G < 0$, $\mathcal{H}_{\text{relu}}$ with $G > \gamma$, $\mathcal{H}_{\text{NN}}$, and $\mathcal{H}_{\text{all}}$.*

While convex surrogates are natural for the $0/1$ loss, the current practice in designing practical algorithms for the adversarial loss involves using convex supremum-based surrogates (Madry et al., 2017; Wong et al., 2020; Shafahi et al., 2019). We next investigate such losses.

4.2 Negative results: supremum-based convex losses

We study losses of the type $\tilde{\phi}(f, \mathbf{x}, y) = \sup_{\mathbf{x}': \|\mathbf{x}-\mathbf{x}'\| \le \gamma} \phi(y f(\mathbf{x}'))$, with ϕ convex, which are often used in practice as surrogates for the adversarial $0/1$ loss. The following theorems presents negative results for supremum-based convex surrogate losses for the common hypothesis sets $\mathcal{H}$.

Theorem 9. *Let ϕ be a convex and non-increasing margin-based loss. Consider the surrogate loss defined by $\tilde{\phi}(f, \mathbf{x}, y) = \sup_{\mathbf{x}': \|\mathbf{x}-\mathbf{x}'\| \le \gamma} \phi(y f(\mathbf{x}'))$. Then $\tilde{\phi}$ is not $\mathcal{H}$-calibrated with respect to ℓ_γ, for $\mathcal{H} = \mathcal{H}_{\text{lin}}$, $\mathcal{H}_g$ with a non-decreasing and continuous function g such that $g(-\gamma) + G > 0$ and $g(\gamma) - G < 0$, and $\mathcal{H}_{\text{relu}}$ with $G > \gamma$.*

Theorem 10. *Let $\mathcal{H}$ be a hypothesis set containing 0 that is regular for adversarial calibration. If a margin-based loss ϕ is convex and non-increasing, then the surrogate loss defined by $\tilde{\phi}(f, \mathbf{x}, y) = \sup_{\mathbf{x}': \|\mathbf{x}-\mathbf{x}'\| \le \gamma} \phi(y f(\mathbf{x}'))$ is not $\mathcal{H}$-calibrated with respect to ℓ_γ.*

The theorems above provides evidence that the current practice of making networks adversarially robust via minimizing convex supremum-based surrogates may have serious deficiencies. This may also explain why in practice the adversarial accuracies that are achievable are much lower than the corresponding natural accuracies of the model (Madry et al., 2017). In general, optimizing non-calibrated or non-consistent surrogates could lead to undesirable solutions even under strong assumptions (such as the Bayes risk being zero). See Section 6, where we empirically demonstrate this in a variety of settings. By Theorem 10 and the fact that $\mathcal{H}_{\text{NN}}$ and $\mathcal{H}_{\text{all}}$ both contain 0 and are regular for adversarial calibration, we can derive the following corollary.

Corollary 11. *Let ϕ be a convex and non-increasing margin-based loss. Consider the surrogate loss defined by $\tilde{\phi}(f, \mathbf{x}, y) = \sup_{\mathbf{x}': \|\mathbf{x}-\mathbf{x}'\| \le \gamma} \phi(y f(\mathbf{x}'))$. Then $\tilde{\phi}$ is not $\mathcal{H}$-calibrated with respect to ℓ_γ, for $\mathcal{H} = \mathcal{H}_{\text{NN}}$, and $\mathcal{H} = \mathcal{H}_{\text{all}}$.*

The proofs of Theorem 7, Theorem 9 and Theorem 10 are included in Appendix E.2. The key in proving the above theorems is to analyze the calibration function $\delta_{\max}(\epsilon, \mathbf{x}, \eta)$ as defined in (6) of losses (ℓ, ℓ_γ) at $\eta = \frac{1}{2}$, $\epsilon = \frac{1}{2}$ and distinguishing $\mathbf{x}_0 \in \mathcal{X}$. Naturally, this requires us to understand the inner risk $\mathcal{C}_\ell(f, \mathbf{x}, \eta)$ that in turn depends on the worst case perturbation of a given data point according to ℓ. Our key insight (Lemma 25) is that $\delta_{\max}(\epsilon, \mathbf{x}_0, \eta)$ can be characterized by two quantities $\underline{M}(f, \mathbf{x}_0, \gamma) = \inf_{\mathbf{x}': \|\mathbf{x}_0-\mathbf{x}'\| \le \gamma} f(\mathbf{x}')$, $\overline{M}(f, \mathbf{x}_0, \gamma) = \sup_{\mathbf{x}': \|\mathbf{x}_0-\mathbf{x}'\| \le \gamma} f(\mathbf{x}')$. Requiring $\delta_{\max}\left(\frac{1}{2}, \mathbf{x}_0, \frac{1}{2}\right) > 0$ corresponds to an appropriate convex function not achieving a minimum in a set that has global optimum, thereby reaching a contradiction.

4.3 Positive results

In this section, we aim to provide alternative losses which could be calibrated with respect to ℓ_γ.

4.3.1 Characterization

In light of the above negative results, we need to consider non-convex surrogates. One possible candidate is the family of losses introduced by Bao et al. (2020a) that satisfy the property that the generic conditional ϕ-risk $\bar{\mathcal{C}}_\phi(t, \eta)$ is quasi-concave in $t \in \mathbb{R}$ for all $\eta \in [0, 1]$. Theorem 12 below is a correction to the main positive result, Theorem 11 in (Bao et al., 2020a), where we prove the theorem under the correct calibration definition.

Theorem 12. *Let a margin-based loss ϕ be bounded, continuous, non-increasing, and satisfy the property that $\bar{\mathcal{C}}_\phi(t, \eta)$ is quasi-concave in $t \in \mathbb{R}$ for all $\eta \in [0, 1]$. Assume that $\phi(-t) > \phi(t)$ for any $\gamma < t \le 1$. Then ϕ is $\mathcal{H}_{\text{lin}}$-calibrated with respect to ℓ_γ if and only if for any $\gamma < t \le 1$,*

$$\phi(\gamma) + \phi(-\gamma) > \phi(t) + \phi(-t)\,. \tag{10}$$

The proof of Theorem 12 is included in Appendix E.4, where we make use of Lemma 27 and Lemma 28, which are powerful since they apply to any symmetric hypothesis sets. These lemmas would be used for proving more general positive results, as we will show later. Note Theorem 11 in (Bao et al., 2020a) does not hold any more under the correct calibration Definition 2, since their condition $\phi(\gamma) + \phi(-\gamma) > \phi(1) + \phi(-1)$ is much weaker than (10).

The following theorem extends the above to show that under certain conditions, such surrogate losses are $\mathcal{H}$-calibrated for the class of generalized linear models with respect to the adversarial $0/1$ loss.

Theorem 13. *Let g be a non-decreasing and continuous function such that $g(1+\gamma) < G$ and $g(-1-\gamma) > -G$ for some $G \geq 0$. Let a margin-based loss ϕ be bounded, continuous, non-increasing, and satisfy the property that $\bar{\mathcal{C}}_\phi(t,\eta)$ is quasi-concave in $t \in \mathbb{R}$ for all $\eta \in [0,1]$. Assume that $\phi(g(-t) - G) > \phi(G - g(-t))$ and $g(-t) + g(t) \geq 0$ for any $0 \leq t \leq 1$. Then ϕ is $\mathcal{H}_g$-calibrated with respect to ℓ_γ if and only if for any $0 \leq t \leq 1$,*

$$\phi(G - g(-t)) + \phi(g(-t) - G) = \phi(g(t) + G) + \phi(-g(t) - G)$$
$$\text{and} \quad \min\left\{\phi(\overline{A}(t)) + \phi(-\overline{A}(t)), \phi(\underline{A}(t)) + \phi(-\underline{A}(t))\right\} > \phi(G - g(-t)) + \phi(g(-t) - G),$$

where $\overline{A}(t) = \max_{s\in[-t,t]} g(s) - g(s-\gamma)$ and $\underline{A}(t) = \min_{s\in[-t,t]} g(s) - g(s+\gamma)$.

See Appendix E.5 for the proof. The conditions in the theorem above are necessary and sufficient and thus characterize calibration for such surrogate losses. To interpret the conditions better, consider ReLU functions. In that case, the assumptions can be further simplified to get the following corollary.

Corollary 14. *Assume that $G > 1 + \gamma$. Let a margin-based loss ϕ be bounded, continuous, non-increasing, and satisfy the property that $\bar{\mathcal{C}}_\phi(t,\eta)$ is quasi-concave in $t \in \mathbb{R}$ for all $\eta \in [0,1]$. Assume that $\phi(-G) > \phi(G)$. Then ϕ is $\mathcal{H}_{\text{relu}}$-calibrated with respect to ℓ_γ if and only if for any $0 \leq t \leq 1$,*

$$\phi(G) + \phi(-G) = \phi(t+G) + \phi(-t-G) \quad \text{and} \quad \phi(\gamma) + \phi(-\gamma) > \phi(G) + \phi(-G).$$

4.3.2 Calibration

To demonstrate the applicability of Theorem 13, we consider a specific surrogate loss namely the *ρ-margin loss* $\phi_\rho(t) := \min\left\{1, \max\left\{0, 1 - \frac{t}{\rho}\right\}\right\}$, $\rho > 0$, which is a generalization of the ramp loss (see, (Mohri et al., 2018)). We also define its supremum-based counterpart as $\tilde{\phi}_\rho(f, \mathbf{x}, y) := \sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\leq\gamma} \phi_\rho(yf(\mathbf{x}'))$. Using Theorem 12, Theorem 13 and Corollary 14 in Section 4.3.1, we can conclude that the ρ-margin loss is calibrated under reasonable conditions for linear hypothesis sets and non-decreasing g-based hypothesis sets, since $\phi_\rho(t)$ is bounded, continuous, non-increasing, and satisfies $\bar{\mathcal{C}}_{\phi_\rho}(t,\eta)$ is quasi-concave in $t \in \mathbb{R}$ for all $\eta \in [0,1]$. This is stated formally below.

Theorem 15. *The surrogate ϕ_ρ is $\mathcal{H}_{\text{lin}}$-calibrated with respect to ℓ_γ if and only if $\rho > \gamma$. Given a non-decreasing and continuous function g such that $g(1+\gamma) < G$ and $g(-1-\gamma) > -G$ for some $G \geq 0$, assume that $g(-t)+g(t) \geq 0$ for any $0 \leq t \leq 1$, then ϕ_ρ is $\mathcal{H}_g$-calibrated with respect to ℓ_γ if and only if for any $0 \leq t \leq 1$, $\phi_\rho(G-g(-t)) = \phi_\rho(g(t)+G)$ and $\min\left\{\phi_\rho(\overline{A}(t)), \phi_\rho(-\underline{A}(t))\right\} > \phi_\rho(G-g(-t))$, where $\overline{A}(t) = \max_{s\in[-t,t]} g(s) - g(s-\gamma)$ and $\underline{A}(t) = \min_{s\in[-t,t]} g(s) - g(s+\gamma)$. Assume that $G > 1+\gamma$, then ϕ_ρ is $\mathcal{H}_{\text{relu}}$-calibrated with respect to ℓ_γ if and only if $G \geq \rho > \gamma$.*

Recall that in Theorem 10 we ruled out the possibility of finding $\mathcal{H}$-calibrated supremum-based convex surrogate losses with respect to the adversarial $0/1$ loss. However, we show that the supremum-based ρ-margin loss is indeed $\mathcal{H}$-calibrated, where $\mathcal{H}$ is any symmetric hypothesis set.

Theorem 16. *Let $\mathcal{H}$ be a symmetric hypothesis set, then $\tilde{\phi}_\rho$ is $\mathcal{H}$-calibrated with respect to ℓ_γ.*

The proof of Theorem 16 is included in Appendix E.4. By Theorem 16 and the fact that $\mathcal{H}_{\text{lin}}$, $\mathcal{H}_{\text{NN}}$ and $\mathcal{H}_{\text{all}}$ are all symmetric, we derive the following.

Corollary 17. *$\tilde{\phi}_\rho$ is $\mathcal{H}$-calibrated with respect to ℓ_γ, for $\mathcal{H} = \mathcal{H}_{\text{lin}}$, $\mathcal{H}_{\text{NN}}$, and $\mathcal{H}_{\text{all}}$.*

The results of this section suggest that the ρ-margin loss and supremum-based ρ-margin loss may be good surrogates for the adversarial $0/1$ loss. However, calibration, in general, is not equivalent to consistency, our eventual goal. In the next section, we study conditions under which we can expect these surrogates losses to be $\mathcal{H}$-consistent as well.

5 $\mathcal{H}$-Consistency

In this section, we study the $\mathcal{H}$-consistency of surrogate loss functions. The results of the previous section suggest that convex losses or supremum-based convex losses would not be $\mathcal{H}$-consistent. However, $\mathcal{H}$-calibrated losses, such as the ρ-margin loss and supremum-based ρ-margin loss present an intriguing possibility. Bao et al. (2020a) made a claim that since the losses they proposed are $\mathcal{H}_{\text{lin}}$-calibrated they are also $\mathcal{H}_{\text{lin}}$-consistent. We first present a result that implies this claim is incorrect. In fact, our result stated below shows that without assumptions on the data distribution, no continuous margin based loss or a supremum-based continuous surrogate could be $\mathcal{H}_{\text{lin}}$-consistent.

5.1 Negative results

Theorem 18. *No continuous margin-based loss function ϕ is $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ. Furthermore, for any continuous and non-increasing margin-based loss ϕ, surrogates of the form*

$$\tilde{\phi}(f, \mathbf{x}, y) = \sup_{\mathbf{x}' : \|\mathbf{x}-\mathbf{x}'\| \leq \gamma} \phi(y f(\mathbf{x}'))$$

are not $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ.

The above theorem is proven in Appendix E.6. In particular, Theorem 18 contradicts the $\mathcal{H}$-consistency claim of Bao et al. (2020a) for their proposed losses when $\mathcal{H}$ is the family of linear functions. Furthermore, the theorem rules out $\mathcal{H}$-consistency of supremum-based surrogates.

5.2 Positive results

In this section, we investigate the nature of the assumptions on the data distributions that may lead to $\mathcal{H}$-consistency of surrogate losses. We take inspiration from the work of Long and Servedio (2013) and Zhang and Agarwal (2020) who study $\mathcal{H}$-consistency for the standard $0/1$ loss. These studies establish consistency under a realizability assumption on the data distribution stated below that requires the Bayes $(\ell_0, \mathcal{H})$-risk to be zero.

Definition 19 ($\mathcal{H}$-realizability)**.** *A distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$ is $\mathcal{H}$-realizable if it labels points according to a deterministic model in $\mathcal{H}$, i.e., if $\exists f \in \mathcal{H}$ such that $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{P}}(\operatorname{sgn}(f(\mathbf{x})) = y) = 1$.*

As with $\mathcal{H}$-realizability, we will assume that, under the data distribution, the Bayes $(\ell_\gamma, \mathcal{H})$-risk is zero. We show that the $\mathcal{H}$-calibrated losses studied in previous sections are $\mathcal{H}$-consistent under natural conditions along with the realizability assumption.

5.2.1 Non-supremum-based surrogates

Theorem 20. *Let $\mathcal{P}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{H}$ a hypothesis set for which $\mathcal{R}^*_{\ell_\gamma,\mathcal{H}} = 0$. Let ϕ be a margin-based loss. If for $\eta \geq 0$, there exists $f^* \in \mathcal{H} \subset \mathcal{H}_{\text{all}}$ such that $\mathcal{R}_\phi(f^*) \leq \mathcal{R}^*_{\phi,\mathcal{H}_{\text{all}}} + \eta < +\infty$ and ϕ is $\mathcal{H}$-calibrated with respect to ℓ_γ, then for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $f \in \mathcal{H}$,*

$$\mathcal{R}_\phi(f) + \eta < \mathcal{R}^*_{\phi,\mathcal{H}} + \delta \implies \mathcal{R}_{\ell_\gamma}(f) < \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} + \epsilon.$$

For the family of linear models, some convex losses may also be $\mathcal{H}_{\text{lin}}$-consistent verifying the conditions ($\eta = 0$) in Theorem 20. However, $\mathcal{H}_{\text{lin}}$-calibrated losses can be $\mathcal{H}_{\text{lin}}$-consistent under more benign assumptions, where the realizability condition $\mathcal{R}^*_{\ell_\gamma,\mathcal{H}_{\text{lin}}} = 0$ can be further relaxed.

Theorem 21. *Let $\mathcal{P}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. Assume that there exists $g^* \in \mathcal{H}_{\text{lin}}$ such that $\mathcal{R}_{\ell_\gamma}(g^*) = \mathcal{R}^*_{\ell_\gamma,\mathcal{H}_{\text{all}}}$. Let ϕ be a margin-based loss. If for $\eta \geq 0$, there exists $f^* \in \mathcal{H}_{\text{lin}} \subset \mathcal{H}_{\text{all}}$ such that $\mathcal{R}_\phi(f^*) \leq \mathcal{R}^*_{\phi,\mathcal{H}_{\text{all}}} + \eta < +\infty$ and ϕ is $\mathcal{H}_{\text{lin}}$-calibrated with respect to ℓ_γ, then for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $f \in \mathcal{H}_{\text{lin}}$ we have*

$$\mathcal{R}_\phi(f) + \eta < \mathcal{R}^*_{\phi,\mathcal{H}_{\text{lin}}} + \delta \implies \mathcal{R}_{\ell_\gamma}(f) < \mathcal{R}^*_{\ell_\gamma,\mathcal{H}_{\text{lin}}} + \epsilon.$$

The proofs of Theorem 20 and Theorem 21 are presented in Appendix E.7. Using Theorem 15 in Section 4.3.2 and theorems above, we immediately conclude that the calibrated ρ-margin loss is consistent with respect to ℓ_γ for all distributions that satisfy our realizability assumptions.

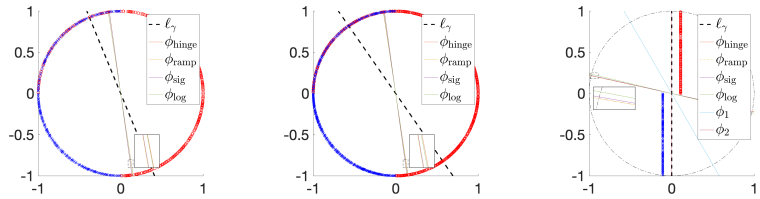

Figure 2: Left: Unit Circle with 1,000 and 2,000 samples. Right: Segment with 5,00 samples.

Theorem 22. *If $\rho > \gamma$, then ϕ_ρ is $\mathcal{H}_{\text{lin}}$-consistent wrt ℓ_γ for all distributions such that there exists $g^* \in \mathcal{H}_{\text{lin}}$ with $\mathcal{R}_{\ell_\gamma}(g^*) = \mathcal{R}^*_{\ell_\gamma,\mathcal{H}_{\text{all}}}$ and there exists $f^* \in \mathcal{H}_{\text{lin}}$ such that $\mathcal{R}_{\phi_\rho}(f^*) = \mathcal{R}^*_{\phi_\rho,\mathcal{H}_{\text{all}}}$. If g verifies the calibration condition in Theorem 15, then ϕ_ρ is $\mathcal{H}_g$-consistent wrt ℓ_γ for all distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$ that satisfies $\mathcal{R}^*_{\ell_\gamma,\mathcal{H}_g} = 0$ and there exists $f^* \in \mathcal{H}_g$ such that $\mathcal{R}_{\phi_\rho}(f^*) = \mathcal{R}^*_{\phi_\rho,\mathcal{H}_{\text{all}}}$. If $G > 1 + \gamma$ and $G \geq \rho > \gamma$, then ϕ_ρ is $\mathcal{H}_{\text{relu}}$-consistent wrt ℓ_γ for all distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$ that satisfies $\mathcal{R}^*_{\ell_\gamma,\mathcal{H}_{\text{relu}}} = 0$ and there exists $f^* \in \mathcal{H}_{\text{relu}}$ such that $\mathcal{R}_{\phi_\rho}(f^*) = \mathcal{R}^*_{\phi_\rho,\mathcal{H}_{\text{all}}}$.*

5.2.2 Supremum-based surrogates

We can also extend the above to obtain $\mathcal{H}$-consistency of supremum-based convex surrogates. However we need the stronger condition that $\mathcal{R}_\phi$ is minimized exactly inside $\mathcal{H}$.

Theorem 23. *Given a distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$ and a hypothesis set $\mathcal{H}$ such that $\mathcal{R}^*_{\ell_\gamma,\mathcal{H}} = 0$. Let ϕ be a non-increasing margin-based loss. If there exists $f^* \in \mathcal{H} \subset \mathcal{H}_{\text{all}}$ such that $\mathcal{R}_\phi(f^*) = \mathcal{R}^*_{\phi,\mathcal{H}_{\text{all}}} < +\infty$ and $\tilde{\phi}(f, \mathbf{x}, y) = \sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\leq\gamma} \phi(yf(\mathbf{x}'))$ is $\mathcal{H}$-calibrated with respect to ℓ_γ, then for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $f \in \mathcal{H}$ we have*

$$\mathcal{R}_{\tilde{\phi}}(f) < \mathcal{R}^*_{\tilde{\phi},\mathcal{H}} + \delta \implies \mathcal{R}_{\ell_\gamma}(f) < \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} + \epsilon.$$

The proof of Theorem 23 is presented in Appendix E.7. Again, when combined with Theorem 16 in Section 4.3.2 we conclude that the $\mathcal{H}$-calibrated supremum-based ρ-margin loss is also $\mathcal{H}$-consistent with respect to ℓ_γ for all distributions that satisfy our realizability assumptions.

Theorem 24. *Let $\mathcal{H}$ be a symmetric hypothesis set, then $\tilde{\phi}_\rho$ is $\mathcal{H}$-consistent with respect to ℓ_γ for all distributions $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$ that satisfy: $\mathcal{R}^*_{\ell_\gamma,\mathcal{H}} = 0$ and there exists $f^* \in \mathcal{H}$ such that $\mathcal{R}_{\phi_\rho}(f^*) = \mathcal{R}^*_{\phi_\rho,\mathcal{H}_{\text{all}}} < +\infty$.*

6 Experiments

Here, we present experiments on simulated data to support our theoretical findings. The goal is two-fold. First, we empirically demonstrate that indeed $\mathcal{H}$-calibrated surrogates in (Bao et al., 2020a) may not be $\mathcal{H}$-consistent unless assumptions on the data distribution are made, even when $\mathcal{H}$ is the class of linear functions. This is consistent with our negative result in Theorem 18 and provides an empirical counterexample to the claim made in (Bao et al., 2020a). Second, we study the necessity of the realizability assumptions we adopted in Section 5.2 to establish $\mathcal{H}$-consistency of surrogates satisfying the conditions in Theorem 12.

We generate data points $\mathbf{x} \in \mathbb{R}^2$ on the unit circle and consider $\mathcal{H}$ to be linear models $\mathcal{H}_{\text{lin}}$. We denote $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$, $\mathbf{w} = (\cos(t), \sin(t))^\top, t \in [0, 2\pi)$, $f \in \mathcal{H}_{\text{lin}}$. All risks are approximated by their empirical counterparts computed over 10^7 i.i.d. samples. To demonstrate the need for some assumptions for $\mathcal{H}$-consistency, we construct a scenario we call the **Unit Circle** case. We consider four surrogates: ϕ_{hinge}, ϕ_{ramp}, ϕ_{sig} and $\phi_{\log}$ defined in Appendix C.1. In general, we refer all of these surrogates as ϕ_{sur}. We generate data points $\mathbf{x}$ from the uniform distribution on the unit circle. Define $\mathbf{x}$ as $\mathbf{x} = (\cos(\theta), \sin(\theta))^\top$, $\theta \in [0, 2\pi)$. Set the label of a point $\mathbf{x}$ as follows: if $\theta \in \left(\frac{\pi}{2}, \pi\right)$, then $y = -1$ with probability $\frac{3}{4}$ and $y = 1$ with probability $\frac{1}{4}$; if $\theta \in \left(0, \frac{\pi}{2}\right)$ or $\left(\frac{3\pi}{2}, 2\pi\right)$, then $y = 1$; if $\theta \in \left(\pi, \frac{3\pi}{2}\right)$, then $y = -1$. Set $\gamma = \frac{\sqrt{2}}{2}$. In this case, the Bayes $(\ell_\gamma, \mathcal{H}_{\text{lin}})$-risk is $\mathcal{R}^*_{\ell_\gamma,\mathcal{H}_{\text{lin}}} \approx 0.5000 \neq 0$ and is achieved by $w_{\ell_\gamma} = (\cos(\theta), \sin(\theta))^\top$ with $\theta \approx 0.7855$. The results obtained by optimizing the different surrogate losses are reported in Table 1(a) and the plots for

Table 1: (a) Unit Circle; (b) Segments.

ϕ_{sur}	$\mathcal{R}_{\ell_\gamma}(f^*)$	$\theta_{\phi_{\text{sur}}}$	$\mathcal{H}_{\text{lin}}$-cal.	$\mathcal{H}_{\text{lin}}$-cons.
ϕ_{hinge}	0.5257	0.1420	✗	✗
ϕ_{ramp}	0.5263	0.1288	✓	✗
ϕ_{sig}	0.5261	0.1320	✓	✗
ϕ_{log}	0.5258	0.1414	✗	✗

(a)

ϕ_{sur}	$\mathcal{R}_{\ell_\gamma}(f^*)$	$\mathcal{R}_{\phi_{\text{sur}}}(f^*)$	$\theta_{\phi_{\text{sur}}}$	$\mathcal{H}_{\text{lin}}$-cal.	$\mathcal{H}_{\text{lin}}$-cons.
ϕ_{hinge}	0.0781	0.6907	1.3548	✗	✗
ϕ_{ramp}	0.0781	0.3454	1.3548	✓	✗
ϕ_{sig}	0.0777	0.4247	1.3498	✓	✗
ϕ_{log}	0.0763	0.8078	1.3341	✗	✗
ϕ_1	0.0111	0	$\frac{\pi}{6}$	✗	✗
ϕ_2	0	0	0	✓	✓

(b)

1,000 samples and 2,000 samples are shown in Figure 2. Table 1(a) shows that neither calibrated nor non-calibrated (convex) surrogates are $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ for this distribution. Figure 2 shows that the classifiers obtained by optimizing the four surrogates are almost the same but deviate a lot from the optimal Bayes classifier for ℓ_γ. This shows that indeed calibrated surrogates may not be consistent and contradicts Figure 12 of (Bao et al., 2020a). The discrepancy results from an incorrect calculation of the adversarial Bayes risk in (Bao et al., 2020a).

Next, we justify the realizability assumptions made in Section 5.2 for obtaining $\mathcal{H}$-consistency of surrogate losses. To do so, we design a scenario that we call the **Segments** case. Here, we consider six surrogates, the four studied above and two more surrogates ϕ_1 and ϕ_2 defined in Appendix C.1. The loss ϕ_1 is a convex loss and ϕ_2 is the ρ-margin ramp loss for some $\rho > \gamma$. In general, we refer to all of these surrogates as ϕ_{sur}. We show in Appendix C.2 that ϕ_{hinge}, ϕ_{log} and ϕ_1 are not $\mathcal{H}_{\text{lin}}$-calibrated while ϕ_{ramp}, ϕ_{sig} and ϕ_2 are $\mathcal{H}_{\text{lin}}$-calibrated with respect to ℓ_γ.

Let $I_{\hat{\gamma}} = \sqrt{1-\hat{\gamma}^2}$ and consider: $\mathbb{P}(Y=1) = \mathbb{P}(Y=-1) = \frac{1}{2}$, and $X \mid Y = 1$ is the uniform distribution on the line segment $\{(\hat{\gamma}, z) \mid z \in [0, I_{\hat{\gamma}}]\}$ and $X \mid Y = -1$ is the uniform distribution on the line segment $\{(-\hat{\gamma}, z) \mid z \in [-I_{\hat{\gamma}}, 0]\}$ where $\hat{\gamma} = \gamma + \frac{1-\gamma}{100} = \frac{1+99\gamma}{100}$, $\gamma \in (0,1)$. We choose $\gamma = 0.1$ and set $\mathbf{w}^* = (1,0)^\top$. It is easy to check that $\mathbf{w}^*$ achieves the Bayes (ℓ_γ, $\mathcal{H}_{\text{lin}}$)-risk $\mathcal{R}^*_{\ell_\gamma,\mathcal{H}_{\text{lin}}} = 0$. The results for the six different surrogate losses are indicated in Table 1(b) and the plot for 5,00 samples are shown in Figure 2. For ϕ_{hinge}, ϕ_{ramp}, ϕ_{sig} and ϕ_{log}, the Bayes (ϕ_{sur}, $\mathcal{H}_{\text{lin}}$)-risk $\mathcal{R}^*_{\phi_{\text{sur}},\mathcal{H}_{\text{lin}}} \neq 0$. Table 1(b) shows that they are not $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ. For ϕ_1 and ϕ_2, the Bayes (ϕ_{sur}, $\mathcal{H}_{\text{lin}}$)-risk $\mathcal{R}^*_{\phi_{\text{sur}},\mathcal{H}_{\text{lin}}} = 0$. Table 1(b) shows that ϕ_1 is not $\mathcal{H}_{\text{lin}}$-consistent (recall that ϕ_1 is not calibrated) but ϕ_2 is $\mathcal{H}_{\text{lin}}$-consistent for this distribution. Hence, even when $\mathcal{R}^*_{\ell_\gamma,\mathcal{H}_{\text{lin}}} = 0$, unless a condition is also imposed on $\mathcal{R}^*_{\phi_{\text{sur}},\mathcal{H}_{\text{lin}}}$, one cannot expect consistency, thereby justifying our realizability assumption. Note that $\mathcal{R}^*_{\phi_{\text{sur}},\mathcal{H}_{\text{lin}}} = \mathcal{R}^*_{\ell_\gamma,\mathcal{H}_{\text{lin}}} = 0$ is a special case verifying the conditions of Theorem 20 for $\eta = 0$. For this distribution, ϕ_{ramp} is not $\mathcal{H}_{\text{lin}}$-consistent while ϕ_2 is $\mathcal{H}_{\text{lin}}$-consistent, although both are $\mathcal{H}_{\text{lin}}$-calibrated. We compare them in Figure 3, showing that minimizing $\mathcal{H}_{\text{lin}}$-consistent surrogate ϕ_2 minimizes the adversarial generalization error for large sample sizes but the same does not hold for non $\mathcal{H}_{\text{lin}}$-consistent surrogate ϕ_{ramp}.

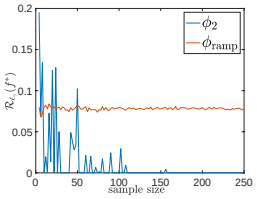

Figure 3: Adv. true risk of consistent and calibrated inconsistent losses vs. sample size.

7 Conclusion

We presented a detailed study of calibration and consistency for adversarial robustness. These results can help guide the design of algorithms for learning robust predictors, an increasingly important problem in applications. Our theoretical results show in particular that many of the surrogate losses typically used in practice do not benefit from any guarantee. Our empirical results further illustrate that in the context of a general example. Our results also show that some of the calibration results presented in previous work do not bear any significance, since we prove that in fact they do not guarantee consistency. Instead, we give a series of positive calibration and consistency results for several families of surrogate functions, under some realizability assumptions.

Acknowledgements

This work was partly funded by NSF CCF-1535987 and NSF IIS-1618662.

References

Attias, I., Kontorovich, A., and Mansour, Y. (2018). Improved generalization bounds for robust learning. *arXiv preprint arXiv:1810.02180*.

Awasthi, P., Dutta, A., and Vijayaraghavan, A. (2019). On robustness to adversarial examples and polynomial optimization. In *Advances in Neural Information Processing Systems*, pages 13737–13747.

Awasthi, P., Frank, N., and Mohri, M. (2020). Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pages 431–441.

Bao, H., Scott, C., and Sugiyama, M. (2020a). Calibrated surrogate losses for adversarially robust classification. In *Conference on Learning Theory*, pages 408–451.

Bao, H., Scott, C., and Sugiyama, M. (2020b). Corrigendum to: Calibrated surrogate losses for adversarially robust classification. *arXiv preprint arXiv:2005.13748*.

Bartlett, P. L., Bubeck, S., and Cherapanamjeri, Y. (2021). Adversarial examples in multi-layer random relu networks. *arXiv preprint arXiv:2106.12611*.

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

Boyd, S. P. and Vandenberghe, L. (2014). *Convex Optimization*. Cambridge University Press.

Bubeck, S., Lee, Y. T., Price, E., and Razenshteyn, I. (2018a). Adversarial examples from cryptographic pseudo-random generators. *arXiv preprint arXiv:1811.06418*.

Bubeck, S., Price, E., and Razenshteyn, I. (2018b). Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*.

Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 39–57.

Carmon, Y., Raghunathan, A., Schmidt, L., Liang, P., and Duchi, J. C. (2019). Unlabeled data improves adversarial robustness. *arXiv preprint arXiv:1905.13736*.

Cullina, D., Bhagoji, A. N., and Mittal, P. (2018). PAC-learning in the presence of evasion adversaries. *arXiv preprint arXiv:1806.01471*.

Diakonikolas, I., Kane, D. M., and Manurangsi, P. (2020). The complexity of adversarially robust proper learning of halfspaces with agnostic noise. *arXiv preprint arXiv:2007.15220*.

Feige, U., Mansour, Y., and Schapire, R. (2015). Learning and inference in the presence of corrupted inputs. In *Conference on Learning Theory*, pages 637–657.

Feige, U., Mansour, Y., and Schapire, R. E. (2018). Robust inference for multiclass classification. In *Algorithmic Learning Theory*, pages 368–386.

Gao, W. and Zhou, Z.-H. (2015). On the consistency of auc pairwise optimization. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Khim, J. and Loh, P.-L. (2018). Adversarial risk bounds for binary classification via function transformation. *arXiv preprint arXiv:1810.09519*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105.

Long, P. and Servedio, R. (2013). Consistency versus realizable H-consistency for multiclass classification. In *International Conference on Machine Learning*, pages 801–809.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning*. MIT Press, second edition.

Montasser, O., Hanneke, S., and Srebro, N. (2019). Vc classes are adversarially robustly learnable, but only improperly. *arXiv preprint arXiv:1902.04217*.

Montasser, O., Hanneke, S., and Srebro, N. (2020). Reducing adversarially robust learning to non-robust pac learning. *arXiv preprint arXiv:2010.12039*.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. (2019). Adversarial training for free! In *Advances in Neural Information Processing Systems*, pages 3353–3364.

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

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tewari, A. and Bartlett, P. L. (2007). On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(36):1007–1025.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. (2018). Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*.

Uematsu, K. and Lee, Y. (2011). On theoretically optimal ranking functions in bipartite ranking. *Department of Statistics, The Ohio State University, Tech. Rep*, 863.

Wong, E., Rice, L., and Kolter, J. Z. (2020). Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*.

Yin, D., Ramchandran, K., and Bartlett, P. L. (2019). Rademacher complexity for adversarially robust generalization. In *International Conference of Machine Learning*, pages 7085–7094.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. (2019). Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*.

Zhang, M. and Agarwal, S. (2020). Bayes consistency vs. h-consistency: The interplay between surrogate loss functions and the scoring function class. In *Advances in Neural Information Processing Systems*, pages 16927–16936.

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

Contents of Appendix

A Related Work **14**

B Further comments on (Bao et al., 2020a) **15**

C Details of Experiments **16**

C.1 Definition of Surrogates 16

C.2 Theoretical Analysis of Surrogates 16

D Future work **16**

E Deferred Proofs **17**

E.1 Proof of Theorem 6 17

E.2 Proof of Theorem 7, Theorem 9 and Theorem 10 17

E.3 Properties of generic conditional ϕ-risk 21

E.4 Proof of Theorem 12 and Theorem 16 23

E.5 Proof of Theorem 13 and Corollary 14 29

E.6 Proof of Theorem 18 34

E.7 Proof of Theorem 20, Theorem 21 and Theorem 23 38

A Related Work

The notions of calibration and consistency with respect to the $0/1$ loss have been widely studied in the statistical learning theory literature to analyze the properties of surrogate losses (Zhang, 2004; Bartlett et al., 2006). Bartlett et al. (2006) showed that margin-based convex surrogates, that is mappings of the form $(f, \mathbf{x}, y) \mapsto \phi(yf(\mathbf{x}))$, where f is a real-valued predictor and $\phi\colon \mathbb{R} \to \mathbb{R}_+$ a function differentiable at 0 with $\phi'(0) < 0$, are calibrated with respect to the class of all measurable functions. Extensions of calibration and consistency to multi-class settings have also been studied (Tewari and Bartlett, 2007). In the special case of the $0/1$ loss and margin-based convex surrogates, calibration immediately implies consistency for the class of all measurable functions. One can then even derive quantitative bounds relating the excess ϕ-risk to the excess $0/1$ loss of any function f (Zhang, 2004; Bartlett et al., 2006).

The case of adversarial loss is more complex. This is because, in particular, the loss of a predictor f at point $\mathbf{x}$ does not just depend on its value $f(\mathbf{x})$ at that point but also on its values in a neighborhood of $\mathbf{x}$. Steinwart (2007) proposed a general framework to study and characterize calibration and consistency, in particular via a *calibration function*. He also defined a *minimizability* condition under which calibration implies consistency. But, while *minimizability* holds for the $0/1$ loss and margin-based convex surrogates over the class of all measurable functions, the condition does not hold in general for the adversarial loss. Our work borrows tools from the work of Steinwart (2007). However, to establish $\mathcal{H}$-*consistency* in the context of the adversarial loss, additional insights are needed and often stronger assumptions on the data distribution are required. These assumptions are captured in the notion of *realizable* $\mathcal{H}$-*consistency* that requires that the optimal risk of both the $0/1$ loss and the surrogate loss being achieved inside the class $\mathcal{H}$. Our positive results for $\mathcal{H}$-*consistency* rely on similar but weaker assumptions. Long and Servedio (2013) gave examples of surrogate losses that are not $\mathcal{H}$-*consistent* when $\mathcal{H}$ is the class of all measurable functions but satisfy *realizable* $\mathcal{H}$-*consistency* when $\mathcal{H}$ is the class of linear functions. Zhang and Agarwal (2020) studied the notion of *improper realizable* $\mathcal{H}$-*consistency* of linear classes where the surrogate ϕ can be optimized over a larger class such as that of piecewise linear functions. The relation between calibration and its implication for consistency has also been explored in ranking problems (Uematsu and Lee, 2011; Gao and Zhou, 2015). In particular, these works show that calibrated surrogate losses for classification problems are not consistent for optimizing ranking losses such as AUC. Hence in these works calibration not implying consistency stems from the mismatch between using a calibrated surrogate for a different loss (classification loss) and applying it for a different purpose. However, in the context of our work the situation is more subtle. Even a calibrated surrogate (with respect to the adversarial 0-1 loss) may not be consistent in general.

These notions of calibration and consistency are relatively unexplored for the adversarial $0/1$ loss. Bao et al. (2020a) recently initiated the study of these notions for the adversarial loss. We give a more detailed discussion regarding that work in Appendix B.

There has also been recent works on theoretically understanding different aspects of adversarial robustness. Tsipras et al. (2018) give constructions under which every classifier with small $0/1$ loss has a large adversarial $0/1$ loss thereby pointing to a tension between the two criteria. This tradeoff has been explored in subsequent work (Zhang et al., 2019; Carmon et al., 2019). Bubeck et al. (2018b), Bubeck et al. (2018a) and Awasthi et al. (2019) quantify computational bottlenecks in learning classifiers with small adversarial loss. The recent work of Bartlett et al. (2021) shows that for randomly innitialized neural networks, low perturbation magnitude adversarial examples exist, with high probability, nearby every data point. There has also been a line of work analyzing the sample complexity of optimizing adversarial surrogate losses using notions of VC-dimension and Rademacher complexity appropriately extended to the adversarial case (Yin et al., 2019; Khim and Loh, 2018; Awasthi et al., 2020; Montasser et al., 2019; Cullina et al., 2018). Another recent line of concerns constructing computationally efficient adversarially robust classifiers for linear classifiers (Diakonikolas et al., 2020) and exploring the connections between adversarial learning and agnostic PAC learning (Montasser et al., 2020). Finally, an alternative adversarial setting has been theoretically studied in (Feige et al., 2015, 2018; Attias et al., 2018), where the adversary has at his disposal a finite set of perturbations for each input.

B Further comments on (Bao et al., 2020a)

Our study is somewhat inspired by and benefits from the prior work of Bao et al. (2020a). However, there are some issues worth pointing out.

Bao et al. (2020a) analyzed $\mathcal{H}$-*calibration* for adversarially robust classification in the special case where $\mathcal{H}$ is the family of linear models. In particular, the authors studied the γ-margin loss defined by $\phi_\gamma(f, \mathbf{x}, y) = \mathbb{1}_{yf(\mathbf{x})\leq\gamma}$, which only coincides with the *adversarial* $0/1$ *loss* in the case of linear hypotheses. The authors showed that, when $\mathcal{H}$ is linear, convex margin-based losses are not $\mathcal{H}$-calibrated and proposed a class of $\mathcal{H}$-calibrated surrogates modulo subtle definition differences.

However, several clarifications are needed. First, the definition of calibration adopted by the authors does not coincide with the standard definition (Steinwart, 2007) in the case of the linear models they study, although it does match that definition in the case of the family of all measurable functions (Steinwart, 2007, Section 3.2): the minimal inner risk in the definition should be defined for a fixed $\mathbf{x}$ and the infimum should be over f, instead of an infimum over both f and $\mathbf{x}$. Second, and this is crucial, $\mathcal{H}$-*calibration*, in general, does not imply $\mathcal{H}$-*consistency*, unless a property such as *minimizability* holds (Steinwart, 2007, Theorem 2.8). *Minimizability* holds for standard binary classification and the family of all measurable functions (Steinwart, 2007, Theorem 3.2). However, it does not hold, in general, for adversarially robust classification and a specific hypothesis set $\mathcal{H}$. As a result, the claim made by the authors that the calibrated surrogates they propose are $\mathcal{H}$-*consistent* is proved to be incorrect as a by-product of our results, which further suggests that the adversarial setting is more complex and requires a more delicate analysis. Third, the authors analyzed $\mathcal{H}$-*calibration* with respect to the γ-margin loss $\phi_\gamma\colon \mathbf{x} \mapsto \mathbb{1}_{yf(\mathbf{x})\leq\gamma}$ in the case where $\mathcal{H} \supset [-1, 1]$ is the general family of functions. However, as already mentioned, ϕ_γ coincides with the *adversarial* $0/1$ *loss* ℓ_γ, introduced in (2), only in the special case where $\mathcal{H}$ is the family of linear models and adversarial perturbations measured in ℓ_2 norm are considered (Bao et al., 2020a, Proposition 1). Fourth, for the negative results, the authors presented a calibration analysis of convex margin-based losses, which is natural for standard $0/1$ loss, but the current practice in designing algorithms for the *adversarial* $0/1$ *loss* typically consists of using convex *supremum-based* surrogates (Madry et al., 2017; Wong et al., 2020; Shafahi et al., 2019). Finally, the experiments in (Bao et al., 2020a) are problematic. Equation (12) in Appendix D.1., which they used to compute the Bayes risks, is wrong since $\mathcal{R}_\ell(f^*) = \mathcal{R}^*_{\ell,\mathcal{H}_{\mathrm{lin}}}$ cannot imply $\mathcal{C}_\ell(f^*, \mathbf{x}, \eta) = \mathcal{C}^*_{\ell,\mathcal{H}_{\mathrm{lin}}}(\mathbf{x}, \eta)$ in general. Let us point out that some of the issues just brought up have been discussed in a corrigendum following comments and questions we addressed to the authors (Bao et al., 2020b), but some others do not seem to have been fully addressed there.

In contrast with that prior work, instead of studying the γ-margin loss and the specific family of linear models though, we directly study the adversarial $0/1$ loss and general hypothesis sets. In particular, our calibration results fix the results presented in (Bao et al., 2020a) for linear hypothesis sets by using the correct definition, and significantly generalize them to the nonlinear hypothesis sets. For any non-decreasing and continuous g-based hypothesis set, including the ReLU-based hypothesis set, we also study the type of margin-based surrogates losses whose *generic conditional risk* has quasi-concave property and further establish several useful properties of such losses building on the work of Bao et al. (2020a). Moreover, our results imply that the convex *supremum-based* surrogates commonly used in practice for optimizing the adversarial loss are not $\mathcal{H}$-*consistent* and that minimizing such losses may not lead to a more favorable adversarial loss. Instead, we suggest alternative surrogate losses that we prove are $\mathcal{H}$-*consistent* for any symmetric hypothesis set including the multi-layer neural networks, the supremum-based ρ-*margin loss*, which can be useful to the design of effective algorithms. To show the claim in (Bao et al., 2020a) that the $\mathcal{H}$-*calibrated* surrogates they propose are $\mathcal{H}$-*consistent* is inaccurate, we carefully design a distribution on the unit disk, where any continuous surrogate can be led astray to a classifier that is far from the optimal classifier of the *adversarial* $0/1$ *loss*. This counterexample in fact rules out the $\mathcal{H}$-*consistency* of a larger class of surrogates, unless assumptions on the data distribution are imposed. In contrast, we give natural $\mathcal{H}$-*consistency* guarantees taking inspiration from the work of Long and Servedio (2013) and Zhang and Agarwal (2020). With the correct approximation of Bayes risks, our experiments further empirically demonstrate that indeed the $\mathcal{H}$-*calibrated* losses proposed in (Bao et al., 2020a) are not $\mathcal{H}$-*consistent* and justify our proposed conditions for $\mathcal{H}$-*consistency*.

C Details of Experiments

As shown by Bao et al. (2020a), the adversarial $0/1$ loss $\ell_\gamma = \mathbb{1}_{yf(\mathbf{x})\leq\gamma}$ when $f \in \mathcal{H}_{\mathrm{lin}}$. In this experiment, we approximate $\mathcal{R}^*_{\ell_\gamma,\mathcal{H}_{\mathrm{lin}}}$ over a grid. For surrogate losses, we approximate $f^* = \mathrm{argmin}_{f\in\mathcal{H}_{\mathrm{lin}}} \mathcal{R}_{\phi_{\mathrm{sur}}}(f)$ over the same grid. The experiments were run on a standard laptop with a 2.4 GHz Quad-Core Intel Core i5 Processor.

C.1 Definition of Surrogates

- Shifted Hinge loss: $\phi_{\mathrm{hinge}}(t) = \max\{0, 1 - t + 0.2\}$;
- Shifted Ramp loss: $\phi_{\mathrm{ramp}}(t) = \min\left\{1, \max\left\{0, \frac{1-t+0.92}{2}\right\}\right\}$;
- Shifted Sigmoid loss: $\phi_{\mathrm{sig}}(t) = \frac{1}{1+e^{t-0.2}}$;
- Shifted Logistic loss: $\phi_{\log}(t) = \log_2(1 + e^{-t+0.2})$;
- One convex loss: $\phi_1(t) = \max\left\{0, \frac{\gamma}{2} - t\right\}$; and
- ρ-margin loss: $\phi_2(t) = \min\left\{1, \max\left\{0, 1 - \frac{t}{\hat{\gamma}}\right\}\right\}$ for $\hat{\gamma} > \gamma$.

C.2 Theoretical Analysis of Surrogates

ϕ_{hinge}, $\phi_{\log}$, and ϕ_1 are convex surrogates and thus are not $\mathcal{H}_{\mathrm{lin}}$-calibrated with respect to ℓ_γ by Corollary 8. However, ϕ_{ramp}, ϕ_{sig} and ϕ_2 are $\mathcal{H}_{\mathrm{lin}}$-calibrated with respect to ℓ_γ since they verify the conditions in Theorem 12.

Note that $\mathbb{E}_{(X,Y)}[\phi_2(Y\mathbf{w}\cdot X)] = 0$ if and only if $w = (1,0)^\top$. Therefore, ϕ_2 is $\mathcal{H}_{\mathrm{lin}}$-consistent for the distribution **Segments**. However, for $w = (1,0)^\top$ or $w = (\cos(\theta), \sin(\theta))^\top$ where $\theta = \frac{\pi}{6}$, we have $\mathbb{E}_{(X,Y)}[\phi_1(Y\mathbf{w}\cdot X)] = 0$. Note when $\mathbf{w} = (\cos(\theta), \sin(\theta))^\top$ where $\theta = \frac{\pi}{6}$, we have $\mathbb{E}_{(X,Y)}[\ell_\gamma(Y\mathbf{w}\cdot X)] \neq 0$. Therefore, ϕ_1 is not $\mathcal{H}_{\mathrm{lin}}$-consistent for the distribution **Segments**.

D Future work

While our calibration and consistency results are very general and apply to several widely used hypothesis sets, other hypothesis sets might require a further study. Nevertheless, we believe that our proof techniques should provide a sufficient tool for the analysis of such other cases.

E Deferred Proofs

For convenience, let $\Delta\mathcal{C}_{\ell,\mathcal{H}}(f,\mathbf{x},\eta) := \mathcal{C}_{\ell}(f,\mathbf{x},\eta) - \mathcal{C}^*_{\ell,\mathcal{H}}(\mathbf{x},\eta)$, $\underline{M}(f,\mathbf{x},\gamma) := \inf_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} f(\mathbf{x}')$ and $\overline{M}(f,\mathbf{x},\gamma) := -\inf_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} -f(\mathbf{x}') = \sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} f(\mathbf{x}')$.

E.1 Proof of Theorem 6

Theorem 6. *Let $\mathcal{H}$ be a symmetric hypothesis set. If $\mathcal{H}$ is not regular for adversarial calibration, then any surrogate loss ℓ is $\mathcal{H}$-calibrated with respect to ℓ_γ.*

Proof. Since $\mathcal{H}$ is symmetric, for any $\mathbf{x}\in\mathcal{X}$, $f\in\mathcal{H}$, $\inf_{\|\mathbf{x}'-\mathbf{x}\|\le\gamma} f(\mathbf{x}') \le 0 \le \sup_{\|\mathbf{x}'-\mathbf{x}\|\le\gamma} f(\mathbf{x}')$. Thus by the definition of inner risk (4) and adversarial 0-1 loss ℓ_γ (9), for any $\mathbf{x}\in\mathcal{X}$, $f\in\mathcal{H}$,

$$\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f,\mathbf{x},\eta) = \eta\mathbb{1}_{\inf_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} f(\mathbf{x}')\le 0} + (1-\eta)\mathbb{1}_{\sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} f(\mathbf{x}')\ge 0} = 1 = \mathcal{C}^*_{\ell_\gamma,\mathcal{H}}(\mathbf{x},\eta),$$

which implies any surrogate loss ℓ is $\mathcal{H}$-calibrated with respect to ℓ_γ by (5). □

E.2 Proof of Theorem 7, Theorem 9 and Theorem 10

We first characterize the calibration function $\delta_{\max}(\epsilon,\mathbf{x},\eta)$ of losses (ℓ,ℓ_γ) at $\eta=\frac{1}{2}$, $\epsilon=\frac{1}{2}$ and distinguishing $\mathbf{x}_0\in\mathcal{X}$ given a hypothesis set $\mathcal{H}$ which is regular for adversarial calibration.

Lemma 25. *Let $\mathcal{H}$ be a hypothesis set that is regular for adversarial calibration. For distinguishing $\mathbf{x}_0\in\mathcal{X}$, the calibration function $\delta_{\max}(\epsilon,\mathbf{x},\eta)$ of losses (ℓ,ℓ_γ) satisfies*

$$\delta_{\max}\left(\frac{1}{2},\mathbf{x}_0,\frac{1}{2}\right) = \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x}_0,\gamma)\le 0\le\overline{M}(f,\mathbf{x}_0,\gamma)} \Delta\mathcal{C}_{\ell,\mathcal{H}}(f,\mathbf{x}_0,\frac{1}{2}).$$

Proof. By the definition of inner risk (4) and adversarial 0-1 loss ℓ_γ (9), the inner ℓ_γ-risk is

$$\begin{aligned}\mathcal{C}_{\ell_\gamma}(f,\mathbf{x},\eta) &= \eta\mathbb{1}_{\{\underline{M}(f,\mathbf{x},\gamma)\le 0\}} + (1-\eta)\mathbb{1}_{\{\overline{M}(f,\mathbf{x},\gamma)\ge 0\}}\\ &= \begin{cases}1 & \text{if } \underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma),\\ \eta & \text{if } \overline{M}(f,\mathbf{x},\gamma)<0,\\ 1-\eta & \text{if } \underline{M}(f,\mathbf{x},\gamma)>0.\end{cases}\end{aligned}$$

For distinguishing $\mathbf{x}_0$ and $\eta\in[0,1]$, $\{f\in\mathcal{H}:\overline{M}(f,\mathbf{x}_0,\gamma)\}<0\}$ and $\{f\in\mathcal{H}:\underline{M}(f,\mathbf{x}_0,\gamma)>0\}$ are not empty sets. Thus

$$\mathcal{C}^*_{\ell_\gamma,\mathcal{H}}(\mathbf{x}_0,\eta) = \inf_{f\in\mathcal{H}}\mathcal{C}_{\ell_\gamma}(f,\mathbf{x}_0,\eta) = \min\{\eta,1-\eta\}.$$

Note for $f\in\{f\in\mathcal{H}:\underline{M}(f,\mathbf{x}_0,\gamma)\le 0\le\overline{M}(f,\mathbf{x}_0,\gamma)\}$, $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f,\mathbf{x}_0,\eta) = \max\{\eta,1-\eta\}$; for $f\in\{f\in\mathcal{H}:\overline{M}(f,\mathbf{x}_0,\gamma)\}<0\}$, $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f,\mathbf{x}_0,\eta) = \eta-\min\{\eta,1-\eta\} = \max\{0,2\eta-1\} = |2\eta-1|\mathbb{1}_{(2\eta-1)(\underline{M}(f,\mathbf{x}_0,\gamma))\le 0}$ since $\underline{M}(f,\mathbf{x}_0,\gamma)\le\overline{M}(f,\mathbf{x}_0,\gamma)<0$; for $f\in\{f\in\mathcal{H}:\underline{M}(f,\mathbf{x}_0,\gamma)>0\}$, $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f,\mathbf{x}_0,\eta) = (1-\eta)-\min\{\eta,1-\eta\} = \max\{0,1-2\eta\} = |2\eta-1|\mathbb{1}_{(2\eta-1)(\underline{M}(f,\mathbf{x}_0,\gamma))\le 0}$. Therefore,

$$\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f,\mathbf{x}_0,\eta) = \begin{cases}\max\{\eta,1-\eta\} & \text{if } \underline{M}(f,\mathbf{x}_0,\gamma)\le 0\le\overline{M}(f,\mathbf{x}_0,\gamma),\\ |2\eta-1|\mathbb{1}_{(2\eta-1)(\underline{M}(f,\mathbf{x}_0,\gamma))\le 0} & \text{if } \underline{M}(f,\mathbf{x}_0,\gamma)>0 \text{ or } \overline{M}(f,\mathbf{x}_0,\gamma)<0.\end{cases}$$

By (6), for a fixed $\eta\in[0,1]$ and $\mathbf{x}\in\mathcal{X}$, the calibration function of losses (ℓ,ℓ_γ) is

$$\delta_{\max}(\epsilon,\mathbf{x},\eta) = \inf_{f\in\mathcal{H}}\left\{\Delta\mathcal{C}_{\ell,\mathcal{H}}(f,\mathbf{x},\eta)\mid\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f,\mathbf{x},\eta)\ge\epsilon\right\}.$$

Observe that for all $\eta\in[0,1]$,

$$\max\{\eta,1-\eta\} = \frac{1}{2}[(1-\eta)+\eta+|(1-\eta)-\eta|] = \frac{1}{2}[1+|2\eta-1|]\ge|2\eta-1|. \tag{11}$$

For distinguishing $\mathbf{x}_0$, $\eta = \frac{1}{2}$ and $\epsilon = \frac{1}{2}$, $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f, \mathbf{x}_0, \frac{1}{2}) \geq \frac{1}{2}$ if and only if $\underline{M}(f, \mathbf{x}_0, \gamma) \leq 0 \leq \overline{M}(f, \mathbf{x}_0, \gamma)$ since $|2\eta - 1| < \epsilon \leq \max\{\eta, 1-\eta\}$. Therefore,

$$\delta_{\max}\left(\frac{1}{2}, \mathbf{x}_0, \frac{1}{2}\right) = \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x}_0,\gamma)\leq 0\leq\overline{M}(f,\mathbf{x}_0,\gamma)} \Delta\mathcal{C}_{\ell,\mathcal{H}}(f, \mathbf{x}_0, \frac{1}{2}).$$

□

Theorem 7. *Assume $\mathcal{H}$ is such that there exists a distinguishing $\mathbf{x}_0 \in \mathcal{X}$ and $f_0 \in \mathcal{H}$ such that $f_0(\mathbf{x}_0) = 0$. If a margin-based loss $\phi: \mathbb{R} \to \mathbb{R}_+$ is convex, then it is not $\mathcal{H}$-calibrated with respect to ℓ_γ.*

Proof. By Lemma 25, for distinguishing $\mathbf{x}_0 \in \mathcal{X}$, the calibration function $\delta_{\max}(\epsilon, \mathbf{x}, \eta)$ of losses (ϕ, ℓ_γ) satisfies

$$\delta_{\max}\left(\frac{1}{2}, \mathbf{x}_0, \frac{1}{2}\right) = \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x}_0,\gamma)\leq 0\leq\overline{M}(f,\mathbf{x}_0,\gamma)} \Delta\mathcal{C}_{\phi,\mathcal{H}}(f, \mathbf{x}_0, \frac{1}{2}).$$

Suppose that ϕ is $\mathcal{H}$-calibrated with respect to ℓ_γ. By Proposition 4, ϕ is $\mathcal{H}$-calibrated with respect to ℓ_γ if and only if its calibration function $\delta_{\max}$ satisfies $\delta_{\max}(\epsilon, \mathbf{x}, \eta) > 0$ for all $\mathbf{x} \in \mathcal{X}$, $\eta \in [0, 1]$ and $\epsilon > 0$. In particular, the condition requires $\delta_{\max}\left(\frac{1}{2}, \mathbf{x}_0, \frac{1}{2}\right) > 0$, that is,

$$\inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x}_0,\gamma)\leq 0\leq\overline{M}(f,\mathbf{x}_0,\gamma)} \Delta\mathcal{C}_{\phi,\mathcal{H}}(f, \mathbf{x}_0, \frac{1}{2}) > 0,$$

which is equivalent to

$$\inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x}_0,\gamma)\leq 0\leq\overline{M}(f,\mathbf{x}_0,\gamma)} \mathcal{C}_{\phi}(f, \mathbf{x}_0, \frac{1}{2}) > \inf_{f\in\mathcal{H}} \mathcal{C}_{\phi}(f, \mathbf{x}_0, \frac{1}{2}), \tag{12}$$

By the definition of inner risk (4),

$$\mathcal{C}_{\phi}(f, \mathbf{x}_0, \frac{1}{2}) = \frac{1}{2}(\phi(f(\mathbf{x}_0)) + \phi(-f(\mathbf{x}_0))). \tag{13}$$

Since ϕ is convex, by Jensen's inequality, for any $f \in \mathcal{H}$, the following holds:

$$\mathcal{C}_{\phi}(f, \mathbf{x}_0, \frac{1}{2}) \geq \phi\left(\frac{1}{2}f(\mathbf{x}_0) - \frac{1}{2}f(\mathbf{x}_0)\right) = \phi(0).$$

For $f = f_0$, we have $f_0(\mathbf{x}_0) = 0$ and by (13),

$$\mathcal{C}_{\phi}(f_0, \mathbf{x}_0, \frac{1}{2}) = \frac{1}{2}(\phi(0) + \phi(0)) = \phi(0).$$

Moreover, when $f = f_0$, $\underline{M}(f_0, \mathbf{x}_0, \gamma) \leq f_0(\mathbf{x}_0) = 0 \leq \overline{M}(f_0, \mathbf{x}_0, \gamma)$. Thus

$$\inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x}_0,\gamma)\leq 0\leq\overline{M}(f,\mathbf{x}_0,\gamma)} \mathcal{C}_{\phi}(f, \mathbf{x}_0, \frac{1}{2}) = \inf_{f\in\mathcal{H}} \mathcal{C}_{\phi}(f, \mathbf{x}_0, \frac{1}{2}) = \phi(0),$$

where the minimum can be achieved by $f = f_0$, contradicting (12). Therefore, ϕ is not $\mathcal{H}$-calibrated with respect to ℓ_γ. □

Theorem 9. *Let ϕ be a convex and non-increasing margin-based loss. Consider the surrogate loss defined by $\tilde{\phi}(f, \mathbf{x}, y) = \sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\leq\gamma} \phi(yf(\mathbf{x}'))$. Then $\tilde{\phi}$ is not $\mathcal{H}$-calibrated with respect to ℓ_γ, for $\mathcal{H} = \mathcal{H}_{\text{lin}}$, $\mathcal{H}_g$ with a non-decreasing and continuous function g such that $g(-\gamma) + G > 0$ and $g(\gamma) - G < 0$, and $\mathcal{H}_{\text{relu}}$ with $G > \gamma$.*

Proof. By Lemma 25, for distinguishing $\mathbf{x}_0 \in \mathcal{X}$, the calibration function $\delta_{\max}(\epsilon, \mathbf{x}, \eta)$ of losses $(\tilde{\phi}, \ell_\gamma)$ satisfies

$$\delta_{\max}\left(\frac{1}{2}, \mathbf{x}_0, \frac{1}{2}\right) = \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x}_0,\gamma)\leq 0\leq\overline{M}(f,\mathbf{x}_0,\gamma)} \Delta\mathcal{C}_{\tilde{\phi},\mathcal{H}}(f, \mathbf{x}_0, \frac{1}{2}).$$

Next we first consider the case where $\mathcal{H} = \mathcal{H}_{\text{lin}}$. Take distinguishing $\mathbf{x}_0 \in \mathcal{X}$ and $f_0 \in \mathcal{H}_{\text{lin}}$ such that $f_0(\mathbf{x}_0) = 0$. As shown by Awasthi et al. (2020), for $f \in \mathcal{H}_{\text{lin}} = \{\mathbf{x} \to \mathbf{w} \cdot \mathbf{x} \mid \|\mathbf{w}\| = 1\}$,

$$\underline{M}(f, \mathbf{x}, \gamma) = \inf_{\mathbf{x}' : \|\mathbf{x}-\mathbf{x}'\| \le \gamma} f(\mathbf{x}') = \inf_{\mathbf{x}' : \|\mathbf{x}-\mathbf{x}'\| \le \gamma} (\mathbf{w} \cdot \mathbf{x}') = \mathbf{w} \cdot \mathbf{x} - \gamma \|\mathbf{w}\| = f(\mathbf{x}) - \gamma,$$

$$\overline{M}(f, \mathbf{x}, \gamma) = -\inf_{\mathbf{x}' : \|\mathbf{x}-\mathbf{x}'\| \le \gamma} -f(\mathbf{x}') = -\inf_{\mathbf{x}' : \|\mathbf{x}-\mathbf{x}'\| \le \gamma} (-\mathbf{w} \cdot \mathbf{x}') = \mathbf{w} \cdot \mathbf{x} + \gamma \|\mathbf{w}\| = f(\mathbf{x}) + \gamma.$$

Suppose that $\tilde{\phi}$ is $\mathcal{H}_{\text{lin}}$-calibrated with respect to ℓ_γ. By Proposition 4, $\tilde{\phi}$ is $\mathcal{H}_{\text{lin}}$-calibrated with respect to ℓ_γ if and only if its calibration function $\delta_{\max}$ satisfies $\delta_{\max}(\epsilon, \mathbf{x}, \eta) > 0$ for all $\mathbf{x} \in \mathcal{X}$, $\eta \in [0, 1]$ and $\epsilon > 0$. In particular, the condition requires $\delta_{\max}\left(\frac{1}{2}, \mathbf{x}_0, \frac{1}{2}\right) > 0$, that is,

$$\inf_{f \in \mathcal{H}_{\text{lin}} : -\gamma \le f(\mathbf{x}_0) \le \gamma} \Delta\mathcal{C}_{\tilde{\phi}, \mathcal{H}_{\text{lin}}}(f, \mathbf{x}_0, \frac{1}{2}) > 0,$$

which is equivalent to

$$\inf_{f \in \mathcal{H}_{\text{lin}} : -\gamma \le f(\mathbf{x}_0) \le \gamma} \mathcal{C}_{\tilde{\phi}}(f, \mathbf{x}_0, \frac{1}{2}) > \inf_{f \in \mathcal{H}_{\text{lin}}} \mathcal{C}_{\tilde{\phi}}(f, \mathbf{x}_0, \frac{1}{2}), \tag{14}$$

By (19), for $f \in \mathcal{H}_{\text{lin}}$,

$$\mathcal{C}_{\tilde{\phi}}(f, \mathbf{x}_0, \frac{1}{2}) = \frac{1}{2}\phi(f(\mathbf{x}_0) - \gamma) + \frac{1}{2}\phi(-f(\mathbf{x}_0) - \gamma). \tag{15}$$

Since ϕ is convex, by Jensen's inequality, for any $f \in \mathcal{H}_{\text{lin}}$, the following holds:

$$\mathcal{C}_{\tilde{\phi}}(f, \mathbf{x}_0, \frac{1}{2}) \ge \phi\left(\frac{1}{2}(f(\mathbf{x}_0) - \gamma) - \frac{1}{2}(f(\mathbf{x}_0) + \gamma)\right) = \phi(-\gamma).$$

For $f = f_0$, we have $f_0(\mathbf{x}_0) = 0$ and by (15),

$$\mathcal{C}_{\tilde{\phi}}(f_0, \mathbf{x}_0, \frac{1}{2}) = \frac{1}{2}(\phi(-\gamma) + \phi(-\gamma)) = \phi(-\gamma).$$

Moreover, when $f = f_0$, $-\gamma \le f_0(\mathbf{x}_0) = 0 \le \gamma$. Thus

$$\inf_{f \in \mathcal{H} : -\gamma \le f(\mathbf{x}_0) \le \gamma} \mathcal{C}_{\tilde{\phi}}(f, \mathbf{x}_0, \frac{1}{2}) = \inf_{f \in \mathcal{H}} \mathcal{C}_{\tilde{\phi}}(f, \mathbf{x}_0, \frac{1}{2}) = \phi(-\gamma),$$

where the minimum can be achieved by $f = f_0$, contradicting (14). Therefore, $\tilde{\phi}$ is not $\mathcal{H}_{\text{lin}}$-calibrated with respect to ℓ_γ.

Then we consider the case where $\mathcal{H} = \mathcal{H}_g$. By the assumption on g, $0 \in \mathcal{X}$ is distinguishing. As shown by Awasthi et al. (2020), for $f \in \mathcal{H}_g$,

$$\underline{M}(f, \mathbf{x}, \gamma) = g(\mathbf{w} \cdot \mathbf{x} - \gamma) + b, \quad \overline{M}(f, \mathbf{x}, \gamma) = g(\mathbf{w} \cdot \mathbf{x} + \gamma) + b.$$

Suppose that $\tilde{\phi}$ is $\mathcal{H}_g$-calibrated with respect to ℓ_γ.By Proposition 4, $\tilde{\phi}$ is $\mathcal{H}_g$-calibrated with respect to ℓ_γ if and only if its calibration function $\delta_{\max}$ satisfies $\delta_{\max}(\epsilon, \mathbf{x}, \eta) > 0$ for all $\mathbf{x} \in \mathcal{X}$, $\eta \in [0, 1]$ and $\epsilon > 0$. In particular, the condition requires $\delta_{\max}\left(\frac{1}{2}, 0, \frac{1}{2}\right) > 0$, that is,

$$\inf_{f \in \mathcal{H}_g : g(-\gamma)+b \le 0 \le g(\gamma)+b} \Delta\mathcal{C}_{\tilde{\phi}, \mathcal{H}_g}(f, 0, \frac{1}{2}) > 0,$$

which is equivalent to

$$\inf_{f \in \mathcal{H}_g : g(-\gamma)+b \le 0 \le g(\gamma)+b} \mathcal{C}_{\tilde{\phi}}(f, 0, \frac{1}{2}) > \inf_{f \in \mathcal{H}_g} \mathcal{C}_{\tilde{\phi}}(f, 0, \frac{1}{2}), \tag{16}$$

By (19), for $f \in \mathcal{H}_g$,

$$\mathcal{C}_{\tilde{\phi}}(f, 0, \frac{1}{2}) = \frac{1}{2}\phi(g(-\gamma) + b) + \frac{1}{2}\phi(-g(\gamma) - b). \tag{17}$$

Since ϕ is convex, by Jensen's inequality, for any $f \in \mathcal{H}_g$, the following holds:

$$\mathcal{C}_{\tilde{\phi}}(f, 0, \frac{1}{2}) \ge \phi\left(\frac{1}{2}(g(-\gamma) + b) + \frac{1}{2}(-g(\gamma) - b)\right) = \phi\left(\frac{g(-\gamma) - g(\gamma)}{2}\right).$$

Take $f_0 \in \mathcal{H}_g$ with $b_0 = \frac{-g(\gamma)-g(-\gamma)}{2}$, we have $g(-\gamma) + b_0 = -g(\gamma) - b_0 = \frac{g(-\gamma)-g(\gamma)}{2}$ and by (17),

$$\mathcal{C}_{\tilde{\phi}}(f_0, 0, \frac{1}{2}) = \frac{1}{2}\phi(g(-\gamma) + b_0) + \frac{1}{2}\phi(-g(\gamma) - b_0) = \phi\left(\frac{g(-\gamma) - g(\gamma)}{2}\right).$$

Moreover, when $f = f_0$, $g(-\gamma) + b_0 \le 0 \le g(\gamma) + b_0$. Thus

$$\inf_{f \in \mathcal{H}_g:\ g(-\gamma)+b \le 0 \le g(\gamma)+b} \mathcal{C}_{\tilde{\phi}}(f, 0, \frac{1}{2}) = \inf_{f \in \mathcal{H}_g} \mathcal{C}_{\tilde{\phi}}(f, 0, \frac{1}{2}) = \phi\left(\frac{g(-\gamma) - g(\gamma)}{2}\right),$$

where the minimum can be achieved by $f = f_0$, contradicting (16). Therefore, $\tilde{\phi}$ is not $\mathcal{H}_g$-calibrated with respect to ℓ_γ. □

Theorem 10. *Let $\mathcal{H}$ be a hypothesis set containing 0 that is regular for adversarial calibration. If a margin-based loss ϕ is convex and non-increasing, then the surrogate loss defined by $\tilde{\phi}(f, \mathbf{x}, y) = \sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} \phi(yf(\mathbf{x}'))$ is not $\mathcal{H}$-calibrated with respect to ℓ_γ.*

Proof. By Lemma 25, for distinguishing $\mathbf{x}_0 \in \mathcal{X}$, the calibration function $\delta_{\max}(\epsilon, \mathbf{x}, \eta)$ of losses $(\tilde{\phi}, \ell_\gamma)$ satisfies

$$\delta_{\max}\left(\frac{1}{2}, \mathbf{x}_0, \frac{1}{2}\right) = \inf_{f \in \mathcal{H}:\ \underline{M}(f,\mathbf{x}_0,\gamma) \le 0 \le \overline{M}(f,\mathbf{x}_0,\gamma)} \Delta\mathcal{C}_{\tilde{\phi},\mathcal{H}}(f, \mathbf{x}_0, \frac{1}{2}).$$

Suppose that $\tilde{\phi}$ is $\mathcal{H}$-calibrated with respect to ℓ_γ. By Proposition 4, $\tilde{\phi}$ is $\mathcal{H}$-calibrated with respect to ℓ_γ if and only if its calibration function $\delta_{\max}$ satisfies $\delta_{\max}(\epsilon, \mathbf{x}, \eta) > 0$ for all $\mathbf{x} \in \mathcal{X}$, $\eta \in [0, 1]$ and $\epsilon > 0$. In particular, the condition requires $\delta_{\max}\left(\frac{1}{2}, \mathbf{x}_0, \frac{1}{2}\right) > 0$, that is,

$$\inf_{f \in \mathcal{H}:\ \underline{M}(f,\mathbf{x}_0,\gamma) \le 0 \le \overline{M}(f,\mathbf{x}_0,\gamma)} \Delta\mathcal{C}_{\tilde{\phi},\mathcal{H}}(f, \mathbf{x}_0, \frac{1}{2}) > 0,$$

which is equivalent to

$$\inf_{f \in \mathcal{H}:\ \underline{M}(f,\mathbf{x}_0,\gamma) \le 0 \le \overline{M}(f,\mathbf{x}_0,\gamma)} \mathcal{C}_{\tilde{\phi}}(f, \mathbf{x}_0, \frac{1}{2}) > \inf_{f \in \mathcal{H}} \mathcal{C}_{\tilde{\phi}}(f, \mathbf{x}_0, \frac{1}{2}), \tag{18}$$

As shown by Awasthi et al. (2020), $\tilde{\phi}$ has the equivalent form

$$\tilde{\phi}(f, \mathbf{x}, y) = \phi\left(\inf_{\|\mathbf{x}'-\mathbf{x}\|\le\gamma} (yf(\mathbf{x}'))\right).$$

By the definition of inner risk (4),

$$\mathcal{C}_{\tilde{\phi}}(f, \mathbf{x}_0, \frac{1}{2}) = \frac{1}{2}(\phi(\underline{M}(f, \mathbf{x}_0, \gamma)) + \phi(-\overline{M}(f, \mathbf{x}_0, \gamma))). \tag{19}$$

Since ϕ is convex, by Jensen's inequality, for any $f \in \mathcal{H}$, the following holds:

$$\mathcal{C}_{\tilde{\phi}}(f, \mathbf{x}_0, \frac{1}{2}) \ge \phi\left(\frac{1}{2}\underline{M}(f, \mathbf{x}_0, \gamma) - \frac{1}{2}\overline{M}(f, \mathbf{x}_0, \gamma)\right) = \phi\left(\frac{1}{2}(\underline{M}(f, \mathbf{x}_0, \gamma) - \overline{M}(f, \mathbf{x}_0, \gamma))\right) \ge \phi(0),$$

where the last inequality used the fact that

$$\frac{1}{2}(\underline{M}(f, \mathbf{x}_0, \gamma) - \overline{M}(f, \mathbf{x}_0, \gamma)) \le 0$$

and ϕ is non-increasing. For $f = 0$, we have $\underline{M}(f, \mathbf{x}_0, \gamma) = \overline{M}(f, \mathbf{x}_0, \gamma) = 0$ and by (19),

$$\mathcal{C}_{\tilde{\phi}}(f, \mathbf{x}_0, \frac{1}{2}) = \frac{1}{2}(\phi(0) + \phi(0)) = \phi(0).$$

Moreover, when $\underline{M}(f, \mathbf{x}_0, \gamma) = \overline{M}(f, \mathbf{x}_0, \gamma) = 0$, $\underline{M}(f, \mathbf{x}_0, \gamma) \le 0 \le \overline{M}(f, \mathbf{x}_0, \gamma)$ is satisfied. Thus

$$\inf_{f \in \mathcal{H}:\ \underline{M}(f,\mathbf{x}_0,\gamma) \le 0 \le \overline{M}(f,\mathbf{x}_0,\gamma)} \mathcal{C}_{\tilde{\phi}}(f, \mathbf{x}_0, \frac{1}{2}) = \inf_{f \in \mathcal{H}} \mathcal{C}_{\tilde{\phi}}(f, \mathbf{x}_0, \frac{1}{2}) = \phi(0),$$

where the minimum can be achieved by $f = 0$, contradicting (18). Therefore, $\tilde{\phi}$ is not $\mathcal{H}$-calibrated with respect to ℓ_γ. □

E.3 Properties of generic conditional ϕ-risk

In this section, we characterize the properties of the generic conditional ϕ-risk $\bar{\mathcal{C}}_\phi(t,\eta)$ when margin-based loss ϕ is bounded, continuous, non-increasing and satisfy $\bar{\mathcal{C}}_\phi(t,\eta)$ is quasi-concave in $t \in \mathbb{R}$ for all $\eta \in [0,1]$, which would be useful in the proof of Theorem 12 and Theorem 13. Without loss of generality, assume that g is continuous, non-decreasing and satisfies $g(-1-\gamma)+G>0$, $g(1+\gamma)-G<0$.

Lemma 26. *Let ϕ be a margin-based loss. If ϕ is bounded, continuous, non-increasing and satisfy $\bar{\mathcal{C}}_\phi(t,\eta)$ is quasi-concave in $t \in \mathbb{R}$ for all $\eta \in [0,1]$, then*

1. *$\bar{\mathcal{C}}_\phi(t,\frac{1}{2})$ is even and non-increasing in t when $t \geq 0$.*
2. *For $l,u \in \mathbb{R}(l \leq u)$, $\inf_{t\in[l,u]} \bar{\mathcal{C}}_\phi(t,\eta) = \min\{\bar{\mathcal{C}}_\phi(l,\eta), \bar{\mathcal{C}}_\phi(u,\eta)\}$ for all $\eta \in [0,1]$.*
3. *For all $\eta \in (\frac{1}{2},1]$, $\bar{\mathcal{C}}_\phi(t,\eta)$ is non-increasing in t when $t \geq 0$.*
4. *For all $\eta \in [0,\frac{1}{2})$, $\bar{\mathcal{C}}_\phi(t,\eta)$ is non-decreasing in t when $t \leq 0$.*
5. *If $\phi(-t) > \phi(t)$ for any $\gamma < t \leq 1$, then, for all $\eta \in (\frac{1}{2},1]$ and any $\gamma < t \leq 1$, $\bar{\mathcal{C}}_\phi(-t,\eta) > \bar{\mathcal{C}}_\phi(t,\eta)$.*
6. *If $\phi(-t) > \phi(t)$ for any $\gamma < t \leq 1$, then, for all $\eta \in [0,\frac{1}{2})$ and any $\gamma < t \leq 1$, $\bar{\mathcal{C}}_\phi(-t,\eta) < \bar{\mathcal{C}}_\phi(t,\eta)$.*
7. *If $\phi(g(-t)-G) > \phi(G-g(-t))$, $g(-t)+g(t) \geq 0$ for any $0 \leq t \leq 1$, then, for all $\eta \in (\frac{1}{2},1]$ and any $0 \leq t \leq 1$, $\bar{\mathcal{C}}_\phi(g(-t)-G,\eta) > \bar{\mathcal{C}}_\phi(g(t)+G,\eta)$.*
8. *If $\phi(g(-t)-G) > \phi(G-g(-t))$, $g(-t)+g(t) \geq 0$ for any $0 \leq t \leq 1$, then, for any $0 \leq t \leq 1$, $\bar{\mathcal{C}}_\phi(g(-t)-G,\eta) < \bar{\mathcal{C}}_\phi(g(t)+G,\eta)$ for all $\eta \in [0,\frac{1}{2})$ if and only if $\phi(G-g(-t))+\phi(g(-t)-G) = \phi(g(t)+G)+\phi(-g(t)-G)$.*

Proof. Part 1 and Part 3 of Lemma 26 are stated in (Bao et al., 2020a, Lemma 13). Part 2 is implied straightforwardly by the assumption $\bar{\mathcal{C}}_\phi(t,\eta)$ is quasi-concave in $t \in \mathbb{R}$ for all $\eta \in [0,1]$ and the characterization of continuous and quasi-convex functions in (Boyd and Vandenberghe, 2014).

Consider Part 4. For $\eta \in [0,\frac{1}{2})$, and $t_1, t_2 \leq 0$. Suppose that $t_1 < t_2$, then

$$\begin{aligned}
&\phi(t_1)-\phi(-t_1)-\phi(t_2)+\phi(-t_2)\\
\geq&\phi(t_2)-\phi(-t_2)-\phi(t_2)+\phi(-t_2)\\
=&0
\end{aligned}$$

since ϕ is non-increasing. By Part 1 of Lemma 26, $\phi(t)+\phi(-t)$ is non-decreasing in t when $t \leq 0$. Therefore, for $\eta \in [0,\frac{1}{2})$,

$$\begin{aligned}
&\bar{\mathcal{C}}_\phi(t_1,\eta)-\bar{\mathcal{C}}_\phi(t_2,\eta)\\
=&(\phi(t_1)-\phi(-t_1)-\phi(t_2)+\phi(-t_2))\eta+\phi(-t_1)-\phi(-t_2)\\
\leq&(\phi(t_1)-\phi(-t_1)-\phi(t_2)+\phi(-t_2))\frac{1}{2}+\phi(-t_1)-\phi(-t_2)\\
=&\frac{1}{2}(\phi(t_1)+\phi(-t_1)-\phi(t_2)-\phi(-t_2))\\
\leq&0.
\end{aligned}$$

Consider Part 5, For $\eta \in (\frac{1}{2},1]$ and any $\gamma < t \leq 1$,

$$\begin{aligned}
\bar{\mathcal{C}}_\phi(-t,\eta)-\bar{\mathcal{C}}_\phi(t,\eta) &= \eta\phi(-t)+(1-\eta)\phi(t)-\eta\phi(t)-(1-\eta)\phi(-t)\\
&= (2\eta-1)\left[\phi(-t)-\phi(t)\right] > 0
\end{aligned}$$

since $\eta > \frac{1}{2}$ and $\phi(-t) > \phi(t)$ for any $\gamma < t \leq 1$.

Consider Part 6, For $\eta \in [0, \frac{1}{2})$ and any $\gamma < t \leq 1$,

$$\begin{aligned}\bar{\mathcal{C}}_\phi(t,\eta) - \bar{\mathcal{C}}_\phi(-t,\eta) &= \eta\phi(t) + (1-\eta)\phi(-t) - \eta\phi(-t) - (1-\eta)\phi(t) \\ &= (1-2\eta)\left[\phi(-t) - \phi(t)\right] > 0\end{aligned}$$

since $\eta < \frac{1}{2}$ and $\phi(-t) > \phi(t)$ for any $\gamma < t \leq 1$.

Consider Part 7. For $\eta \in (\frac{1}{2}, 1]$ and any $0 \leq t \leq 1$,

$$\begin{aligned}&\bar{\mathcal{C}}_\phi(g(-t) - G, \eta) - \bar{\mathcal{C}}_\phi(g(t) + G, \eta) \\ \geq&\bar{\mathcal{C}}_\phi(g(-t) - G, \eta) - \bar{\mathcal{C}}_\phi(G - g(-t), \eta) && (g(-t) + g(t) \geq 0, \text{ Part 3 of Lemma 26}) \\ =&(2\eta - 1)[\phi(g(-t) - G) - \phi(G - g(-t))] \\ >&0 && (\phi(g(-t) - G) > \phi(G - g(-t)))\end{aligned}$$

Consider Part 8. Since ϕ is non-increasing, for any $0 \leq t \leq 1$,

$$\begin{aligned}&\phi(g(-t) - G) - \phi(G - g(-t)) + \phi(-g(t) - G) - \phi(g(t) + G) \\ \geq&\phi(g(-t) - G) - \phi(G - g(-t)) + \phi(g(t) + G) - \phi(g(t) + G) && (g(t) + G > 0) \\ =&\phi(g(-t) - G) - \phi(G - g(-t)) \\ >&0 && (\phi(g(-t) - G) > \phi(G - g(-t)))\end{aligned}$$

$\Longleftarrow$: Suppose $\phi(G - g(-t)) + \phi(g(-t) - G) = \phi(g(t) + G) + \phi(-g(t) - G)$, then for $\eta \in [0, \frac{1}{2})$,

$$\begin{aligned}&\bar{\mathcal{C}}_\phi(g(-t) - G, \eta) - \bar{\mathcal{C}}_\phi(g(t) + G, \eta) \\ =&(\phi(g(-t) - G) - \phi(G - g(-t)) + \phi(-g(t) - G) - \phi(g(t) + G))\eta \\ &\quad + \phi(G - g(-t)) - \phi(-g(t) - G) \\ <&(\phi(g(-t) - G) - \phi(G - g(-t)) + \phi(-g(t) - G) - \phi(g(t) + G))\frac{1}{2} \\ &\quad + \phi(G - g(-t)) - \phi(-g(t) - G) \\ =&\frac{1}{2}(\phi(G - g(-t)) + \phi(g(-t) - G) - \phi(g(t) + G) - \phi(-g(t) - G)) \\ =&0.\end{aligned}$$

$\Longrightarrow$: Suppose $\bar{\mathcal{C}}_\phi(g(-t) - G, \eta) < \bar{\mathcal{C}}_\phi(g(t) + G, \eta)$ for $\eta \in [0, \frac{1}{2})$, then

$$\begin{aligned}&\bar{\mathcal{C}}_\phi(g(-t) - G, \eta) - \bar{\mathcal{C}}_\phi(g(t) + G, \eta) \\ =&(\phi(g(-t) - G) - \phi(G - g(-t)) + \phi(-g(t) - G) - \phi(g(t) + G))\eta \\ &\quad + \phi(G - g(-1)) - \phi(-g(1) - G) \\ <&0\end{aligned}$$

for $\eta \in [0, \frac{1}{2})$. By taking $\eta \to \frac{1}{2}$, we have

$$\begin{aligned}&\frac{1}{2}(\phi(G - g(-t)) + \phi(g(-t) - G) - \phi(g(t) + G) - \phi(-g(t) - G)) \\ =&(\phi(g(-t) - G) - \phi(G - g(-t)) + \phi(-g(t) - G) - \phi(g(t) + G))\frac{1}{2} \\ &\quad + \phi(G - g(-t)) - \phi(-g(t) - G) \\ \leq&0.\end{aligned}$$

By Part 1 of Lemma 26, we have

$$\begin{aligned}&\phi(G - g(-t)) + \phi(g(-t) - G) - \phi(g(t) + G) - \phi(-g(t) - G) \\ \geq&\phi(g(t) + G) + \phi(-g(t) - G) - \phi(g(t) + G) - \phi(-g(t) - G) && (g(-t) + g(t) \geq 0) \\ =&0.\end{aligned}$$

Therefore, $\phi(G - g(-t)) + \phi(g(-t) - G) - \phi(g(t) + G) - \phi(-g(t) - G) = 0$, i.e., $\phi(G - g(-t)) + \phi(g(-t) - G) = \phi(g(t) + G) + \phi(-g(t) - G)$. □

E.4 Proof of Theorem 12 and Theorem 16

We will make use of general form (9) of the adversarial $0/1$ loss:

$$\ell_\gamma(f,\mathbf{x},y) = \sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} \mathbb{1}_{yf(\mathbf{x}')\le 0} = \mathbb{1}_{\inf_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} yf(\mathbf{x}')\le 0}\,.$$

Next, we first characterize the calibration function $\delta_{\max}(\epsilon,\mathbf{x},\eta)$ of losses (ℓ,ℓ_γ) given a symmetric hypothesis set $\mathcal{H}$.

Lemma 27. *Let $\mathcal{H}$ be a symmetric hypothesis set. For a surrogate loss ℓ, the calibration function $\delta_{\max}(\epsilon,\mathbf{x},\eta)$ of the losses (ℓ,ℓ_γ) is*

$$\delta_{\max}(\epsilon,\mathbf{x},\eta) = \begin{cases} +\infty & \text{if } \mathbf{x}\in\mathcal{X}_1 \text{ or } \mathbf{x}\in\mathcal{X}_2,\ \epsilon>\max\{\eta,1-\eta\},\\ \inf\limits_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma)} \Delta\mathcal{C}_{\ell,\mathcal{H}}(f,\mathbf{x},\eta) & \text{if } \mathbf{x}\in\mathcal{X}_2,\ |2\eta-1|<\epsilon\le\max\{\eta,1-\eta\},\\ \inf\limits_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma) \text{ or } (2\eta-1)(\underline{M}(f,\mathbf{x},\gamma))\le 0} \Delta\mathcal{C}_{\ell,\mathcal{H}}(f,\mathbf{x},\eta) & \text{if } \mathbf{x}\in\mathcal{X}_2,\ \epsilon\le|2\eta-1|, \end{cases}$$

where $\mathcal{X}_1=\{\mathbf{x}\in\mathcal{X}:\underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma),\ \forall f\in\mathcal{H}\}$, $\mathcal{X}_2=\{\mathbf{x}\in\mathcal{X}:$ there exists $f'\in\mathcal{H}$ such that $\underline{M}(f',\mathbf{x},\gamma)>0\}$ and $\mathcal{X}=\mathcal{X}_1\cup\mathcal{X}_2$, $\mathcal{X}_1\cap\mathcal{X}_2=\varnothing$.

Proof. By the definition of inner risk (4) and adversarial 0-1 loss ℓ_γ (9), the inner ℓ_γ-risk is

$$\begin{aligned}\mathcal{C}_{\ell_\gamma}(f,\mathbf{x},\eta) &= \eta\mathbb{1}_{\{\underline{M}(f,\mathbf{x},\gamma)\le 0\}} + (1-\eta)\mathbb{1}_{\{\overline{M}(f,\mathbf{x},\gamma)\ge 0\}}\\ &= \begin{cases} 1 & \text{if } \underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma),\\ \eta & \text{if } \overline{M}(f,\mathbf{x},\gamma)<0,\\ 1-\eta & \text{if } \underline{M}(f,\mathbf{x},\gamma)>0. \end{cases}\end{aligned}$$

Let $\mathcal{X}_1=\{\mathbf{x}\in\mathcal{X}:\underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma),\ \forall f\in\mathcal{H}\}$, $\mathcal{X}_2=\{\mathbf{x}\in\mathcal{X}:$ there exists $f'\in\mathcal{H}$ such that $\underline{M}(f',\mathbf{x},\gamma)>0\}$. It is obvious that $\mathcal{X}_1\cap\mathcal{X}_2=\varnothing$. Since $\mathcal{H}$ is symmetric, for any $\mathbf{x}\in\mathcal{X}$, either there exists $f'\in\mathcal{H}$ such that $\underline{M}(f',\mathbf{x},\gamma)>0$ and $\overline{M}(-f',\mathbf{x},\gamma)<0$, or $\underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma)$ for any $f\in\mathcal{H}$. Thus $\mathcal{X}=\mathcal{X}_1\cup\mathcal{X}_2$. Note when $\mathbf{x}\in\mathcal{X}_1$, $\{f\in\mathcal{H}:\overline{M}(f,\mathbf{x},\gamma)<0\}$ and $\{f\in\mathcal{H}:\underline{M}(f,\mathbf{x},\gamma)>0\}$ are both empty sets. Therefore, the minimal inner ℓ_γ-risk is

$$\mathcal{C}^*_{\ell_\gamma,\mathcal{H}}(\mathbf{x},\eta) = \begin{cases} 1, & \mathbf{x}\in\mathcal{X}_1\,,\\ \min\{\eta,1-\eta\}, & \mathbf{x}\in\mathcal{X}_2\,. \end{cases}$$

Note when $\mathbf{x}\in\mathcal{X}_1$, $\mathcal{C}_{\ell_\gamma}(f,\mathbf{x},\eta)=1$ for any $f\in\mathcal{H}$, thus $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f,\mathbf{x},\eta)=0$. When $\mathbf{x}\in\mathcal{X}_2$, for $f\in\{f\in\mathcal{H}:\underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma)\}$, $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f,\mathbf{x},\eta)=1-\min\{\eta,1-\eta\}=\max\{\eta,1-\eta\}$; for $f\in\{f\in\mathcal{H}:\overline{M}(f,\mathbf{x},\gamma)<0\}$, $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f,\mathbf{x},\eta)=\eta-\min\{\eta,1-\eta\}=\max\{0,2\eta-1\}=|2\eta-1|\mathbb{1}_{(2\eta-1)(\underline{M}(f,\mathbf{x},\gamma))\le 0}$ since $\underline{M}(f,\mathbf{x},\gamma)\le\overline{M}(f,\mathbf{x},\gamma)<0$; for $f\in\{f\in\mathcal{H}:\underline{M}(f,\mathbf{x},\gamma)>0\}$, $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f,\mathbf{x},\eta)=1-\eta-\min\{\eta,1-\eta\}=\max\{0,1-2\eta\}=|2\eta-1|\mathbb{1}_{(2\eta-1)(\underline{M}(f,\mathbf{x},\gamma))\le 0}$ since $\underline{M}(f,\mathbf{x},\gamma)>0$. Therefore,

$$\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f,\mathbf{x},\eta) = \begin{cases} \max\{\eta,1-\eta\} & \text{if } \mathbf{x}\in\mathcal{X}_2,\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma),\\ |2\eta-1|\mathbb{1}_{(2\eta-1)(\underline{M}(f,\mathbf{x},\gamma))\le 0} & \text{if } \mathbf{x}\in\mathcal{X}_2,\ \underline{M}(f,\mathbf{x},\gamma)>0 \text{ or } \overline{M}(f,\mathbf{x},\gamma)<0,\\ 0 & \text{if } \mathbf{x}\in\mathcal{X}_1. \end{cases} \tag{20}$$

By (6), for a fixed $\eta\in[0,1]$ and $\mathbf{x}\in\mathcal{X}$, the calibration function of losses (ℓ,ℓ_γ) is

$$\delta_{\max}(\epsilon,\mathbf{x},\eta) = \inf_{f\in\mathcal{H}}\left\{\Delta\mathcal{C}_{\ell,\mathcal{H}}(f,\mathbf{x},\eta)\mid\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f,\mathbf{x},\eta)\ge\epsilon\right\}$$

If $\mathbf{x}\in\mathcal{X}_1$, then for all $f\in\mathcal{H}$, $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f,\mathbf{x},\eta)=0<\epsilon$, which implies that $\delta_{\max}(\epsilon,\mathbf{x},\eta)=+\infty$. Next we consider case where $\mathbf{x}\in\mathcal{X}_2$. By the observation (11), if $\epsilon>\max\{\eta,1-\eta\}$, then for all $f\in\mathcal{H}$, $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f,\mathbf{x},\eta)<\epsilon$, which implies that $\delta_{\max}(\epsilon,\mathbf{x},\eta)=+\infty$; if $|2\eta-1|<\epsilon\le\max\{\eta,1-\eta\}$, then $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f,\mathbf{x},\eta)\ge\epsilon$ if and only if $\underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma)$, which leads to

$$\delta_{\max}(\epsilon,\mathbf{x},\eta) = \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma)} \Delta\mathcal{C}_{\ell,\mathcal{H}}(f,\mathbf{x},\eta);$$

if $\epsilon \le |2\eta - 1|$, then $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f,\mathbf{x},\eta) \ge \epsilon$ if and only if $\underline{M}(f,\mathbf{x},\gamma) \le 0 \le \overline{M}(f,\mathbf{x},\gamma)$ or $(2\eta - 1)(\underline{M}(f,\mathbf{x},\gamma)) \le 0$, which leads to

$$\delta_{\max}(\epsilon,\mathbf{x},\eta) = \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma) \text{ or } (2\eta-1)(\underline{M}(f,\mathbf{x},\gamma))\le 0} \Delta\mathcal{C}_{\ell,\mathcal{H}}(f,\mathbf{x},\eta).$$

□

We then give the equivalent conditions of calibration based on inner ℓ-risk for any symmetric hypothesis set $\mathcal{H}$.

Lemma 28. *Let $\mathcal{H}$ be a symmetric hypothesis set and ℓ be a surrogate loss function. If $\mathcal{X}_2 = \varnothing$, any loss ℓ is $\mathcal{H}$-calibrated with respect to ℓ_γ. If $\mathcal{X}_2 \ne \varnothing$, then ℓ is $\mathcal{H}$-calibrated with respect to ℓ_γ if and only if for any $\mathbf{x} \in \mathcal{X}_2$,*

$$\inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma)} \mathcal{C}_\ell(f,\mathbf{x},\frac{1}{2}) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f,\mathbf{x},\frac{1}{2}), \textit{and}$$

$$\inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f,\mathbf{x},\eta) \textit{ for all } \eta \in (\frac{1}{2},1], \textit{and}$$

$$\inf_{f\in\mathcal{H}:\ \overline{M}(f,\mathbf{x},\gamma)\ge 0} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f,\mathbf{x},\eta) \textit{ for all } \eta \in [0,\frac{1}{2}).$$

where $\mathcal{X}_2 = \{\mathbf{x} \in \mathcal{X} :$ there exists $f' \in \mathcal{H}$ such that $\underline{M}(f',\mathbf{x},\gamma) > 0\}$.

Proof. Let $\delta_{\max}$ be the calibration function of (ℓ,ℓ_γ) given hypothesis set $\mathcal{H}$. By Lemma 27,

$$\delta_{\max}(\epsilon,\mathbf{x},\eta) = \begin{cases} +\infty & \text{if } \mathbf{x}\in\mathcal{X}_1 \text{ or } \mathbf{x}\in\mathcal{X}_2,\ \epsilon > \max\{\eta,1-\eta\}, \\ \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma)} \Delta\mathcal{C}_{\ell,\mathcal{H}}(f,\mathbf{x},\eta) & \text{if } \mathbf{x}\in\mathcal{X}_2,\ |2\eta-1| < \epsilon \le \max\{\eta,1-\eta\}, \\ \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma) \text{ or } (2\eta-1)(\underline{M}(f,\mathbf{x},\gamma))\le 0} \Delta\mathcal{C}_{\ell,\mathcal{H}}(f,\mathbf{x},\eta) & \text{if } \mathbf{x}\in\mathcal{X}_2,\ \epsilon \le |2\eta-1|, \end{cases}$$

where $\mathcal{X}_1 = \{\mathbf{x} \in \mathcal{X} : \underline{M}(f,\mathbf{x},\gamma) \le 0 \le \overline{M}(f,\mathbf{x},\gamma),\ \forall f \in \mathcal{H}\}$, $\mathcal{X}_2 = \{\mathbf{x} \in \mathcal{X} :$ there exists $f' \in \mathcal{H}$ such that $\underline{M}(f',\mathbf{x},\gamma) > 0\}$ and $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$, $\mathcal{X}_1 \cap \mathcal{X}_2 = \varnothing$. By Proposition 4, ℓ is $\mathcal{H}$-calibrated with respect to ℓ_γ if and only if its calibration function $\delta_{\max}$ satisfies $\delta_{\max}(\epsilon,\mathbf{x},\eta) > 0$ for all $\mathbf{x} \in \mathcal{X}$, $\eta \in [0,1]$ and $\epsilon > 0$. Since $\delta(\epsilon,\mathbf{x},\eta) = +\infty > 0$ when $\mathbf{x} \notin \mathcal{X}_2$, any loss ℓ is $\mathcal{H}$-calibrated with respect to ℓ_γ when $\mathcal{X}_2 = \varnothing$. Furtheremore, when $\mathcal{X}_2 \ne \varnothing$, we only need to analyze $\delta(\epsilon,\mathbf{x},\eta)$ when $\mathbf{x} \in \mathcal{X}_2$. For $\eta = \frac{1}{2}$, we have for any $\mathbf{x} \in \mathcal{X}_2$,

$$\delta_{\max}(\epsilon,\mathbf{x},\frac{1}{2}) > 0 \text{ for all } \epsilon > 0 \Leftrightarrow \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma)} \mathcal{C}_\ell(f,\mathbf{x},\frac{1}{2}) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f,\mathbf{x},\frac{1}{2}). \tag{21}$$

For $1 \ge \eta > \frac{1}{2}$, we have $|2\eta - 1| = 2\eta - 1$, $\max\{\eta, 1-\eta\} = \eta$, and

$$\inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma) \text{ or } (2\eta-1)(\underline{M}(f,\mathbf{x},\gamma))\le 0} \Delta\mathcal{C}_{\ell,\mathcal{H}}(f,\mathbf{x},\eta) = \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0} \Delta\mathcal{C}_{\ell,\mathcal{H}}(f,\mathbf{x},\eta).$$

Therefore, $\delta_{\max}(\epsilon,\mathbf{x},\frac{1}{2}) > 0$ for all $\mathbf{x} \in \mathcal{X}_2, \epsilon > 0$ and $\eta \in (\frac{1}{2},1]$ if and only if for all $\mathbf{x} \in \mathcal{X}_2$,

$$\begin{cases} \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma)} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f,\mathbf{x},\eta) & \text{for all } \eta \in (\frac{1}{2},1] \text{ such that } 2\eta - 1 < \epsilon \le \eta, \\ \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f,\mathbf{x},\eta) & \text{for all } \eta \in (\frac{1}{2},1] \text{ such that } \epsilon \le 2\eta - 1, \end{cases}$$

for all $\epsilon > 0$, which is equivalent to for all $\mathbf{x} \in \mathcal{X}_2$,

$$\begin{cases} \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma)} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f,\mathbf{x},\eta) & \text{for all } \eta \in (\frac{1}{2},1] \text{ such that } \epsilon \le \eta < \frac{\epsilon+1}{2}, \\ \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f,\mathbf{x},\eta) & \text{for all } \eta \in (\frac{1}{2},1] \text{ such that } \frac{\epsilon+1}{2} \le \eta, \end{cases} \tag{22}$$

for all $\epsilon > 0$. Observe that

$$\left\{\eta \in (\frac{1}{2},1] \middle| \epsilon \le \eta < \frac{\epsilon+1}{2}, \epsilon > 0\right\} = \left\{\frac{1}{2} < \eta \le 1\right\}, \text{ and}$$

$$\left\{\eta \in (\frac{1}{2},1] \middle| \frac{\epsilon+1}{2} \le \eta, \epsilon > 0\right\} = \left\{\frac{1}{2} < \eta \le 1\right\}, \text{ and}$$

$$\inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le\overline{M}(f,\mathbf{x},\gamma)} \mathcal{C}_\ell(f,\mathbf{x},\eta) \ge \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0} \mathcal{C}_\ell(f,\mathbf{x},\eta) \text{ for all } \eta.$$

Therefore, we reduce the above condition (22) as for all $\mathbf{x} \in \mathcal{X}_2$,

$$\inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f,\mathbf{x},\eta) \text{ for all } \eta \in (\frac{1}{2},1]\,. \tag{23}$$

For $\frac{1}{2} > \eta \ge 0$, we have $|2\eta - 1| = 1 - 2\eta$, $\max\{\eta, 1-\eta\} = 1-\eta$, and

$$\inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le \overline{M}(f,\mathbf{x},\gamma) \text{ or } (2\eta-1)(\underline{M}(f,\mathbf{x},\gamma))\le 0} \Delta\mathcal{C}_{\ell,\mathcal{H}}(f,\mathbf{x},\eta) = \inf_{f\in\mathcal{H}:\ \overline{M}(f,\mathbf{x},\gamma)\ge 0} \Delta\mathcal{C}_{\ell,\mathcal{H}}(f,\mathbf{x},\eta)\,.$$

Therefore, $\delta_{\max}(\epsilon,\mathbf{x},\frac{1}{2}) > 0$ for all $\mathbf{x}\in\mathcal{X}_2, \epsilon > 0$ and $\eta \in [0,\frac{1}{2})$ if and only if for all $\mathbf{x}\in\mathcal{X}_2$,

$$\begin{cases} \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le \overline{M}(f,\mathbf{x},\gamma)} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f,\mathbf{x},\eta) & \text{for all } \eta\in[0,\frac{1}{2}) \text{ such that } 1-2\eta < \epsilon \le 1-\eta, \\ \inf_{f\in\mathcal{H}:\ \overline{M}(f,\mathbf{x},\gamma)\ge 0} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f,\mathbf{x},\eta) & \text{for all } \eta\in[0,\frac{1}{2}) \text{ such that } \epsilon \le 1-2\eta, \end{cases}$$

for all $\epsilon > 0$, which is equivalent to for all $\mathbf{x}\in\mathcal{X}_2$,

$$\begin{cases} \inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le \overline{M}(f,\mathbf{x},\gamma)} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f,\mathbf{x},\eta) & \text{for all } \eta\in[0,\frac{1}{2}) \text{ such that } \frac{1-\epsilon}{2} < \eta \le 1-\epsilon, \\ \inf_{f\in\mathcal{H}:\ \overline{M}(f,\mathbf{x},\gamma)\ge 0} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f,\mathbf{x},\eta) & \text{for all } \eta\in[0,\frac{1}{2}) \text{ such that } \eta \le \frac{1-\epsilon}{2}, \end{cases} \tag{24}$$

for all $\epsilon > 0$. Observe that

$$\left\{\eta\in[0,\frac{1}{2}) \middle| \frac{1-\epsilon}{2} < \eta \le 1-\epsilon, \epsilon > 0\right\} = \left\{0 \le \eta < \frac{1}{2}\right\}, \text{ and}$$

$$\left\{\eta\in[0,\frac{1}{2}) \middle| \eta \le \frac{1-\epsilon}{2}, \epsilon > 0\right\} = \left\{0 \le \eta < \frac{1}{2}\right\}, \text{ and}$$

$$\inf_{f\in\mathcal{H}:\ \underline{M}(f,\mathbf{x},\gamma)\le 0\le \overline{M}(f,\mathbf{x},\gamma)} \mathcal{C}_\ell(f,\mathbf{x},\eta) \ge \inf_{f\in\mathcal{H}:\ \overline{M}(f,\mathbf{x},\gamma)\ge 0} \mathcal{C}_\ell(f,\mathbf{x},\eta) \text{ for all } \eta\,.$$

Therefore, we reduce the above condition (24) as for all $\mathbf{x}\in\mathcal{X}_2$,

$$\inf_{f\in\mathcal{H}:\ \overline{M}(f,\mathbf{x},\gamma)\ge 0} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f,\mathbf{x},\eta) \text{ for all } \eta\in[0,\frac{1}{2})\,. \tag{25}$$

To sum up, by (21), (23) and (25), we conclude the proof. □

Since $\mathcal{H}_{\text{lin}}$ is a symmetric hypothesis set, we could make use of Lemma 27 and Lemma 28 for proving Theorem 12.

Theorem 12. *Let a margin-based loss ϕ be bounded, continuous, non-increasing, and satisfy the property that $\bar{\mathcal{C}}_\phi(t,\eta)$ is quasi-concave in $t\in\mathbb{R}$ for all $\eta\in[0,1]$. Assume that $\phi(-t) > \phi(t)$ for any $\gamma < t \le 1$. Then ϕ is $\mathcal{H}_{\text{lin}}$-calibrated with respect to ℓ_γ if and only if for any $\gamma < t \le 1$,*

$$\phi(\gamma) + \phi(-\gamma) > \phi(t) + \phi(-t)\,. \tag{10}$$

Proof. As shown by Awasthi et al. (2020), for $f\in\mathcal{H}_{\text{lin}} = \{\mathbf{x}\to\mathbf{w}\cdot\mathbf{x} \mid \|\mathbf{w}\| = 1\}$,

$$\underline{M}(f,\mathbf{x},\gamma) = \inf_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} f(\mathbf{x}') = \inf_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} (\mathbf{w}\cdot\mathbf{x}') = \mathbf{w}\cdot\mathbf{x} - \gamma\|\mathbf{w}\| = f(\mathbf{x}) - \gamma,$$

$$\overline{M}(f,\mathbf{x},\gamma) = -\inf_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} -f(\mathbf{x}') = -\inf_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} (-\mathbf{w}\cdot\mathbf{x}') = \mathbf{w}\cdot\mathbf{x} + \gamma\|\mathbf{w}\| = f(\mathbf{x}) + \gamma.$$

Thus for $\mathcal{H}_{\text{lin}}$, $\mathcal{X}_2 = \{\mathbf{x}\in\mathcal{X} :$ there exists $f'\in\mathcal{H}_{\text{lin}}$ such that $\underline{M}(f',\mathbf{x},\gamma) > 0\} = \{\mathbf{x}\in\mathcal{X} :$ there exists $f'\in\mathcal{H}_{\text{lin}}$ such that $f'(\mathbf{x}) > \gamma\} = \{\mathbf{x} : \gamma < \|\mathbf{x}\| \le 1\}$ since $f(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x}\in[-\|\mathbf{x}\|,\|\mathbf{x}\|]$ when $f\in\mathcal{H}_{\text{lin}}$. Note $\mathcal{H}_{\text{lin}}$ is a symmetric hypothesis set. Therefore, by Lemma 28, ϕ is $\mathcal{H}_{\text{lin}}$-calibrated with respect to ℓ_γ if and only if for any $\mathbf{x}\in\mathcal{X}$ such that $\gamma < \|\mathbf{x}\| \le 1$,

$$\begin{aligned} &\inf_{f\in\mathcal{H}_{\text{lin}}:\ |f(\mathbf{x})|\le\gamma} \mathcal{C}_\phi(f,\mathbf{x},\frac{1}{2}) > \inf_{f\in\mathcal{H}_{\text{lin}}} \mathcal{C}_\phi(f,\mathbf{x},\frac{1}{2}), \text{and} \\ &\inf_{f\in\mathcal{H}_{\text{lin}}:\ f(\mathbf{x})\le\gamma} \mathcal{C}_\phi(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}_{\text{lin}}} \mathcal{C}_\phi(f,\mathbf{x},\eta) \text{ for all } \eta\in(\frac{1}{2},1], \text{and} \\ &\inf_{f\in\mathcal{H}_{\text{lin}}:\ f(\mathbf{x})\ge-\gamma} \mathcal{C}_\phi(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}_{\text{lin}}} \mathcal{C}_\phi(f,\mathbf{x},\eta) \text{ for all } \eta\in[0,\frac{1}{2})\,. \end{aligned} \tag{26}$$

By the definition of inner risk (4), the inner ϕ-risk is

$$\mathcal{C}_\phi(f, \mathbf{x}, \eta) = \eta\phi(f(\mathbf{x})) + (1-\eta)\phi(-f(\mathbf{x}))\,.$$

Note $f(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} \in [-\|\mathbf{x}\|, \|\mathbf{x}\|]$ when $f \in \mathcal{H}_{\text{lin}}$. Therefore, (26) is equivalent to for any $\mathbf{x} \in \mathcal{X}$ such that $\gamma < \|\mathbf{x}\| \le 1$,

$$\begin{aligned}
&\inf_{-\gamma\le t\le\gamma} \bar{\mathcal{C}}_\phi(t, \frac{1}{2}) > \inf_{-\|\mathbf{x}\|\le t\le\|\mathbf{x}\|} \bar{\mathcal{C}}_\phi(t, \frac{1}{2})\,, \text{and} \\
&\inf_{-\|\mathbf{x}\|\le t\le\gamma} \bar{\mathcal{C}}_\phi(t, \eta) > \inf_{-\|\mathbf{x}\|\le t\le\|\mathbf{x}\|} \bar{\mathcal{C}}_\phi(t, \eta) \text{ for all } \eta \in (\frac{1}{2}, 1]\,, \text{and} \\
&\inf_{-\gamma\le t\le\|\mathbf{x}\|} \bar{\mathcal{C}}_\phi(t, \eta) > \inf_{-\|\mathbf{x}\|\le t\le\|\mathbf{x}\|} \bar{\mathcal{C}}_\phi(t, \eta) \text{ for all } \eta \in [0, \frac{1}{2})\,.
\end{aligned} \tag{27}$$

Suppose that ϕ is $\mathcal{H}_{\text{lin}}$-calibrated with respect to ℓ_γ. Since by Part 1 of Lemma 26,

$$\inf_{-\gamma\le t\le\gamma} \bar{\mathcal{C}}_\phi(t, \frac{1}{2}) = \bar{\mathcal{C}}_\phi(\gamma, \frac{1}{2}), \qquad \inf_{-\|\mathbf{x}\|\le t\le\|\mathbf{x}\|} \bar{\mathcal{C}}_\phi(t, \frac{1}{2}) = \bar{\mathcal{C}}_\phi(\|\mathbf{x}\|, \frac{1}{2})\,,$$

we obtain $\phi(\gamma) + \phi(-\gamma) = 2\bar{\mathcal{C}}_\phi(\gamma, \frac{1}{2}) > 2\bar{\mathcal{C}}_\phi(t, \frac{1}{2}) = \phi(t) + \phi(-t)$ for any $\gamma < t \le 1$.

Now for the other direction, assume that $\phi(\gamma) + \phi(-\gamma) > \phi(t) + \phi(-t)$ for any $\gamma < t \le 1$. For $\eta = \frac{1}{2}$, by Part 1 of Lemma 26, we obtain for any $\mathbf{x} \in \mathcal{X}$ such that $\gamma < \|\mathbf{x}\| \le 1$,

$$\inf_{-\gamma\le t\le\gamma} \bar{\mathcal{C}}_\phi(t, \frac{1}{2}) = \bar{\mathcal{C}}_\phi(\gamma, \frac{1}{2}) > \bar{\mathcal{C}}_\phi(\|\mathbf{x}\|, \frac{1}{2}) = \inf_{-\|\mathbf{x}\|\le t\le\|\mathbf{x}\|} \bar{\mathcal{C}}_\phi(t, \frac{1}{2})\,.$$

For $\eta \in (\frac{1}{2}, 1]$ and any $\mathbf{x} \in \mathcal{X}$ such that $\gamma < \|\mathbf{x}\| \le 1$,

$$\begin{aligned}
&\inf_{-\|\mathbf{x}\|\le t\le\gamma} \bar{\mathcal{C}}_\phi(t, \eta) = \min\{\bar{\mathcal{C}}_\phi(\gamma, \eta), \bar{\mathcal{C}}_\phi(-\|\mathbf{x}\|, \eta)\} \quad \text{(Part 2 of Lemma 26)} \\
&\inf_{-\|\mathbf{x}\|\le t\le\|\mathbf{x}\|} \bar{\mathcal{C}}_\phi(t, \eta) = \min\{\bar{\mathcal{C}}_\phi(\|\mathbf{x}\|, \eta), \bar{\mathcal{C}}_\phi(-\|\mathbf{x}\|, \eta)\} \quad \text{(Part 2 of Lemma 26)} \\
&= \bar{\mathcal{C}}_\phi(\|\mathbf{x}\|, \eta) \quad \text{(Part 5 of Lemma 26)}
\end{aligned}$$

Note for $\eta \in (\frac{1}{2}, 1]$ and any $\mathbf{x} \in \mathcal{X}$ such that $\gamma < \|\mathbf{x}\| \le 1$, since ϕ is non-increasing,

$$\phi(\gamma) - \phi(-\gamma) - \phi(\|\mathbf{x}\|) + \phi(-\|\mathbf{x}\|) \ge \phi(\|\mathbf{x}\|) - \phi(-\|\mathbf{x}\|) - \phi(\|\mathbf{x}\|) + \phi(-\|\mathbf{x}\|) = 0.$$

Thus

$$\begin{aligned}
\bar{\mathcal{C}}_\phi(\gamma, \eta) - \bar{\mathcal{C}}_\phi(\|\mathbf{x}\|, \eta) &= \eta\phi(\gamma) + (1-\eta)\phi(-\gamma) - \eta\phi(\|\mathbf{x}\|) - (1-\eta)\phi(-\|\mathbf{x}\|) \\
&= (\phi(\gamma) - \phi(-\gamma) - \phi(\|\mathbf{x}\|) + \phi(-\|\mathbf{x}\|))\,\eta + \phi(-\gamma) - \phi(-\|\mathbf{x}\|) \\
&\ge (\phi(\gamma) - \phi(-\gamma) - \phi(\|\mathbf{x}\|) + \phi(-\|\mathbf{x}\|))\,\frac{1}{2} + \phi(-\gamma) - \phi(-\|\mathbf{x}\|) \\
&= \frac{1}{2}\left[\phi(\gamma) + \phi(-\gamma) - \phi(\|\mathbf{x}\|) - \phi(-\|\mathbf{x}\|)\right] \\
&> 0\,.
\end{aligned}$$

In addition, we have for $\eta \in (\frac{1}{2}, 1]$ and any $\mathbf{x} \in \mathcal{X}$ such that $\gamma < \|\mathbf{x}\| \le 1$,

$$\bar{\mathcal{C}}_\phi(-\|\mathbf{x}\|, \eta) > \bar{\mathcal{C}}_\phi(\|\mathbf{x}\|, \eta). \quad \text{(Part 5 of Lemma 26)}$$

Therefore for $\eta \in (\frac{1}{2}, 1]$ and any $\mathbf{x} \in \mathcal{X}$ such that $\gamma < \|\mathbf{x}\| \le 1$,

$$\inf_{-\|\mathbf{x}\|\le t\le\gamma} \bar{\mathcal{C}}_\phi(t, \eta) = \min\{\bar{\mathcal{C}}_\phi(\gamma, \eta), \bar{\mathcal{C}}_\phi(-\|\mathbf{x}\|, \eta)\} > \bar{\mathcal{C}}_\phi(\|\mathbf{x}\|, \eta) = \inf_{-\|\mathbf{x}\|\le t\le\|\mathbf{x}\|} \bar{\mathcal{C}}_\phi(t, \eta)\,.$$

For $\eta \in [0, \frac{1}{2})$ and any $\mathbf{x} \in \mathcal{X}$ such that $\gamma < \|\mathbf{x}\| \le 1$,

$$\begin{aligned}
&\inf_{-\gamma\le t\le\|\mathbf{x}\|} \bar{\mathcal{C}}_\phi(t, \eta) = \min\{\bar{\mathcal{C}}_\phi(-\gamma, \eta), \bar{\mathcal{C}}_\phi(\|\mathbf{x}\|, \eta)\} \quad \text{(Part 2 of Lemma 26)} \\
&\inf_{-\|\mathbf{x}\|\le t\le\|\mathbf{x}\|} \bar{\mathcal{C}}_\phi(t, \eta) = \min\{\bar{\mathcal{C}}_\phi(\|\mathbf{x}\|, \eta), \bar{\mathcal{C}}_\phi(-\|\mathbf{x}\|, \eta)\} \quad \text{(Part 2 of Lemma 26)} \\
&= \bar{\mathcal{C}}_\phi(-\|\mathbf{x}\|, \eta) \quad \text{(Part 6 of Lemma 26)}
\end{aligned}$$

Note for $\eta \in [0, \frac{1}{2})$ and any $\mathbf{x} \in \mathcal{X}$ such that $\gamma < \|\mathbf{x}\| \le 1$, since ϕ is non-increasing,

$$\phi(-\gamma) - \phi(\gamma) - \phi(-\|\mathbf{x}\|) + \phi(\|\mathbf{x}\|) \le \phi(-\|\mathbf{x}\|) - \phi(\|\mathbf{x}\|) - \phi(-\|\mathbf{x}\|) + \phi(\|\mathbf{x}\|) = 0\,.$$

Thus

$$\begin{aligned}
\bar{\mathcal{C}}_\phi(-\gamma,\eta) - \bar{\mathcal{C}}_\phi(-\|\mathbf{x}\|,\eta) &= \eta\phi(-\gamma) + (1-\eta)\phi(\gamma) - \eta\phi(-\|\mathbf{x}\|) - (1-\eta)\phi(\|\mathbf{x}\|) \\
&= \left(\phi(-\gamma) - \phi(\gamma) - \phi(-\|\mathbf{x}\|) + \phi(\|\mathbf{x}\|)\right)\eta + \phi(\gamma) - \phi(\|\mathbf{x}\|) \\
&\ge \left(\phi(-\gamma) - \phi(\gamma) - \phi(-\|\mathbf{x}\|) + \phi(\|\mathbf{x}\|)\right)\frac{1}{2} + \phi(\gamma) - \phi(\|\mathbf{x}\|) \\
&= \frac{1}{2}\left[\phi(\gamma) + \phi(-\gamma) - \phi(\|\mathbf{x}\|) - \phi(-\|\mathbf{x}\|)\right] \\
&> 0\,.
\end{aligned}$$

In addition, we have for $\eta \in [0, \frac{1}{2})$ and any $\mathbf{x} \in \mathcal{X}$ such that $\gamma < \|\mathbf{x}\| \le 1$,

$$\bar{\mathcal{C}}_\phi(\|\mathbf{x}\|,\eta) > \bar{\mathcal{C}}_\phi(-\|\mathbf{x}\|,\eta). \quad \text{(Part 6 of Lemma 26)}$$

Therefore for $\eta \in [0, \frac{1}{2})$ and any $\mathbf{x} \in \mathcal{X}$ such that $\gamma < \|\mathbf{x}\| \le 1$,

$$\inf_{-\gamma \le t \le \|\mathbf{x}\|} \bar{\mathcal{C}}_\phi(t,\eta) = \min\left\{\bar{\mathcal{C}}_\phi(-\gamma,\eta), \bar{\mathcal{C}}_\phi(\|\mathbf{x}\|,\eta)\right\} > \bar{\mathcal{C}}_\phi(-\|\mathbf{x}\|,\eta) = \inf_{-\|\mathbf{x}\| \le t \le \|\mathbf{x}\|} \bar{\mathcal{C}}_\phi(t,\eta)\,.$$

□

Theorem 16. *Let $\mathcal{H}$ be a symmetric hypothesis set, then $\tilde{\phi}_\rho$ is $\mathcal{H}$-calibrated with respect to ℓ_γ.*

Proof. By Lemma 28, if $\mathcal{X}_2 = \varnothing$, $\tilde{\phi}_\rho$ is $\mathcal{H}$-calibrated with respect to ℓ_γ. Next consider the case where $\mathcal{X}_2 \ne \varnothing$. By Lemma 28, $\tilde{\phi}_\rho$ is $\mathcal{H}$-calibrated with respect to ℓ_γ if and only if for all $\mathbf{x} \in \mathcal{X}_2$,

$$\begin{aligned}
&\inf_{f \in \mathcal{H}:\, \underline{M}(f,\mathbf{x},\gamma) \le 0 \le \overline{M}(f,\mathbf{x},\gamma)} \mathcal{C}_{\tilde{\phi}_\rho}(f,\mathbf{x},\frac{1}{2}) > \inf_{f \in \mathcal{H}} \mathcal{C}_{\tilde{\phi}_\rho}(f,\mathbf{x},\frac{1}{2})\,, \text{and} \\
&\inf_{f \in \mathcal{H}:\, \underline{M}(f,\mathbf{x},\gamma) \le 0} \mathcal{C}_{\tilde{\phi}_\rho}(f,\mathbf{x},\eta) > \inf_{f \in \mathcal{H}} \mathcal{C}_{\tilde{\phi}_\rho}(f,\mathbf{x},\eta) \text{ for all } \eta \in (\frac{1}{2},1]\,, \text{and} \\
&\inf_{f \in \mathcal{H}:\, \overline{M}(f,\mathbf{x},\gamma) \ge 0} \mathcal{C}_{\tilde{\phi}_\rho}(f,\mathbf{x},\eta) > \inf_{f \in \mathcal{H}} \mathcal{C}_{\tilde{\phi}_\rho}(f,\mathbf{x},\eta) \text{ for all } \eta \in [0,\frac{1}{2})\,,
\end{aligned}$$

where $\mathcal{X}_2 = \{\mathbf{x} \in \mathcal{X} :$ there exists $f' \in \mathcal{H}$ such that $\underline{M}(f',\mathbf{x},\gamma) > 0\}$. As shown by Awasthi et al. (2020), $\tilde{\phi}_\rho$ has the equivalent form

$$\tilde{\phi}_\rho(f,\mathbf{x},y) = \phi_\rho\left(\inf_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\| \le \gamma} (yf(\mathbf{x}'))\right).$$

Thus by the definition of inner risk (4), the inner $\tilde{\phi}_\rho$-risk is

$$\mathcal{C}_{\tilde{\phi}_\rho}(f,\mathbf{x},\eta) = \eta\phi_\rho(\underline{M}(f,\mathbf{x},\gamma)) + (1-\eta)\phi_\rho(-\overline{M}(f,\mathbf{x},\gamma))\,.$$

For any $\mathbf{x} \in \mathcal{X}_2$, let $M_\mathbf{x} = \sup_{f \in \mathcal{H}} \underline{M}(f,\mathbf{x},\gamma) > 0$. Since $\mathcal{H}$ is symmetric, we have $-M_\mathbf{x} = \inf_{f \in \mathcal{H}} \overline{M}(f,\mathbf{x},\gamma) < 0$. Since ϕ_ρ is continuous, for any $\mathbf{x} \in \mathcal{X}_2$ and $\epsilon > 0$, there exists $f^\epsilon_\mathbf{x} \in \mathcal{H}$ such that $\phi_\rho(\underline{M}(f^\epsilon_\mathbf{x},\mathbf{x},\gamma)) < \phi_\rho(M_\mathbf{x}) + \epsilon$ and $\overline{M}(f^\epsilon_\mathbf{x},\mathbf{x},\gamma) \ge \underline{M}(f^\epsilon_\mathbf{x},\mathbf{x},\gamma) > 0$, $\underline{M}(-f^\epsilon_\mathbf{x},\mathbf{x},\gamma) \le \overline{M}(-f^\epsilon_\mathbf{x},\mathbf{x},\gamma) = -\underline{M}(f^\epsilon_\mathbf{x},\mathbf{x},\gamma) < 0$. Next we analyze three cases:

- When $\eta = \frac{1}{2}$, since ϕ_ρ is non-increasing,

$$\begin{aligned}
&\inf_{f \in \mathcal{H}:\, \underline{M}(f,\mathbf{x},\gamma) \le 0 \le \overline{M}(f,\mathbf{x},\gamma)} \mathcal{C}_{\tilde{\phi}_\rho}(f,\mathbf{x},\frac{1}{2}) \\
&= \inf_{f \in \mathcal{H}:\, \underline{M}(f,\mathbf{x},\gamma) \le 0 \le \overline{M}(f,\mathbf{x},\gamma)} \frac{1}{2}\phi_\rho(\underline{M}(f,\mathbf{x},\gamma)) + \frac{1}{2}\phi_\rho(-\overline{M}(f,\mathbf{x},\gamma)) \\
&\ge \frac{1}{2}\phi_\rho(0) + \frac{1}{2}\phi_\rho(0) = \phi_\rho(0) = 1\,.
\end{aligned}$$

For any $\mathbf{x} \in \mathcal{X}_2$, there exists $f' \in \mathcal{H}$ such that $\underline{M}(f', \mathbf{x}, \gamma) > 0$ and $-\overline{M}(f', \mathbf{x}, \gamma) \leq -\underline{M}(f', \mathbf{x}, \gamma) < 0$, we obtain

$$\mathcal{C}_{\tilde{\phi}_\rho}(f', \mathbf{x}, \frac{1}{2}) = \frac{1}{2}\phi_\rho(\underline{M}(f', \mathbf{x}, \gamma)) + \frac{1}{2}\phi_\rho(-\overline{M}(f', \mathbf{x}, \gamma)) = \frac{1}{2}\phi_\rho(\underline{M}(f', \mathbf{x}, \gamma)) + \frac{1}{2} < 1.$$

Therefore for any $\mathbf{x} \in \mathcal{X}_2$,

$$\inf_{f \in \mathcal{H}} \mathcal{C}_{\tilde{\phi}_\rho}(f, \mathbf{x}, \frac{1}{2}) \leq \mathcal{C}_{\tilde{\phi}_\rho}(f', \mathbf{x}, \frac{1}{2}) < 1 \leq \inf_{f \in \mathcal{H}:\ \underline{M}(f, \mathbf{x}, \gamma) \leq 0 \leq \overline{M}(f, \mathbf{x}, \gamma)} \mathcal{C}_{\tilde{\phi}_\rho}(f, \mathbf{x}, \frac{1}{2}). \tag{28}$$

- When $\eta \in (\frac{1}{2}, 1]$, since ϕ_ρ is non-increasing, for any $\mathbf{x} \in \mathcal{X}_2$,

$$\begin{aligned}
\inf_{f \in \mathcal{H}:\ \underline{M}(f, \mathbf{x}, \gamma) \leq 0} \mathcal{C}_{\tilde{\phi}_\rho}(f, \mathbf{x}, \eta) &= \inf_{f \in \mathcal{H}:\ \underline{M}(f, \mathbf{x}, \gamma) \leq 0} \eta\phi_\rho(\underline{M}(f, \mathbf{x}, \gamma)) + (1-\eta)\phi_\rho(-\overline{M}(f, \mathbf{x}, \gamma)) \\
&= \eta + \inf_{f \in \mathcal{H}:\ \underline{M}(f, \mathbf{x}, \gamma) \leq 0} (1-\eta)\phi_\rho(-\overline{M}(f, \mathbf{x}, \gamma)) \\
&\geq \eta + (1-\eta)\phi_\rho(M_{\mathbf{x}}).
\end{aligned}$$

On the other hand, for any $\mathbf{x} \in \mathcal{X}_2$ and $\epsilon > 0$,

$$\begin{aligned}
\mathcal{C}_{\tilde{\phi}_\rho}(f^\epsilon_{\mathbf{x}}, \mathbf{x}, \eta) &= \eta\phi_\rho(\underline{M}(f^\epsilon_{\mathbf{x}}, \mathbf{x}, \gamma)) + (1-\eta)\phi_\rho(-\overline{M}(f^\epsilon_{\mathbf{x}}, \mathbf{x}, \gamma)) \\
&< \eta\phi_\rho(M_{\mathbf{x}}) + \epsilon + (1-\eta).
\end{aligned}$$

Since $\eta > \frac{1}{2}$ and $M_{\mathbf{x}} > 0$, we have

$$\begin{aligned}
&\inf_{f \in \mathcal{H}:\ \underline{M}(f, \mathbf{x}, \gamma) \leq 0} \mathcal{C}_{\tilde{\phi}_\rho}(f, \mathbf{x}, \eta) - \mathcal{C}_{\tilde{\phi}_\rho}(f^\epsilon_{\mathbf{x}}, \mathbf{x}, \eta) \\
&> [\eta + (1-\eta)\phi_\rho(M_{\mathbf{x}})] - [\eta\phi_\rho(M_{\mathbf{x}}) + \epsilon + (1-\eta)] \\
&= (2\eta - 1)(1 - \phi_\rho(M_{\mathbf{x}})) - \epsilon \\
&> 0,
\end{aligned}$$

where we take $0 < \epsilon < (2\eta - 1)(1 - \phi_\rho(M_{\mathbf{x}}))$.

Therefore for any $\eta \in (\frac{1}{2}, 1]$ and $\mathbf{x} \in \mathcal{X}_2$, there exists $0 < \epsilon < (2\eta - 1)(1 - \phi_\rho(M_{\mathbf{x}}))$ such that

$$\inf_{f \in \mathcal{H}} \mathcal{C}_{\tilde{\phi}_\rho}(f, \mathbf{x}, \eta) \leq \mathcal{C}_{\tilde{\phi}_\rho}(f^\epsilon_{\mathbf{x}}, \mathbf{x}, \eta) < \inf_{f \in \mathcal{H}:\ \underline{M}(f, \mathbf{x}, \gamma) \leq 0} \mathcal{C}_{\tilde{\phi}_\rho}(f, \mathbf{x}, \eta). \tag{29}$$

- When $\eta \in [0, \frac{1}{2})$, since ϕ_ρ is non-increasing, for any $\mathbf{x} \in \mathcal{X}_2$,

$$\begin{aligned}
\inf_{f \in \mathcal{H}:\ \overline{M}(f, \mathbf{x}, \gamma) \geq 0} \mathcal{C}_{\tilde{\phi}_\rho}(f, \mathbf{x}, \eta) &= \inf_{f \in \mathcal{H}:\ \overline{M}(f, \mathbf{x}, \gamma) \geq 0} \eta\phi_\rho(\underline{M}(f, \mathbf{x}, \gamma)) + (1-\eta)\phi_\rho(-\overline{M}(f, \mathbf{x}, \gamma)) \\
&= 1 - \eta + \inf_{f \in \mathcal{H}:\ \overline{M}(f, \mathbf{x}, \gamma) \geq 0} \eta\phi_\rho(\underline{M}(f, \mathbf{x}, \gamma)) \\
&\geq 1 - \eta + \eta\phi_\rho(M_{\mathbf{x}})
\end{aligned}$$

On the other hand, for any $\mathbf{x} \in \mathcal{X}_2$ and $\epsilon > 0$,

$$\begin{aligned}
\mathcal{C}_{\tilde{\phi}_\rho}(-f^\epsilon_{\mathbf{x}}, \mathbf{x}, \eta) &= \eta\phi_\rho(\underline{M}(-f^\epsilon_{\mathbf{x}}, \mathbf{x}, \gamma)) + (1-\eta)\phi_\rho(-\overline{M}(-f^\epsilon_{\mathbf{x}}, \mathbf{x}, \gamma)) \\
&= \eta + (1-\eta)\phi_\rho(\underline{M}(f^\epsilon_{\mathbf{x}}, \mathbf{x}, \gamma)) \\
&< \eta + (1-\eta)\phi_\rho(M_{\mathbf{x}}) + \epsilon
\end{aligned}$$

Since $\eta < \frac{1}{2}$ and $M_{\mathbf{x}} > 0$, we have

$$\begin{aligned}
&\inf_{f \in \mathcal{H}:\ \overline{M}(f, \mathbf{x}, \gamma) \geq 0} \mathcal{C}_{\tilde{\phi}_\rho}(f, \mathbf{x}, \eta) - \mathcal{C}_{\tilde{\phi}_\rho}(-f^\epsilon_{\mathbf{x}}, \mathbf{x}, \eta) \\
&> [1 - \eta + \eta\phi_\rho(M_{\mathbf{x}})] - [\eta + (1-\eta)\phi_\rho(M_{\mathbf{x}}) + \epsilon] \\
&= (1 - 2\eta)(1 - \phi_\rho(M_{\mathbf{x}})) - \epsilon \\
&> 0
\end{aligned}$$

where we take $0 < \epsilon < (1 - 2\eta)(1 - \phi_\rho(M_{\mathbf{x}}))$.

Therefore for any $\eta \in [0, \frac{1}{2})$ and $\mathbf{x} \in \mathcal{X}_2$, there exists $0 < \epsilon < (1 - 2\eta)(1 - \phi_\rho(M_{\mathbf{x}}))$ such that

$$\inf_{f \in \mathcal{H}} \mathcal{C}_{\tilde{\phi}_\rho}(f, \mathbf{x}, \eta) \leq \mathcal{C}_{\tilde{\phi}_\rho}(-f^\epsilon_{\mathbf{x}}, \mathbf{x}, \eta) < \inf_{f \in \mathcal{H}:\ \overline{M}(f, \mathbf{x}, \gamma) \geq 0} \mathcal{C}_{\tilde{\phi}_\rho}(f, \mathbf{x}, \eta). \tag{30}$$

To sum up, by (28), (29) and (30), we conclude that $\tilde{\phi}_\rho$ is $\mathcal{H}$-calibrated with respect to ℓ_γ. □

E.5 Proof of Theorem 13 and Corollary 14

As shown by Awasthi et al. (2020), for $f \in \mathcal{H}_g$, the adversarial $0/1$ loss has the equivalent form

$$\ell_\gamma(f, \mathbf{x}, y) = \mathbb{1}_{\inf_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma}(yg(\mathbf{w}\cdot\mathbf{x}')+by)\le 0} = \mathbb{1}_{yg(\mathbf{w}\cdot\mathbf{x}-\gamma y\|\mathbf{w}\|)+by\le 0} = \mathbb{1}_{yg(\mathbf{w}\cdot\mathbf{x}-\gamma y)+by\le 0}\,. \tag{31}$$

The proofs of Theorem 13 will closely follow the proofs of Theorem 12 and Theorem 16. We will first prove Lemma 29 and Lemma 30 analogous to Lemma 27 and Lemma 28 respectively. Without loss of generality, assume that g is continuous and satisfies $g(-1-\gamma)+G>0$, $g(1+\gamma)-G<0$. Then observe that $g(-\gamma)+G>0$, $g(\gamma)-G<0$ since g is non-decreasing.

Lemma 29. *For a surrogate loss ℓ and hypothesis set $\mathcal{H}_g$, the calibration function of losses (ℓ, ℓ_γ) is*

$$\delta_{\max}(\epsilon, \mathbf{x}, \eta) = \begin{cases} +\infty & \text{if } \epsilon > \max\{\eta, 1-\eta\}, \\ \inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0\le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b} \Delta\mathcal{C}_{\ell,\mathcal{H}_g}(f, \mathbf{x}, \eta) & \text{if } |2\eta-1| < \epsilon \le \max\{\eta, 1-\eta\}, \\ \inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0\le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b \text{ or } (2\eta-1)[g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b]\le 0} \Delta\mathcal{C}_{\ell,\mathcal{H}_g}(f, \mathbf{x}, \eta) & \text{if } \epsilon \le |2\eta-1|. \end{cases}$$

Proof. As with the proof of Lemma 27, we first characterize the inner ℓ-risk and minimal inner ℓ_γ-risk for $\mathcal{H}_g$. By the definition of inner risk (4) and equivalent form of adversarial 0-1 loss ℓ_γ for $\mathcal{H}_g$ (31), the inner ℓ_γ-risk is

$$\begin{aligned} \mathcal{C}_{\ell_\gamma}(f, \mathbf{x}, \eta) &= \eta\mathbb{1}_{g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0} + (1-\eta)\mathbb{1}_{g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b\ge 0} \\ &= \begin{cases} 1 & \text{if } g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b \le 0 \le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b, \\ \eta & \text{if } g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b < 0, \\ 1-\eta & \text{if } g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b > 0. \end{cases} \end{aligned}$$

where we used the fact that g is non-decreasing and $g(\mathbf{w}\cdot\mathbf{x}-\gamma) \le g(\mathbf{w}\cdot\mathbf{x}+\gamma)$. Note for any $\mathbf{x}\in\mathcal{X}$, $\mathbf{w}\cdot\mathbf{x}\in[-\|\mathbf{x}\|, \|\mathbf{x}\|]$. Thus we have $g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b \in [g(-\|\mathbf{x}\|-\gamma)-G, g(\|\mathbf{x}\|-\gamma)+G]$ and $g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b \in [g(-\|\mathbf{x}\|+\gamma)-G, g(\|\mathbf{x}\|+\gamma)+G]$ since g is non-decreasing. By the fact that $g(-\gamma)+G>0$ and $g(\gamma)-G<0$, we obtain the minimal inner ℓ_γ-risk, which is for any $\mathbf{x}\in\mathcal{X}$,

$$\mathcal{C}^*_{\ell_\gamma,\mathcal{H}_g}(\mathbf{x}, \eta) = \min\{\eta, 1-\eta\}\,.$$

As with the derivation of $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(f, \mathbf{x}, \eta)$ (20), we derive $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}_g}(f, \mathbf{x}, \eta)$ as follows. By the observation (11), for any $\mathbf{x}\in\mathcal{X}$, for $f\in\mathcal{H}_g$ such that $g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b \le 0 \le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b$, $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}_g}(f, \mathbf{x}, \eta) = 1-\min\{\eta, 1-\eta\} = \max\{\eta, 1-\eta\}$; for $f\in\mathcal{H}_g$ such that $g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b < 0$, $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}_g}(f, \mathbf{x}, \eta) = \eta-\min\{\eta, 1-\eta\} = \max\{0, 2\eta-1\} = |2\eta-1|\mathbb{1}_{(2\eta-1)[g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b]\le 0}$ since $g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b \le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b < 0$; for $f\in\mathcal{H}_g$ such that $g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b > 0$, $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}_g}(f, \mathbf{x}, \eta) = 1-\eta-\min\{\eta, 1-\eta\} = \max\{0, 1-2\eta\} = |2\eta-1|\mathbb{1}_{(2\eta-1)[g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b]\le 0}$ since $g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b > 0$. Therefore,

$$\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}_g}(f, \mathbf{x}, \eta) = \begin{cases} \max\{\eta, 1-\eta\} & \text{if } g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b \le 0 \le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b\,, \\ |2\eta-1|\mathbb{1}_{(2\eta-1)[g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b]\le 0} & \text{if } g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b < 0 \text{ or } g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b > 0\,. \end{cases}$$

By (6), for a fixed $\eta\in[0,1]$ and $\mathbf{x}\in\mathcal{X}$, the calibration function of losses (ℓ, ℓ_γ) given $\mathcal{H}_g$ is

$$\delta_{\max}(\epsilon, \mathbf{x}, \eta) = \inf_{f\in\mathcal{H}_g}\left\{\Delta\mathcal{C}_{\ell,\mathcal{H}_g}(f, \mathbf{x}, \eta) \mid \Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}_g}(f, \mathbf{x}, \eta) \ge \epsilon\right\}.$$

As with the proof of Lemma 27, we then make use of the observation (11) for deriving the the calibration function. By the observation (11), if $\epsilon > \max\{\eta, 1-\eta\}$, then for all $f\in\mathcal{H}_g$, $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}_g}(f, \mathbf{x}, \eta) < \epsilon$, which implies that $\delta_{\max}(\epsilon, \mathbf{x}, \eta) = \infty$; if $|2\eta-1| < \epsilon \le \max\{\eta, 1-\eta\}$, then $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}_g}(f, \mathbf{x}, \eta) \ge \epsilon$ if and only if $g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b \le 0 \le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b$, which leads to

$$\delta_{\max}(\epsilon, \mathbf{x}, \eta) = \inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0\le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b} \Delta\mathcal{C}_{\ell,\mathcal{H}_g}(f, \mathbf{x}, \eta);$$

if $\epsilon \le |2\eta-1|$, then $\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}_g}(f, \mathbf{x}, \eta) \ge \epsilon$ if and only if $g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b \le 0 \le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b$ or $(2\eta-1)[g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b] \le 0$, which leads to

$$\delta_{\max}(\epsilon, \mathbf{x}, \eta) = \inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0\le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b \text{ or } (2\eta-1)[g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b]\le 0} \Delta\mathcal{C}_{\ell,\mathcal{H}_g}(f, \mathbf{x}, \eta).$$

□

Lemma 30. *Let ℓ be a surrogate loss function. Then ℓ is $\mathcal{H}_g$-calibrated with respect to ℓ_γ if and only if for any $\mathbf{x} \in \mathcal{X}$,*

$$\inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0\le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b} \mathcal{C}_\ell(f,\mathbf{x},\frac{1}{2}) > \inf_{f\in\mathcal{H}_g} \mathcal{C}_\ell(f,\mathbf{x},\frac{1}{2})\,, \textit{and}$$

$$\inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}_g} \mathcal{C}_\ell(f,\mathbf{x},\eta) \textit{ for all } \eta\in(\frac{1}{2},1]\,, \textit{and}$$

$$\inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b\ge 0} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}_g} \mathcal{C}_\ell(f,\mathbf{x},\eta) \textit{ for all } \eta\in[0,\frac{1}{2})\,.$$

Proof. As the proof of Lemma 28 first makes use of Lemma 27 and Proposition 4, we also first make use of Lemma 29 and Proposition 4 in the following proof. Let $\delta_{\max}$ be the calibration function of (ℓ,ℓ_γ) for hypothesis set $\mathcal{H}_g$. By Lemma 29,

$$\delta_{\max}(\epsilon,\mathbf{x},\eta) = \begin{cases} +\infty & \text{if } \epsilon > \max\{\eta,1-\eta\}, \\ \inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0\le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b} \Delta\mathcal{C}_{\ell,\mathcal{H}_g}(f,\mathbf{x},\eta) & \text{if } |2\eta-1|<\epsilon\le\max\{\eta,1-\eta\}, \\ \inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0\le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b \text{ or } (2\eta-1)[g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b]\le 0} \Delta\mathcal{C}_{\ell,\mathcal{H}_g}(f,\mathbf{x},\eta) & \text{if } \epsilon\le|2\eta-1|. \end{cases}$$

By Proposition 4, ℓ is $\mathcal{H}_g$-calibrated with respect to ℓ_γ if and only if its calibration function $\delta_{\max}$ satisfies $\delta_{\max}(\epsilon,\mathbf{x},\eta)>0$ for all $\mathbf{x}\in\mathcal{X}$, $\eta\in[0,1]$ and $\epsilon>0$. The following steps are similar to the steps in the proof of Lemma 28, where we analyze by considering three cases.
For $\eta=\frac{1}{2}$, we have for any $\mathbf{x}\in\mathcal{X}$,

$$\delta_{\max}(\epsilon,\mathbf{x},\frac{1}{2})>0 \text{ for all } \epsilon>0 \Leftrightarrow \inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0\le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b} \mathcal{C}_\ell(f,\mathbf{x},\frac{1}{2}) > \inf_{f\in\mathcal{H}_g} \mathcal{C}_\ell(f,\mathbf{x},\frac{1}{2}). \tag{32}$$

For $1\ge\eta>\frac{1}{2}$, we have $|2\eta-1|=2\eta-1$, $\max\{\eta,1-\eta\}=\eta$, and

$$\inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0\le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b \text{ or } (2\eta-1)[g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b]\le 0} \Delta\mathcal{C}_{\ell,\mathcal{H}_g}(f,\mathbf{x},\eta) = \inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0} \Delta\mathcal{C}_{\ell,\mathcal{H}_g}(f,\mathbf{x},\eta)\,.$$

Therefore, $\delta_{\max}(\epsilon,\mathbf{x},\frac{1}{2})>0$ for any $\mathbf{x}\in\mathcal{X}$, $\epsilon>0$ and $\eta\in(\frac{1}{2},1]$ if and only if for any $\mathbf{x}\in\mathcal{X}$,

$$\begin{cases} \inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0\le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}_g} \mathcal{C}_\ell(f,\mathbf{x},\eta) & \text{for all } \eta\in(\frac{1}{2},1] \text{ such that } 2\eta-1<\epsilon\le\eta, \\ \inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}_g} \mathcal{C}_\ell(f,\mathbf{x},\eta) & \text{for all } \eta\in(\frac{1}{2},1] \text{ such that } \epsilon\le 2\eta-1, \end{cases}$$

for all $\epsilon>0$, which is equivalent to for any $\mathbf{x}\in\mathcal{X}$,

$$\begin{cases} \inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0\le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}_g} \mathcal{C}_\ell(f,\mathbf{x},\eta) & \text{for all } \eta\in(\frac{1}{2},1] \text{ such that } \epsilon\le\eta<\frac{\epsilon+1}{2}, \\ \inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}_g} \mathcal{C}_\ell(f,\mathbf{x},\eta) & \text{for all } \eta\in(\frac{1}{2},1] \text{ such that } \frac{\epsilon+1}{2}\le\eta, \end{cases} \tag{33}$$

for all $\epsilon>0$. Observe that

$$\left\{\eta\in(\frac{1}{2},1]\,\middle|\,\epsilon\le\eta<\frac{\epsilon+1}{2},\epsilon>0\right\} = \left\{\frac{1}{2}<\eta\le 1\right\}\,, \text{ and}$$

$$\left\{\eta\in(\frac{1}{2},1]\,\middle|\,\frac{\epsilon+1}{2}\le\eta,\epsilon>0\right\} = \left\{\frac{1}{2}<\eta\le 1\right\}\,, \text{ and}$$

$$\inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0\le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b} \mathcal{C}_\ell(f,\mathbf{x},\eta) \ge \inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0} \mathcal{C}_\ell(f,\mathbf{x},\eta) \text{ for all } \eta\,.$$

Therefore, we reduce the above condition (33) as for any $\mathbf{x}\in\mathcal{X}$,

$$\inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0} \mathcal{C}_\ell(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}_g} \mathcal{C}_\ell(f,\mathbf{x},\eta) \text{ for all } \eta\in(\frac{1}{2},1]\,. \tag{34}$$

For $\frac{1}{2}>\eta\ge 0$, we have $|2\eta-1|=1-2\eta$, $\max\{\eta,1-\eta\}=1-\eta$, and

$$\inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b\le 0\le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b \text{ or } (2\eta-1)[g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b]\le 0} \Delta\mathcal{C}_{\ell,\mathcal{H}_g}(f,\mathbf{x},\eta) = \inf_{f\in\mathcal{H}_g:\, g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b\ge 0} \Delta\mathcal{C}_{\ell,\mathcal{H}_g}(f,\mathbf{x},\eta)\,.$$

Therefore, $\delta_{\max}(\epsilon, \mathbf{x}, \frac{1}{2}) > 0$ for any $\mathbf{x} \in \mathcal{X}$, $\epsilon > 0$ and $\eta \in [0, \frac{1}{2})$ if and only if for any $\mathbf{x} \in \mathcal{X}$,

$$\begin{cases} \inf\limits_{f \in \mathcal{H}_g:\ g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b \le 0 \le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b} \mathcal{C}_\ell(f, \mathbf{x}, \eta) > \inf\limits_{f \in \mathcal{H}_g} \mathcal{C}_\ell(f, \mathbf{x}, \eta) & \text{for all } \eta \in [0, \frac{1}{2}) \text{ such that } 1 - 2\eta < \epsilon \le 1 - \eta, \\ \inf\limits_{f \in \mathcal{H}_g:\ g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b \ge 0} \mathcal{C}_\ell(f, \mathbf{x}, \eta) > \inf\limits_{f \in \mathcal{H}_g} \mathcal{C}_\ell(f, \mathbf{x}, \eta) & \text{for all } \eta \in [0, \frac{1}{2}) \text{ such that } \epsilon \le 1 - 2\eta, \end{cases}$$

for all $\epsilon > 0$, which is equivalent to for any $\mathbf{x} \in \mathcal{X}$,

$$\begin{cases} \inf\limits_{f \in \mathcal{H}_g:\ g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b \le 0 \le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b} \mathcal{C}_\ell(f, \mathbf{x}, \eta) > \inf\limits_{f \in \mathcal{H}_g} \mathcal{C}_\ell(f, \mathbf{x}, \eta) & \text{for all } \eta \in [0, \frac{1}{2}) \text{ such that } \frac{1-\epsilon}{2} < \eta \le 1 - \epsilon, \\ \inf\limits_{f \in \mathcal{H}_g:\ g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b \ge 0} \mathcal{C}_\ell(f, \mathbf{x}, \eta) > \inf\limits_{f \in \mathcal{H}_g} \mathcal{C}_\ell(f, \mathbf{x}, \eta) & \text{for all } \eta \in [0, \frac{1}{2}) \text{ such that } \eta \le \frac{1-\epsilon}{2}, \end{cases} \tag{35}$$

for all $\epsilon > 0$. Observe that

$$\left\{\eta \in [0, \frac{1}{2}) \middle| \frac{1-\epsilon}{2} < \eta \le 1 - \epsilon, \epsilon > 0\right\} = \left\{0 \le \eta < \frac{1}{2}\right\}, \text{ and}$$

$$\left\{\eta \in [0, \frac{1}{2}) \middle| \eta \le \frac{1-\epsilon}{2}, \epsilon > 0\right\} = \left\{0 \le \eta < \frac{1}{2}\right\}, \text{ and}$$

$$\inf_{f \in \mathcal{H}_g:\ g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b \le 0 \le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b} \mathcal{C}_\ell(f, \mathbf{x}, \eta) \ge \inf_{f \in \mathcal{H}_g:\ g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b \ge 0} \mathcal{C}_\ell(f, \mathbf{x}, \eta) \text{ for all } \eta.$$

Therefore we reduce the above condition (35) as for any $\mathbf{x} \in \mathcal{X}$,

$$\inf_{f \in \mathcal{H}_g:\ g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b \ge 0} \mathcal{C}_\ell(f, \mathbf{x}, \eta) > \inf_{f \in \mathcal{H}_g} \mathcal{C}_\ell(f, \mathbf{x}, \eta) \text{ for all } \eta \in [0, \frac{1}{2}). \tag{36}$$

To sum up, by (32), (34) and (36), we conclude the proof. □

Theorem 13. *Let g be a non-decreasing and continuous function such that $g(1+\gamma) < G$ and $g(-1-\gamma) > -G$ for some $G \ge 0$. Let a margin-based loss ϕ be bounded, continuous, non-increasing, and satisfy the property that $\bar{\mathcal{C}}_\phi(t, \eta)$ is quasi-concave in $t \in \mathbb{R}$ for all $\eta \in [0, 1]$. Assume that $\phi(g(-t) - G) > \phi(G - g(-t))$ and $g(-t) + g(t) \ge 0$ for any $0 \le t \le 1$. Then ϕ is $\mathcal{H}_g$-calibrated with respect to ℓ_γ if and only if for any $0 \le t \le 1$,*

$$\phi(G - g(-t)) + \phi(g(-t) - G) = \phi(g(t) + G) + \phi(-g(t) - G)$$

$$\textit{and} \quad \min\left\{\phi(\overline{A}(t)) + \phi(-\overline{A}(t)), \phi(\underline{A}(t)) + \phi(-\underline{A}(t))\right\} > \phi(G - g(-t)) + \phi(g(-t) - G),$$

where $\overline{A}(t) = \max_{s \in [-t,t]} g(s) - g(s-\gamma)$ and $\underline{A}(t) = \min_{s \in [-t,t]} g(s) - g(s+\gamma)$.

Proof. By Lemma 30, ϕ is $\mathcal{H}_g$-calibrated with respect to ℓ_γ if and only if for any $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned} &\inf_{f \in \mathcal{H}_g:\ g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b \le 0 \le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b} \mathcal{C}_\phi(f, \mathbf{x}, \frac{1}{2}) > \inf_{f \in \mathcal{H}_g} \mathcal{C}_\phi(f, \mathbf{x}, \frac{1}{2}), \text{and} \\ &\inf_{f \in \mathcal{H}_g:\ g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b \le 0} \mathcal{C}_\phi(f, \mathbf{x}, \eta) > \inf_{f \in \mathcal{H}_g} \mathcal{C}_\phi(f, \mathbf{x}, \eta) \text{ for all } \eta \in (\frac{1}{2}, 1], \text{and} \\ &\inf_{f \in \mathcal{H}_g:\ g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b \ge 0} \mathcal{C}_\phi(f, \mathbf{x}, \eta) > \inf_{f \in \mathcal{H}_g} \mathcal{C}_\phi(f, \mathbf{x}, \eta) \text{ for all } \eta \in [0, \frac{1}{2}). \end{aligned} \tag{37}$$

By the definition of inner risk (4), the inner ϕ-risk is

$$\mathcal{C}_\phi(f, \mathbf{x}, \eta) = \eta\phi(f(\mathbf{x})) + (1-\eta)\phi(-f(\mathbf{x})).$$

and $f(\mathbf{x}) = g(\mathbf{w}\cdot\mathbf{x}) + b \in [g(-\|\mathbf{x}\|) - G, g(\|\mathbf{x}\|) + G]$ when $f \in \mathcal{H}_g$ since g is continuous and non-decreasing. Specifically, by the assumption that $g(-1-\gamma) + G > 0$, $g(1+\gamma) - G < 0$, when $f \in \{f \in \mathcal{H}_g:\ g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b \le 0 \le g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b\}$, $f(\mathbf{x}) = g(\mathbf{w}\cdot\mathbf{x})+b \in [\min_{-\|\mathbf{x}\| \le s \le \|\mathbf{x}\|} g(s) - g(s+\gamma), \max_{-\|\mathbf{x}\| \le s \le \|\mathbf{x}\|} g(s) - g(s-\gamma)]$; when $f \in \{f \in \mathcal{H}_g:\ g(\mathbf{w}\cdot\mathbf{x}-\gamma)+b \le 0\}$, $f(\mathbf{x}) = g(\mathbf{w}\cdot\mathbf{x})+b \in [g(-\|\mathbf{x}\|) - G, \max_{-\|\mathbf{x}\| \le s \le \|\mathbf{x}\|} g(s) - g(s-\gamma)]$; when $f \in \{f \in \mathcal{H}_g:\ g(\mathbf{w}\cdot\mathbf{x}+\gamma)+b \ge 0\}$, $f(\mathbf{x}) = g(\mathbf{w}\cdot\mathbf{x})+b \in [\min_{-\|\mathbf{x}\| \le s \le \|\mathbf{x}\|} g(s) - g(s+\gamma), g(\|\mathbf{x}\|) + G]$. For convenience, we denote

$\overline{A}(t) = \max_{-t\le s\le t} g(s) - g(s-\gamma) \ge 0$ and $\underline{A}(t) = \min_{-t\le s\le t} g(s) - g(s+\gamma) \le 0$ for any $0 \le t \le 1$. Therefore, for any $\mathbf{x} \in \mathcal{X}$, (37) is equivalent to

$$\begin{aligned}
&\inf_{\underline{A}(\|\mathbf{x}\|)\le t\le \overline{A}(\|\mathbf{x}\|)} \bar{\mathcal{C}}_\phi(t, \frac{1}{2}) > \inf_{g(-\|\mathbf{x}\|)-G\le t\le g(\|\mathbf{x}\|)+G} \bar{\mathcal{C}}_\phi(t, \frac{1}{2}), \text{and} \\
&\inf_{g(-\|\mathbf{x}\|)-G\le t\le \overline{A}(\|\mathbf{x}\|)} \bar{\mathcal{C}}_\phi(t, \eta) > \inf_{g(-\|\mathbf{x}\|)-G\le t\le g(\|\mathbf{x}\|)+G} \bar{\mathcal{C}}_\phi(t, \eta) \text{ for all } \eta \in (\frac{1}{2}, 1], \text{and} \qquad (38) \\
&\inf_{\underline{A}(\|\mathbf{x}\|)\le t\le g(\|\mathbf{x}\|)+G} \bar{\mathcal{C}}_\phi(t, \eta) > \inf_{g(-\|\mathbf{x}\|)-G\le t\le g(\|\mathbf{x}\|)+G} \bar{\mathcal{C}}_\phi(t, \eta) \text{ for all } \eta \in [0, \frac{1}{2}).
\end{aligned}$$

Suppose that ϕ is $\mathcal{H}_g$-calibrated with respect to ℓ_γ. Since for $\eta \in [0, \frac{1}{2})$,

$$\inf_{\underline{A}(\|\mathbf{x}\|)\le t\le g(\|\mathbf{x}\|)+G} \bar{\mathcal{C}}_\phi(t, \eta) = \min\{\bar{\mathcal{C}}_\phi(\underline{A}(\|\mathbf{x}\|), \eta), \bar{\mathcal{C}}_\phi(g(\|\mathbf{x}\|) + G, \eta)\} \qquad \text{(Part 2 of Lemma 26)}$$

$$\inf_{g(-\|\mathbf{x}\|)-G\le t\le g(\|\mathbf{x}\|)+G} \bar{\mathcal{C}}_\phi(t, \eta) = \min\{\bar{\mathcal{C}}_\phi(g(-\|\mathbf{x}\|) - G, \eta), \bar{\mathcal{C}}_\phi(g(\|\mathbf{x}\|) + G, \eta)\} \qquad \text{(Part 2 of Lemma 26)}$$

we have $\bar{\mathcal{C}}_\phi(g(-\|\mathbf{x}\|) - G, \eta) < \bar{\mathcal{C}}_\phi(g(\|\mathbf{x}\|) + G, \eta)$ for any $\mathbf{x} \in \mathcal{X}$, otherwise

$$\inf_{\underline{A}(\|\mathbf{x}\|)\le t\le g(\|\mathbf{x}\|)+G} \bar{\mathcal{C}}_\phi(t, \eta) \le \bar{\mathcal{C}}_\phi(g(\|\mathbf{x}\|) + G, \eta) = \inf_{g(-\|\mathbf{x}\|)-G\le t\le g(\|\mathbf{x}\|)+G} \bar{\mathcal{C}}_\phi(t, \eta).$$

By Part 8 of Lemma 26, $\phi(G-g(-t))+\phi(g(-t)-G) = \phi(g(t)+G)+\phi(-g(t)-G)$ for all $0 \le t \le 1$. Also, for any $0 \le t \le 1$,

$$\begin{aligned}
&\frac{1}{2}\min\{\phi(\overline{A}(t)) + \phi(-\overline{A}(t)), \phi(\underline{A}(t)) + \phi(-\underline{A}(t))\} \\
=& \inf_{\underline{A}(t)\le t\le \overline{A}(t)} \bar{\mathcal{C}}_\phi(t, \frac{1}{2}) && \text{(Part 2 of Lemma 26)} \\
>& \inf_{g(-t)-G\le t\le g(t)+G} \bar{\mathcal{C}}_\phi(t, \frac{1}{2}) && (38) \\
=& \frac{1}{2}\min\{\phi(G - g(-t)) + \phi(g(-t) - G), \phi(g(t) + G) + \phi(-g(t) - G)\} && \text{(Part 2 of Lemma 26)} \\
=& \frac{1}{2}(\phi(G - g(-t)) + \phi(g(-t) - G))
\end{aligned}$$

Now for the other direction, assume that for any $0 \le t \le 1$,

$$\begin{aligned}
&\phi(G - g(-t)) + \phi(g(-t) - G) = \phi(g(t) + G) + \phi(-g(t) - G) \\
\text{and} \quad &\min\{\phi(\overline{A}(t)) + \phi(-\overline{A}(t)), \phi(\underline{A}(t)) + \phi(-\underline{A}(t))\} > \phi(G - g(-t)) + \phi(g(-t) - G).
\end{aligned}$$

Then for $\eta = \frac{1}{2}$ and any $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
&\inf_{\underline{A}(\|\mathbf{x}\|)\le t\le \overline{A}(\|\mathbf{x}\|)} \bar{\mathcal{C}}_\phi(t, \frac{1}{2}) \\
=& \frac{1}{2}\min\{\phi(\overline{A}(\|\mathbf{x}\|)) + \phi(-\overline{A}(\|\mathbf{x}\|)), \phi(\underline{A}(\|\mathbf{x}\|)) + \phi(-\underline{A}(\|\mathbf{x}\|))\} && \text{(Part 2 of Lemma 26)} \\
>& \frac{1}{2}(\phi(G - g(-\|\mathbf{x}\|)) + \phi(g(-\|\mathbf{x}\|) - G)) && \text{(by assumption)} \\
=& \frac{1}{2}\min\{\phi(G - g(-\|\mathbf{x}\|)) + \phi(g(-\|\mathbf{x}\|) - G), \phi(g(\|\mathbf{x}\|) + G) + \phi(-g(\|\mathbf{x}\|) - G)\} && \text{(by assumption)} \\
=& \inf_{g(-\|\mathbf{x}\|)-G\le t\le g(\|\mathbf{x}\|)+G} \bar{\mathcal{C}}_\phi(t, \frac{1}{2}). && \text{(Part 2 of Lemma 26)}
\end{aligned}$$

For $\eta \in (\frac{1}{2}, 1]$ and any $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
&\inf_{g(-\|\mathbf{x}\|)-G\le t\le \overline{A}(\|\mathbf{x}\|)} \bar{\mathcal{C}}_\phi(t, \eta) = \min\{\bar{\mathcal{C}}_\phi(g(-\|\mathbf{x}\|) - G, \eta), \bar{\mathcal{C}}_\phi(\overline{A}(\|\mathbf{x}\|), \eta)\} && \text{(Part 2 of Lemma 26)} \\
&\inf_{g(-\|\mathbf{x}\|)-G\le t\le g(\|\mathbf{x}\|)+G} \bar{\mathcal{C}}_\phi(t, \eta) = \min\{\bar{\mathcal{C}}_\phi(g(-\|\mathbf{x}\|) - G, \eta), \bar{\mathcal{C}}_\phi(g(\|\mathbf{x}\|) + G, \eta)\} && \text{(Part 2 of Lemma 26)} \\
&= \bar{\mathcal{C}}_\phi(g(\|\mathbf{x}\|) + G, \eta) && \text{(Part 7 of Lemma 26)}
\end{aligned}$$

Since ϕ is non-increasing, we have for any $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
&\phi(-g(\|\mathbf{x}\|) - G) - \phi(g(\|\mathbf{x}\|) + G) + \phi(\overline{A}(\|\mathbf{x}\|)) - \phi(-\overline{A}(\|\mathbf{x}\|)) \\
\geq &\phi(-g(\|\mathbf{x}\|) - G) - \phi(g(\|\mathbf{x}\|) + G) + \phi(g(\|\mathbf{x}\|) + G) - \phi(-g(\|\mathbf{x}\|) - G) \\
= &0.
\end{aligned}$$

Then for $\eta \in (\frac{1}{2}, 1]$ and any $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
&\bar{\mathcal{C}}_\phi(\overline{A}(\|\mathbf{x}\|), \eta) - \bar{\mathcal{C}}_\phi(g(\|\mathbf{x}\|) + G, \eta) \\
=&(\phi(\overline{A}(\|\mathbf{x}\|)) - \phi(-\overline{A}(\|\mathbf{x}\|)) + \phi(-g(\|\mathbf{x}\|) - G) - \phi(g(\|\mathbf{x}\|) + G))\eta + \phi(-\overline{A}(\|\mathbf{x}\|)) - \phi(-g(\|\mathbf{x}\|) - G) \\
\geq&(\phi(\overline{A}(\|\mathbf{x}\|)) - \phi(-\overline{A}(\|\mathbf{x}\|)) + \phi(-g(\|\mathbf{x}\|) - G) - \phi(g(\|\mathbf{x}\|) + G))\frac{1}{2} + \phi(-\overline{A}(\|\mathbf{x}\|)) - \phi(-g(\|\mathbf{x}\|) - G) \\
=&\frac{1}{2}(\phi(\overline{A}(\|\mathbf{x}\|)) - \phi(-\overline{A}(\|\mathbf{x}\|)) - \phi(-g(\|\mathbf{x}\|) - G) - \phi(g(\|\mathbf{x}\|) + G)) \\
>&0.
\end{aligned}$$

In addition, by Part 7 of Lemma 26, for all $\eta \in (\frac{1}{2}, 1]$ and any $\mathbf{x} \in \mathcal{X}$, $\bar{\mathcal{C}}_\phi(g(-\|\mathbf{x}\|) - G, \eta) - \bar{\mathcal{C}}_\phi(g(\|\mathbf{x}\|) + G, \eta) > 0$. As a result, for $\eta \in (\frac{1}{2}, 1]$ and any $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
&\inf_{g(-\|\mathbf{x}\|)-G \leq t \leq \overline{A}(\|\mathbf{x}\|)} \bar{\mathcal{C}}_\phi(t, \eta) - \inf_{g(-\|\mathbf{x}\|)-G \leq t \leq g(\|\mathbf{x}\|)+G} \bar{\mathcal{C}}_\phi(t, \eta) \\
&= \min\left\{\bar{\mathcal{C}}_\phi(g(-\|\mathbf{x}\|) - G, \eta) - \bar{\mathcal{C}}_\phi(g(\|\mathbf{x}\|) + G, \eta), \bar{\mathcal{C}}_\phi(\overline{A}(\|\mathbf{x}\|), \eta) - \bar{\mathcal{C}}_\phi(g(\|\mathbf{x}\|) + G, \eta)\right\} \\
&> 0.
\end{aligned}$$

Finally, for $\eta \in [0, \frac{1}{2})$, by Part 8 of Lemma 26, we have $\bar{\mathcal{C}}_\phi(g(-\|\mathbf{x}\|) - G, \eta) < \bar{\mathcal{C}}_\phi(g(\|\mathbf{x}\|) + G, \eta)$ and

$$\begin{aligned}
&\inf_{\underline{A}(\|\mathbf{x}\|) \leq t \leq g(\|\mathbf{x}\|)+G} \bar{\mathcal{C}}_\phi(t, \eta) = \min\left\{\bar{\mathcal{C}}_\phi(\underline{A}(\|\mathbf{x}\|), \eta), \bar{\mathcal{C}}_\phi(g(\|\mathbf{x}\|) + G, \eta)\right\} && \text{(Part 2 of Lemma 26)} \\
&\inf_{g(-\|\mathbf{x}\|)-G \leq t \leq g(\|\mathbf{x}\|)+G} \bar{\mathcal{C}}_\phi(t, \eta) = \min\left\{\bar{\mathcal{C}}_\phi(g(-\|\mathbf{x}\|) - G, \eta), \bar{\mathcal{C}}_\phi(g(\|\mathbf{x}\|) + G, \eta)\right\} && \text{(Part 2 of Lemma 26)} \\
&= \bar{\mathcal{C}}_\phi(g(-\|\mathbf{x}\|) - G, \eta) && \text{(Part 8 of Lemma 26)}
\end{aligned}$$

Since $\phi(\underline{A}(\|\mathbf{x}\|)) + \phi(-\underline{A}(\|\mathbf{x}\|)) > \phi(G - g(-\|\mathbf{x}\|)) + \phi(g(-\|\mathbf{x}\|) - G)$ and ϕ is non-increasing, we have for any $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
&\phi(G - g(-\|\mathbf{x}\|)) - \phi(g(-\|\mathbf{x}\|) - G) + \phi(\underline{A}(\|\mathbf{x}\|)) - \phi(-\underline{A}(\|\mathbf{x}\|)) \\
=&\phi(G - g(-\|\mathbf{x}\|)) - \phi(-\underline{A}(\|\mathbf{x}\|)) + \phi(\underline{A}(\|\mathbf{x}\|)) - \phi(g(-\|\mathbf{x}\|) - G) \\
<&\phi(\underline{A}(\|\mathbf{x}\|)) - \phi(g(-\|\mathbf{x}\|) - G) + \phi(\underline{A}(\|\mathbf{x}\|)) - \phi(g(-\|\mathbf{x}\|) - G) \\
=&2[\phi(\underline{A}(\|\mathbf{x}\|)) - \phi(g(-\|\mathbf{x}\|) - G)] \\
\leq&0.
\end{aligned}$$

Then for $\eta \in [0, \frac{1}{2})$ and any $\mathbf{x} \in \mathcal{X}$.

$$\begin{aligned}
&\bar{\mathcal{C}}_\phi(\underline{A}(\|\mathbf{x}\|), \eta) - \bar{\mathcal{C}}_\phi(g(-\|\mathbf{x}\|) - G, \eta) \\
=&[\phi(G - g(-\|\mathbf{x}\|)) - \phi(g(-\|\mathbf{x}\|) - G) + \phi(\underline{A}(\|\mathbf{x}\|)) - \phi(-\underline{A}(\|\mathbf{x}\|))]\eta + \phi(-\underline{A}(\|\mathbf{x}\|)) - \phi(G - g(-\|\mathbf{x}\|)) \\
\geq&[\phi(G - g(-\|\mathbf{x}\|)) - \phi(g(-\|\mathbf{x}\|) - G) + \phi(\underline{A}(\|\mathbf{x}\|)) - \phi(-\underline{A}(\|\mathbf{x}\|))]\frac{1}{2} + \phi(-\underline{A}(\|\mathbf{x}\|)) - \phi(G - g(-\|\mathbf{x}\|)) \\
=&\frac{1}{2}[\phi(\underline{A}(\|\mathbf{x}\|)) + \phi(-\underline{A}(\|\mathbf{x}\|)) - \phi(g(-\|\mathbf{x}\|) - G) - \phi(G - g(-\|\mathbf{x}\|))] \\
>&0.
\end{aligned}$$

In addition, by Part 8 of Lemma 26, for all $\eta \in [0, \frac{1}{2})$ and any $\mathbf{x} \in \mathcal{X}$, $\bar{\mathcal{C}}_\phi(g(\|\mathbf{x}\|)+G, \eta) - \bar{\mathcal{C}}_\phi(g(-\|\mathbf{x}\|) - G, \eta) > 0$. As a result, for $\eta \in [0, \frac{1}{2})$ and any $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
&\inf_{\underline{A}(\|\mathbf{x}\|) \leq t \leq g(\|\mathbf{x}\|)+G} \bar{\mathcal{C}}_\phi(t, \eta) - \inf_{g(-\|\mathbf{x}\|)-G \leq t \leq g(\|\mathbf{x}\|)+G} \bar{\mathcal{C}}_\phi(t, \eta) \\
&= \min\left\{\bar{\mathcal{C}}_\phi(g(\|\mathbf{x}\|) + G, \eta) - \bar{\mathcal{C}}_\phi(g(-\|\mathbf{x}\|) - G, \eta), \bar{\mathcal{C}}_\phi(\underline{A}(\|\mathbf{x}\|), \eta) - \bar{\mathcal{C}}_\phi(g(-\|\mathbf{x}\|) - G, \eta)\right\} \\
&> 0.
\end{aligned}$$

□

Corollary 14. *Assume that $G > 1 + \gamma$. Let a margin-based loss ϕ be bounded, continuous, non-increasing, and satisfy the property that $\bar{C}_\phi(t, \eta)$ is quasi-concave in $t \in \mathbb{R}$ for all $\eta \in [0, 1]$. Assume that $\phi(-G) > \phi(G)$. Then ϕ is $\mathcal{H}_{\text{relu}}$-calibrated with respect to ℓ_γ if and only if for any $0 \le t \le 1$,*

$$\phi(G) + \phi(-G) = \phi(t + G) + \phi(-t - G) \quad \textit{and} \quad \phi(\gamma) + \phi(-\gamma) > \phi(G) + \phi(-G).$$

Proof. For hypothesis set $\mathcal{H}_{\text{relu}}$, that is, $g = (\cdot)_+$ in $\mathcal{H}_g$, we have

$$\overline{A}(t) = \max_{s \in [-t,t]} (s)_+ - (s - \gamma)_+ = \begin{cases} t, 0 \le t < \gamma, \\ \gamma, \gamma \le t \le 1. \end{cases} \qquad \text{and} \quad \underline{A}(t) = \min_{s \in [-t,t]} (s)_+ - (s + \gamma)_+ = -\gamma.$$

As a result, using the fact that $\phi(t) + \phi(-t) \ge \phi(\gamma) + \phi(-\gamma)$ when $0 \le t \le \gamma$ by Part 1 of Lemma 26, we conclude the proof by Theorem 13. □

E.6 Proof of Theorem 18

Theorem 18. *No continuous margin-based loss function ϕ is $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ. Furthermore, for any continuous and non-increasing margin-based loss ϕ, surrogates of the form*

$$\tilde{\phi}(f, \mathbf{x}, y) = \sup_{\mathbf{x}' : \|\mathbf{x} - \mathbf{x}'\| \le \gamma} \phi(y f(\mathbf{x}'))$$

are not $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ.

Proof. Let $\mathbf{x}$ follow the uniform distribution on the unit circle. Denote $\mathbf{x} = (\cos(\theta), \sin(\theta))^\top, \theta \in [0, 2\pi)$ and $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$, $\mathbf{w} = (\cos(t), \sin(t))^\top, t \in [0, 2\pi)$, $f \in \mathcal{H}_{\text{lin}} = \{\mathbf{x} \to \mathbf{w} \cdot \mathbf{x} \mid \|\mathbf{w}\|_2 = 1\}$. We set the label of a point $\mathbf{x}$ as follows: if $\theta \in (\sigma, \pi)$, where $\sigma \in (0, \pi)$, then set $y = -1$ with probability $\frac{3}{4}$ and $y = 1$ with probability $\frac{1}{4}$; if $\theta \in (0, \sigma)$ or $(\sigma + \pi, 2\pi)$, then set $y = 1$; if $\theta \in (\pi, \sigma + \pi)$, then set $y = -1$.

Let $\eta : \mathcal{X} \to [0, 1]$ be a measurable function such that $\eta(X) = \mathbb{P}(Y = 1 \mid X)$. For $\ell_\gamma(\tau) = \mathbb{1}_{\tau \le \gamma}$, we want to solve

$$\mathcal{R}^*_{\ell_\gamma, \mathcal{H}_{\text{lin}}} = \min_{f \in \mathcal{H}_{\text{lin}}} \mathcal{R}_{\ell_\gamma}(f) = \min_{f \in \mathcal{H}_{\text{lin}}} \mathbb{E}_X[\ell_\gamma(f(X))\eta + \ell_\gamma(-f(X))(1 - \eta)].$$

Let $\eta' : \Theta \to [0, 1]$ be a measurable function such that $\eta' = \mathbb{P}(Y = 1 | \Theta)$, $\Theta \sim \mathcal{U}(0, 2\pi)$. In our example, we have

$$\eta' = \begin{cases} \frac{1}{4} & \theta \in (\sigma, \pi) \ , \\ 1 & \theta \in (0, \sigma) \text{ or } \theta \in (\sigma + \pi, 2\pi) \ , \\ 0 & \theta \in (\pi, \sigma + \pi) \ . \end{cases}$$

Therefore we obtain

$$\begin{aligned}
\mathcal{R}^*_{\ell_\gamma, \mathcal{H}_{\text{lin}}} &= \min_{t \in [0, 2\pi)} \mathbb{E}_\Theta[\ell_\gamma(\cos(\Theta - t))\eta' + \ell_\gamma(-\cos(\Theta - t))(1 - \eta')] \\
&= \frac{1}{2\pi} \min_{t \in [0, 2\pi)} \int_\sigma^\pi \frac{1}{4}\ell_\gamma(\cos(\theta - t)) + \frac{3}{4}\ell_\gamma(-\cos(\theta - t))\, d\theta + \int_{\sigma - \pi}^\sigma \ell_\gamma(\cos(\theta - t))\, d\theta \\
&\quad + \int_{-\pi}^{\sigma - \pi} \ell_\gamma(-\cos(\theta - t))\, d\theta \\
&= \frac{1}{2\pi} \min_{t \in [0, 2\pi)} \int_\sigma^\pi \frac{1}{4}\ell_\gamma(\cos(\theta - t)) + \frac{3}{4}\ell_\gamma(-\cos(\theta - t)) d\theta + \int_{\sigma - \pi}^0 \ell_\gamma(\cos(\theta - t)) d\theta \\
&\quad + \int_0^\sigma \ell_\gamma(\cos(\theta - t)) d\theta + \int_0^\sigma \ell_\gamma(\cos(\theta - t)) d\theta \\
&= \frac{1}{2\pi} \min_{t \in [0, 2\pi)} \int_\sigma^\pi \frac{1}{4}\ell_\gamma(\cos(\theta - t))\, d\theta + \int_{\sigma - \pi}^0 \frac{3}{4}\ell_\gamma(\cos(\theta - t))\, d\theta + \int_{\sigma - \pi}^0 \ell_\gamma(\cos(\theta - t))\, d\theta \\
&\quad + \int_0^\sigma 2\ell_\gamma(\cos(\theta - t))\, d\theta \\
&= \frac{1}{2\pi} \min_{t \in [0, 2\pi)} \int_\sigma^\pi \frac{1}{4}\ell_\gamma(\cos(\theta - t))\, d\theta + \int_{\sigma - \pi}^0 \frac{7}{4}\ell_\gamma(\cos(\theta - t))\, d\theta + \int_0^\sigma \frac{7}{4}\ell_\gamma(\cos(\theta - t))\, d\theta
\end{aligned}$$

$$
\begin{aligned}
&+ \int_0^{\sigma} \frac{1}{4}\ell_\gamma(\cos(\theta - t))\, d\theta \\
&= \frac{1}{2\pi} \min_{t\in[0,2\pi)} \int_0^{\pi} \frac{1}{4}\ell_\gamma(\cos(\theta - t))\, d\theta + \int_{\sigma-\pi}^{\sigma} \frac{7}{4}\ell_\gamma(\cos(\theta - t))\, d\theta \qquad (39)\\
&= \frac{1}{2\pi} \min_{t\in[0,2\pi)} \int_0^{\pi} \frac{1}{4}\ell_\gamma(\cos(\theta - t))\, d\theta + \int_0^{\pi} \frac{7}{4}\ell_\gamma(-\cos(\theta - t + \sigma))\, d\theta \\
&= \frac{1}{2\pi} \min_{t\in[0,2\pi)} \int_{-t}^{\pi-t} \frac{1}{4}\ell_\gamma(\cos(\theta))\, d\theta + \frac{7}{4}\ell_\gamma(-\cos(\theta + \sigma))\, d\theta \\
&= \frac{1}{2\pi} \min_{t\in[0,2\pi)} \int_{-t}^{\pi-t} \frac{1}{4}\mathbb{1}_{\cos(\theta)\le\gamma} + \frac{7}{4}\mathbb{1}_{-\cos(\theta+\sigma)\le\gamma}\, d\theta\,.
\end{aligned}
$$

Take $\gamma = \cos(\frac{\sigma}{2}) \in (0,1)$. For $\sigma \in (0, \frac{\pi}{2}]$, we analyze six cases:

- When $-t \in [-\frac{3\sigma}{2}, -\frac{\sigma}{2}]$,

$$
\begin{aligned}
&\int_{-t}^{\pi-t} \frac{1}{4}\mathbb{1}_{\cos(\theta)\le\gamma} + \frac{7}{4}\mathbb{1}_{-\cos(\theta+\sigma)\le\gamma}\, d\theta \\
&= \int_{-t}^{-\frac{\sigma}{2}} \frac{1}{4} + \frac{7}{4}\, d\theta + \int_{-\frac{\sigma}{2}}^{\frac{\sigma}{2}} \frac{7}{4}\, d\theta + \int_{\frac{\sigma}{2}}^{-\frac{3\sigma}{2}+\pi} \frac{1}{4} + \frac{7}{4}\, d\theta + \int_{-\frac{3\sigma}{2}+\pi}^{\pi-t} \frac{1}{4}\, d\theta \\
&= 2\pi - \frac{23}{8}\sigma + \frac{7}{4}t \ge 2\pi - 2\sigma
\end{aligned}
$$

  where the equality is achieved when $t = \frac{\sigma}{2}$.

- When $-t \in [-\frac{\sigma}{2}, \frac{\sigma}{2}]$,

$$
\begin{aligned}
&\int_{-t}^{\pi-t} \frac{1}{4}\mathbb{1}_{\cos(\theta)\le\gamma} + \frac{7}{4}\mathbb{1}_{-\cos(\theta+\sigma)\le\gamma}\, d\theta \\
&= \int_{-t}^{\frac{\sigma}{2}} \frac{7}{4}\, d\theta + \int_{\frac{\sigma}{2}}^{-\frac{3\sigma}{2}+\pi} \frac{1}{4} + \frac{7}{4}\, d\theta + \int_{-\frac{3\sigma}{2}+\pi}^{-\frac{\sigma}{2}+\pi} \frac{1}{4}\, d\theta + \int_{-\frac{\sigma}{2}+\pi}^{\pi-t} \frac{1}{4} + \frac{7}{4}\, d\theta \\
&= 2\pi - \frac{15}{8}\sigma - \frac{1}{4}t \ge 2\pi - 2\sigma
\end{aligned}
$$

  where the equality is achieved when $t = \frac{\sigma}{2}$.

- When $-t \in [\frac{\sigma}{2}, -\frac{3\sigma}{2} + \pi]$,

$$
\begin{aligned}
&\int_{-t}^{\pi-t} \frac{1}{4}\mathbb{1}_{\cos(\theta)\le\gamma} + \frac{7}{4}\mathbb{1}_{-\cos(\theta+\sigma)\le\gamma}\, d\theta \\
&= \int_{-t}^{-\frac{3\sigma}{2}+\pi} \frac{1}{4} + \frac{7}{4}\, d\theta + \int_{-\frac{3\sigma}{2}+\pi}^{-\frac{\sigma}{2}+\pi} \frac{1}{4}\, d\theta + \int_{-\frac{\sigma}{2}+\pi}^{\pi-t} \frac{1}{4} + \frac{7}{4}\, d\theta \\
&= 2\pi - \frac{7}{4}\sigma\,.
\end{aligned}
$$

- When $-t \in [-\frac{3\sigma}{2} + \pi, -\frac{\sigma}{2} + \pi]$,

$$
\begin{aligned}
&\int_{-t}^{\pi-t} \frac{1}{4}\mathbb{1}_{\cos(\theta)\le\gamma} + \frac{7}{4}\mathbb{1}_{-\cos(\theta+\sigma)\le\gamma}\, d\theta \\
&= \int_{-t}^{-\frac{\sigma}{2}+\pi} \frac{1}{4}\, d\theta + \int_{-\frac{\sigma}{2}+\pi}^{\pi-t} \frac{1}{4} + \frac{7}{4}\, d\theta \\
&= \frac{\pi}{4} + \frac{7}{8}\sigma - \frac{7}{4}t \ge 2\pi - \frac{7}{4}\sigma
\end{aligned}
$$

  where the equality is achieved when $t = \frac{3\sigma}{2} - \pi$.

- When $-t \in [-\frac{\sigma}{2} + \pi, \frac{\sigma}{2} + \pi]$,

$$\begin{aligned}
&\int_{-t}^{\pi-t} \frac{1}{4}\mathbb{1}_{\cos(\theta)\le\gamma} + \frac{7}{4}\mathbb{1}_{-\cos(\theta+\sigma)\le\gamma}\, d\theta \\
&= \int_{-t}^{-\frac{\sigma}{2}+2\pi} \frac{1}{4} + \frac{7}{4}\, d\theta + \int_{-\frac{\sigma}{2}+2\pi}^{\pi-t} \frac{7}{4}\, d\theta \\
&= \frac{9\pi}{4} - \frac{1}{8}\sigma + \frac{1}{4}t \ge 2\pi - \frac{1}{4}\sigma
\end{aligned}$$

  where the equality is achieved when $t = -\frac{\sigma}{2} - \pi$.

- When $-t \in [\frac{\sigma}{2} + \pi, -\frac{3\sigma}{2} + 2\pi]$,

$$\begin{aligned}
&\int_{-t}^{\pi-t} \frac{1}{4}\mathbb{1}_{\cos(\theta)\le\gamma} + \frac{7}{4}\mathbb{1}_{-\cos(\theta+\sigma)\le\gamma}\, d\theta \\
&= \int_{-t}^{-\frac{\sigma}{2}+2\pi} \frac{1}{4} + \frac{7}{4}\, d\theta + \int_{-\frac{\sigma}{2}+2\pi}^{\frac{\sigma}{2}+2\pi} \frac{7}{4}\, d\theta + \int_{\frac{\sigma}{2}+2\pi}^{\pi-t} \frac{1}{4} + \frac{7}{4}\, d\theta \\
&= 2\pi - \frac{1}{4}\sigma\,.
\end{aligned}$$

Similarly for $\sigma \in [\frac{\pi}{2}, \pi)$, we analyze six cases:

- When $-t \in [-\frac{3\sigma}{2}, \frac{\sigma}{2} - \pi]$,

$$\begin{aligned}
&\int_{-t}^{\pi-t} \frac{1}{4}\mathbb{1}_{\cos(\theta)\le\gamma} + \frac{7}{4}\mathbb{1}_{-\cos(\theta+\sigma)\le\gamma}\, d\theta \\
&= \int_{-t}^{-\frac{\sigma}{2}} \frac{1}{4} + \frac{7}{4}\, d\theta + \int_{-\frac{\sigma}{2}}^{-\frac{3\sigma}{2}+\pi} \frac{7}{4}\, d\theta \\
&= \frac{7}{4}\pi - \frac{11}{4}\sigma + 2t \ge \frac{15}{4}\pi - \frac{15}{4}\sigma
\end{aligned}$$

  where the equality is achieved when $t = \pi - \frac{\sigma}{2}$.

- When $-t \in [\frac{\sigma}{2} - \pi, -\frac{\sigma}{2}]$,

$$\begin{aligned}
&\int_{-t}^{\pi-t} \frac{1}{4}\mathbb{1}_{\cos(\theta)\le\gamma} + \frac{7}{4}\mathbb{1}_{-\cos(\theta+\sigma)\le\gamma}\, d\theta \\
&= \int_{-t}^{-\frac{\sigma}{2}} \frac{1}{4} + \frac{7}{4}\, d\theta + \int_{-\frac{\sigma}{2}}^{-\frac{3\sigma}{2}+\pi} \frac{7}{4}\, d\theta + \int_{\frac{\sigma}{2}}^{\pi-t} \frac{1}{4}\, d\theta \\
&= 2\pi - \frac{23}{8}\sigma + \frac{7}{4}t \ge 2\pi - 2\sigma
\end{aligned}$$

  where the equality is achieved when $t = \frac{\sigma}{2}$.

- When $-t \in [-\frac{\sigma}{2}, -\frac{3\sigma}{2} + \pi]$,

$$\begin{aligned}
&\int_{-t}^{\pi-t} \frac{1}{4}\mathbb{1}_{\cos(\theta)\le\gamma} + \frac{7}{4}\mathbb{1}_{-\cos(\theta+\sigma)\le\gamma}\, d\theta \\
&= \int_{-t}^{-\frac{3\sigma}{2}+\pi} \frac{7}{4}\, d\theta + \int_{\frac{\sigma}{2}}^{-\frac{\sigma}{2}+\pi} \frac{1}{4}\, d\theta + \int_{-\frac{\sigma}{2}+\pi}^{\pi-t} \frac{1}{4} + \frac{7}{4}\, d\theta \\
&= 2\pi - \frac{15}{8}\sigma - \frac{1}{4}t \ge 2\pi - 2\sigma
\end{aligned}$$

  where the equality is achieved when $t = \frac{\sigma}{2}$.

- When $-t \in [-\frac{3\sigma}{2} + \pi, \frac{\sigma}{2}]$,

$$\begin{aligned}
&\int_{-t}^{\pi-t} \frac{1}{4}\mathbb{1}_{\cos(\theta)\le\gamma} + \frac{7}{4}\mathbb{1}_{-\cos(\theta+\sigma)\le\gamma}\, d\theta \\
&= \int_{\frac{\sigma}{2}}^{-\frac{\sigma}{2}+\pi} \frac{1}{4}\, d\theta + \int_{-\frac{\sigma}{2}+\pi}^{\pi-t} \frac{1}{4} + \frac{7}{4}\, d\theta \\
&= \frac{\pi}{4} + \frac{3}{4}\sigma - 2t \ge \frac{9}{4}\pi - \frac{9}{4}\sigma
\end{aligned}$$

  where the equality is achieved when $t = \frac{3\sigma}{2} - \pi$.

- When $-t \in [\frac{\sigma}{2}, -\frac{\sigma}{2} + \pi]$,

$$\begin{aligned}
&\int_{-t}^{\pi-t} \frac{1}{4}\mathbb{1}_{\cos(\theta)\le\gamma} + \frac{7}{4}\mathbb{1}_{-\cos(\theta+\sigma)\le\gamma}\, d\theta \\
&= \int_{-t}^{-\frac{\sigma}{2}+\pi} \frac{7}{4}\, d\theta + \int_{-\frac{\sigma}{2}+\pi}^{\pi-t} \frac{1}{4} + \frac{7}{4}\, d\theta \\
&= \frac{7\pi}{4} + \frac{1}{8}\sigma - \frac{1}{4}t \ge \frac{7\pi}{4} + \frac{1}{4}\sigma
\end{aligned}$$

  where the equality is achieved when $t = -\frac{\sigma}{2}$.

- When $-t \in [-\frac{\sigma}{2} + \pi, -\frac{3\sigma}{2} + 2\pi]$,

$$\begin{aligned}
&\int_{-t}^{\pi-t} \frac{1}{4}\mathbb{1}_{\cos(\theta)\le\gamma} + \frac{7}{4}\mathbb{1}_{-\cos(\theta+\sigma)\le\gamma}\, d\theta \\
&= \int_{-t}^{-\frac{\sigma}{2}+2\pi} \frac{1}{4} + \frac{7}{4}\, d\theta + \int_{-\frac{\sigma}{2}+2\pi}^{\pi-t} \frac{7}{4}\, d\theta \\
&= \frac{9}{4}\pi - \frac{1}{8}\sigma + \frac{1}{4}t \ge \frac{7}{4}\pi + \frac{1}{4}\sigma
\end{aligned}$$

  where the equality is achieved when $t = \frac{3\sigma}{2} - 2\pi$.

Therefore for $\sigma \in (0, \pi)$,

$$\min_{t\in[0,2\pi)} \int_{-t}^{\pi-t} \frac{1}{4}\mathbb{1}_{\cos(\theta)\le\gamma} + \frac{7}{4}\mathbb{1}_{-\cos(\theta+\sigma)\le\gamma}\, d\theta = 2\pi - 2\sigma$$

where the equality is achieved when $t = \frac{\sigma}{2}$. Therefore

$$\mathcal{R}^*_{\ell_\gamma,\mathcal{H}_{\text{lin}}} = \frac{1}{2\pi} \times (2\pi - 2\sigma) = 1 - \frac{\sigma}{\pi}\,,$$

where the unique Bayes classifier satisfies $t_1^* = \frac{\sigma}{2}$.

For continuous margin-based loss ϕ, by (39) we have

$$\begin{aligned}
\mathcal{R}^*_{\phi,\mathcal{H}_{\text{lin}}} &= \frac{1}{2\pi} \min_{t\in[0,2\pi]} \int_0^{\pi} \frac{1}{4}\phi(\cos(\theta - t))\, d\theta + \int_0^{\pi} \frac{7}{4}\phi(\sin(\theta - t))\, d\theta \\
&= \frac{1}{2\pi} \min_{t\in[0,2\pi]} \int_{-t}^{\pi-t} \frac{1}{4}\phi(\cos(\theta)) + \frac{7}{4}\phi(-\cos(\theta + \sigma))\, d\theta.
\end{aligned} \tag{40}$$

If $t^* = \frac{\sigma}{2}$ is the minimizer of $g(t) = \int_{-t}^{\pi-t} \frac{1}{4}\phi(\cos(\theta)) + \frac{7}{4}\phi(-\cos(\theta + \sigma))\, d\theta,\ t \in [0, 2\pi]$, since $\frac{\sigma}{2}$ is not at the boundary of $[0, 2\pi]$, we need

$$g'\left(\frac{\sigma}{2}\right) = 0\,.$$

Since ϕ is continuous, by Leibniz Integral Rule, we have

$$\begin{aligned}
g'\left(\frac{\sigma}{2}\right) &= -\frac{1}{4}\phi\left(\cos\left(\pi - \frac{\sigma}{2}\right)\right) - \frac{7}{4}\phi\left(-\cos\left(\pi + \frac{\sigma}{2}\right)\right) + \frac{1}{4}\phi\left(\cos\left(-\frac{\sigma}{2}\right)\right) + \frac{7}{4}\phi\left(-\cos\left(\frac{\sigma}{2}\right)\right) \\
&= -\frac{1}{4}\phi\left(-\cos\left(\frac{\sigma}{2}\right)\right) - \frac{7}{4}\phi\left(\cos\left(\frac{\sigma}{2}\right)\right) + \frac{1}{4}\phi\left(\cos\left(\frac{\sigma}{2}\right)\right) + \frac{7}{4}\phi\left(-\cos\left(\frac{\sigma}{2}\right)\right) \\
&= \frac{3}{2}\phi\left(-\cos\left(\frac{\sigma}{2}\right)\right) - \frac{3}{2}\phi\left(\cos\left(\frac{\sigma}{2}\right)\right)\,.
\end{aligned}$$

Thus if $t^* = \frac{\sigma}{2}$ is the minimizer of $\mathcal{R}^*_{\phi,\mathcal{H}_{\text{lin}}}$, we need ϕ satisfies

$$\phi\left(-\cos\left(\frac{\sigma}{2}\right)\right) = \phi\left(\cos\left(\frac{\sigma}{2}\right)\right). \tag{41}$$

Therefore, if ϕ is $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ, we need ϕ satisfies (41) for any $\sigma \in (0,\pi)$. Namely ϕ satisfies

$$\phi(-\tau) = \phi(\tau), \quad \tau \in [0,1).$$

Note in our example, $\tau \in [-1,1]$, ϕ is continuous. We obtain that if ϕ is $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ, ϕ must be even function in $[-1,1]$. Next we claim that if ϕ is even function in $[-1,1]$, ϕ is not $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ. Indeed, for the distribution $y = 1$ if $\theta \in (0,\pi)$ and $y = -1$ if $\theta \in (\pi, 2\pi)$, we have

$$\begin{aligned}
\mathcal{R}^*_{\phi,\mathcal{H}_{\text{lin}}} &= \frac{1}{2\pi} \min_{t\in[0,2\pi]} \int_0^{\pi} \phi(\cos(\theta - t)) + \int_{\pi}^{2\pi} \phi(-\cos(\theta - t))\, d\theta \\
&= \frac{1}{\pi} \min_{t\in[0,2\pi]} \int_0^{\pi} \phi(\cos(\theta - t))\, d\theta \\
&= \frac{1}{\pi} \min_{t\in[0,2\pi]} \int_{-t}^{\pi-t} \phi(\cos(\theta))\, d\theta.
\end{aligned} \tag{42}$$

Note that when ϕ is even function in $[-1,1]$, $h(t) = \int_{-t}^{\pi-t} \phi(\cos(\theta))\, d\theta$ satisfies

$$h'(t) = -\phi(-\cos(t)) + \phi(\cos(t)) = 0, \quad t \in [0, 2\pi].$$

Thus $h(t)$ is a constant for $t \in [0,2\pi]$ and $\mathcal{R}^*_{\phi,\mathcal{H}_{\text{lin}}}$ can be attained for any classifier $t \in [0,2\pi]$. However, $\mathcal{R}^*_{\ell_\gamma,\mathcal{H}_{\text{lin}}}$ can not be attained for any classifier $t \in [0,2\pi]$ with respect to this distribution. Therefore when ϕ is even function in $[-1,1]$, ϕ is not $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ. By the claim, we conclude that any continuous margin-based loss ϕ is not $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ.

Furthermore, as shown by Awasthi et al. (2020), for a continuous and non-increasing margin-based loss ϕ, when $f \in \mathcal{H}_{\text{lin}}$, the supremum-based surrogate loss can be expressed as follows:

$$\tilde{\phi}(f,\mathbf{x},y) = \sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} \phi(yf(\mathbf{x}')) = \phi\Big(\inf_{\|\mathbf{s}\|\le 1} (yf(\mathbf{x}+\gamma\mathbf{s}))\Big) = \phi(y(\mathbf{w}\cdot\mathbf{x}) - \gamma) = \psi(y(\mathbf{w}\cdot\mathbf{x})),$$

where $\psi(t) = \phi(t-\gamma)$ is also a continuous margin-based loss. In view of the results above, we conclude that the supremum-based surrogate loss $\tilde{\phi}$ is also not $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ. □

E.7 Proof of Theorem 20, Theorem 21 and Theorem 23

Since the proofs adopt some results of (Steinwart, 2007), we introduce the notation used in (Steinwart, 2007) to make the proofs more clear. In this section, we denote the loss $\ell(f,\mathbf{x},y)$ defined on a particular hypothesis set $\mathcal{H}$ as $\ell_{\mathcal{H}}(f,\mathbf{x},y)$. For a joint distribution $\mathcal{P}$ over $\mathcal{X}\times\mathcal{Y}$, the corresponding conditional distribution and marginal distribution are denoted as $\mathcal{P}(\cdot|\mathbf{x})$ and $\mathcal{P}_X$ respectively. In (Steinwart, 2007), given a distribution $\mathcal{P}$ over $\mathcal{X}\times\mathcal{Y}$, the $\ell_{\mathcal{H}}$-risk and the inner $\ell_{\mathcal{H}}$-risk of a classifier $f \in \mathcal{H}$ for the loss $\ell_{\mathcal{H}}$ are denoted by

$$\mathcal{R}_{\ell_{\mathcal{H}},\mathcal{P}}(f) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{P}}[\ell_{\mathcal{H}}(f,\mathbf{x},y)], \quad \mathcal{C}_{\ell_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) = \mathop{\mathbb{E}}_{y\sim\mathcal{P}(\cdot|\mathbf{x})}[\ell_{\mathcal{H}}(f,\mathbf{x},y)].$$

Accordingly, the minimal $\ell_{\mathcal{H}}$-risk and minimal inner $\ell_{\mathcal{H}}$-risk are denoted by $\mathcal{R}^*_{\ell_{\mathcal{H}},\mathcal{P}}$ and $\mathcal{C}^*_{\ell_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}$. For convenience, we will alternately use the notation of risk and inner risk presented above and in Section 2 for the proofs. Next, we introduce the minimizability proposed in (Steinwart, 2007).

Definition 31 (minimizability)**.** *Given a distribution $\mathcal{P}$ over $\mathcal{X}\times\mathcal{Y}$ and a hypothesis set $\mathcal{H}$. We say that loss $\ell_{\mathcal{H}}(f,\mathbf{x},y)$ is $\mathcal{P}$-minimizable if for all $\epsilon > 0$ there exists $f_\epsilon \in \mathcal{H}$ such that for all $\mathbf{x} \in \mathcal{X}$ we have*

$$\mathcal{C}_{\ell_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f_\epsilon) < \mathcal{C}^*_{\ell_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + \epsilon.$$

The following lemmas are useful in the proofs of Theorem 20 and Theorem 23.

Lemma 32. *Given a distribution $\mathcal{P}$ over $\mathcal{X}\times\mathcal{Y}$ and a hypothesis set $\mathcal{H}$. Let ϕ be a margin-based loss. Then $\phi_{\mathcal{H}_{\text{all}}}$ is $\mathcal{P}$-minimizable. If there exists $f^* \in \mathcal{H} \subset \mathcal{H}_{\text{all}}$ such that $\mathcal{R}^*_{\phi_{\mathcal{H}_{\text{all}}},\mathcal{P}} = \mathcal{R}_{\phi_{\mathcal{H}},\mathcal{P}}(f^*)$, then $\phi_{\mathcal{H}}$ is also $\mathcal{P}$-minimizable in the almost surely sense.*

Proof. By Theorem 3.2 of (Steinwart, 2007), since $\mathcal{C}^*_{\phi_{\mathcal{H}_{\mathrm{all}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} < \infty$ for all $\mathbf{x} \in \mathcal{X}$, $\phi_{\mathcal{H}_{\mathrm{all}}}$ is $\mathcal{P}$-minimizable. Therefore, by Lemma 2.5 of (Steinwart, 2007), we have

$$\mathcal{R}^*_{\phi_{\mathcal{H}_{\mathrm{all}}},\mathcal{P}} = \int_{\mathcal{X}} \mathcal{C}^*_{\phi_{\mathcal{H}_{\mathrm{all}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} \, d\mathcal{P}_X(\mathbf{x}).$$

Then by the assumption,

$$\int_{\mathcal{X}} \mathcal{C}_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f^*) \, d\mathcal{P}_X(\mathbf{x}) = \mathcal{R}_{\phi_{\mathcal{H}},\mathcal{P}}(f^*) = \mathcal{R}^*_{\phi_{\mathcal{H}_{\mathrm{all}}},\mathcal{P}} = \int_{\mathcal{X}} \mathcal{C}^*_{\phi_{\mathcal{H}_{\mathrm{all}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} \, d\mathcal{P}_X(\mathbf{x}).$$

Since

$$\mathcal{C}^*_{\phi_{\mathcal{H}_{\mathrm{all}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} \le \mathcal{C}_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f^*),$$

for almost all $\mathbf{x} \in \mathcal{X}$,

$$\mathcal{C}^*_{\phi_{\mathcal{H}_{\mathrm{all}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} = \mathcal{C}_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f^*).$$

Thus, for all $\epsilon > 0$, for almost all $\mathbf{x} \in \mathcal{X}$ we have

$$\mathcal{C}_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f^*) = \mathcal{C}^*_{\phi_{\mathcal{H}_{\mathrm{all}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} < \mathcal{C}^*_{\phi_{\mathcal{H}_{\mathrm{all}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + \epsilon \le \mathcal{C}^*_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + \epsilon.$$

This completes the proof. □

Lemma 33. *Given a distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$ and a hypothesis set $\mathcal{H}$. Let ϕ be a margin-based loss. If for $\eta \ge 0$, there exists $f^* \in \mathcal{H} \subset \mathcal{H}_{\mathrm{all}}$ such that $\mathcal{R}_{\phi_{\mathcal{H}},\mathcal{P}}(f^*) \le \mathcal{R}^*_{\phi_{\mathcal{H}_{\mathrm{all}}},\mathcal{P}} + \eta$, then $\phi_{\mathcal{H}}$ satisfies*

$$\int_{\mathcal{X}} \mathcal{C}^*_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} \, d\mathcal{P}_X(\mathbf{x}) \le \mathcal{R}^*_{\phi_{\mathcal{H}},\mathcal{P}} \le \int_{\mathcal{X}} \mathcal{C}^*_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} \, d\mathcal{P}_X(\mathbf{x}) + \eta.$$

Proof. By Lemma 32, $\phi_{\mathcal{H}_{\mathrm{all}}}$ is $\mathcal{P}$-minimizable. Then by Lemma 2.5 of (Steinwart, 2007), we have

$$\mathcal{R}^*_{\phi_{\mathcal{H}_{\mathrm{all}}},\mathcal{P}} = \int_{\mathcal{X}} \mathcal{C}^*_{\phi_{\mathcal{H}_{\mathrm{all}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} \, d\mathcal{P}_X(\mathbf{x}).$$

Therefore,

$$\mathcal{R}^*_{\phi_{\mathcal{H}},\mathcal{P}} \le \mathcal{R}_{\phi_{\mathcal{H}},\mathcal{P}}(f^*) \le \int_{\mathcal{X}} \mathcal{C}^*_{\phi_{\mathcal{H}_{\mathrm{all}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} \, d\mathcal{P}_X(\mathbf{x}) + \eta \le \int_{\mathcal{X}} \mathcal{C}^*_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} \, d\mathcal{P}_X(\mathbf{x}) + \eta.$$

Also,

$$\begin{aligned}\int_{\mathcal{X}} \mathcal{C}^*_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} \, d\mathcal{P}_X(\mathbf{x}) &\le \int_{\mathcal{X}} \inf_{f \in \mathcal{H}} \mathcal{C}_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) \, d\mathcal{P}_X(\mathbf{x}) \\ &\le \inf_{f \in \mathcal{H}} \int_{\mathcal{X}} \mathcal{C}_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) \, d\mathcal{P}_X(x) = \mathcal{R}^*_{\phi_{\mathcal{H}},\mathcal{P}}.\end{aligned}$$

□

Lemma 34. *Given a distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$ with random variables X and Y and a hypothesis set $\mathcal{H}$ such that $\mathcal{R}^*_{\ell_\gamma,\mathcal{H}} = \mathcal{R}_{\ell_\gamma}(f^*) = 0$, where $f^* \in \mathcal{H}$ achieves the Bayes risk. Then f^* correctly classifies $\mathbf{x} \in \mathcal{X}$ in the almost surely sense and for almost all $\mathbf{x} \in \mathcal{X}$, any $\mathbf{x}' \in \{\mathbf{x}' \colon \|\mathbf{x}' - \mathbf{x}\| \le \gamma\}$ has same label as $\mathbf{x}$.*

Proof. Since $\mathcal{R}^*_{\ell_{\gamma_{\mathcal{H}}},\mathcal{P}} = \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} = 0$, the distribution $\mathcal{P}$ is $\mathcal{H}$-realizable. Therefore $\mathbb{P}(Y = 1|X = \mathbf{x}) = 1$ or 0. Thus

$$\mathcal{C}_{\ell_{\gamma_{\mathcal{H}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) = \begin{cases} \sup\limits_{\mathbf{x}' \colon \|\mathbf{x}-\mathbf{x}'\| \le \gamma} \mathbb{1}_{\{f(\mathbf{x}') \le 0\}}, & \text{if } \mathbb{P}(Y = 1|X = \mathbf{x}) = 1, \\ \sup\limits_{\mathbf{x}' \colon \|\mathbf{x}-\mathbf{x}'\| \le \gamma} \mathbb{1}_{\{-f(\mathbf{x}') \le 0\}}, & \text{if } \mathbb{P}(Y = 1|X = \mathbf{x}) = 0, \end{cases}$$

Since $\mathcal{R}_{\ell_{\gamma_{\mathcal{H}}},\mathcal{P}}(f^*) = \mathcal{R}_{\ell_\gamma}(f^*) = 0$, we have $\mathcal{C}_{\ell_{\gamma_{\mathcal{H}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f^*) = 0$ for almost all $\mathbf{x} \in \mathcal{X}$. When $\mathbb{P}(Y = 1|X = \mathbf{x}) = 1$, we obtain

$$\sup_{\mathbf{x}' \colon \|\mathbf{x}-\mathbf{x}'\| \le \gamma} \mathbb{1}_{\{f^*(\mathbf{x}') \le 0\}} = 0 \implies f^*(\mathbf{x}') > 0 \text{ for any } \mathbf{x}' \in \{\mathbf{x}' \colon \|\mathbf{x}' - \mathbf{x}\| \le \gamma\}. \tag{43}$$

When $\mathbb{P}(Y = 1|X = \mathbf{x}) = 0$, we obtain

$$\sup_{\mathbf{x}' \colon \|\mathbf{x}-\mathbf{x}'\| \le \gamma} \mathbb{1}_{\{-f^*(\mathbf{x}') \le 0\}} = 0 \implies f^*(\mathbf{x}') < 0 \text{ for any } \mathbf{x}' \in \{\mathbf{x}' \colon \|\mathbf{x}' - \mathbf{x}\| \le \gamma\}. \tag{44}$$

Thus $f^*(\mathbf{x}) > 0$ when $\mathbb{P}(Y = 1|X = \mathbf{x}) = 1$ and $f^*(\mathbf{x}) < 0$ when $\mathbb{P}(Y = 1|X = \mathbf{x}) = 0$ for almost all $\mathbf{x} \in \mathcal{X}$. Therefore f^* correctly classify $\mathbf{x} \in \mathcal{X}$ in the almost surely sense. Furthermore, by (43) and (44), for almost all $\mathbf{x} \in \mathcal{X}$, any $\mathbf{x}' \in \{\mathbf{x}' \colon \|\mathbf{x}' - \mathbf{x}\| \le \gamma\}$ has same label as $\mathbf{x}$. □

Lemma 35. *Given a distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$ and a hypothesis set $\mathcal{H}$ such that $\mathcal{R}^*_{\ell_\gamma,\mathcal{H}} = 0$. Let ϕ be a margin-based loss and $\tilde{\phi}(f, \mathbf{x}, y) = \sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} \phi(yf(\mathbf{x}'))$. If $\phi_{\mathcal{H}}$ is $\mathcal{P}$-minimizable in the almost surely sense, then $\tilde{\phi}_{\mathcal{H}}$ is also $\mathcal{P}$-minimizable in the almost surely sense.*

Proof. As shown by Awasthi et al. (2020), $\tilde{\phi}$ has the equivalent form

$$\tilde{\phi}(f, \mathbf{x}, y) = \phi\left(\inf_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} (yf(\mathbf{x}'))\right).$$

Since $\mathcal{R}^*_{\ell_{\gamma_{\mathcal{H}}},\mathcal{P}} = \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} = 0$, the distribution $\mathcal{P}$ is $\mathcal{H}$-realizable. Therefore $\mathbb{P}(Y = 1|X = \mathbf{x}) = 1$ or 0. Thus

$$\mathcal{C}_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) = \begin{cases} \phi(f(\mathbf{x})), & \text{if } \mathbb{P}(Y = 1|X = \mathbf{x}) = 1, \\ \phi(-f(\mathbf{x})), & \text{if } \mathbb{P}(Y = 1|X = \mathbf{x}) = 0, \end{cases}$$

Note $\tilde{\phi}(f, \mathbf{x}, +1) = \phi\left(\inf_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} f(\mathbf{x}')\right) = \phi(f(m_{f,\mathbf{x}}))$, where w.l.o.g. we assume that f is continuous and $m_{f,\mathbf{x}} \in \{\mathbf{x}' : \|\mathbf{x} - \mathbf{x}'\| \le \gamma\}$ is the point such that $\min_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} f(\mathbf{x}') = f(m_{f,\mathbf{x}})$. Similarly $\tilde{\phi}(f, \mathbf{x}, -1) = \phi\left(-\sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} f(\mathbf{x}')\right) = \phi(-f(M_{f,\mathbf{x}}))$, where w.l.o.g. we assume that f is continuous and $M_{f,\mathbf{x}} \in \{\mathbf{x}' : \|\mathbf{x} - \mathbf{x}'\| \le \gamma\}$ is the point such that $\max_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} f(\mathbf{x}') = f(M_{f,\mathbf{x}})$. Then for $\tilde{\phi}_{\mathcal{H}}$, we have

$$\mathcal{C}_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) = \begin{cases} \phi(f(m_{f,\mathbf{x}})), & \text{if } \mathbb{P}(Y = 1|X = \mathbf{x}) = 1, \\ \phi(-f(M_{f,\mathbf{x}})), & \text{if } \mathbb{P}(Y = 1|X = \mathbf{x})) = 0, \end{cases}$$

Since $\phi_{\mathcal{H}}$ is $\mathcal{P}$-minimizable in the almost surely sense, by the definition for all $\epsilon > 0$, there exists an $f^* \in \mathcal{H}$ such that for almost all $\mathbf{x} \in \mathcal{X}$ we have

$$\mathcal{C}_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f^*) < \mathcal{C}^*_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + \epsilon.$$

When $\mathbb{P}(Y = 1|X = \mathbf{x}) = 1$, we obtain

$$\mathcal{C}_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f^*) = \phi(f^*(m_{f^*,\mathbf{x}})) = \mathcal{C}_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|m_{f^*,\mathbf{x}}),m_{f^*,\mathbf{x}}}(f^*) < \mathcal{C}^*_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|m_{f^*,\mathbf{x}}),m_{f^*,\mathbf{x}}} + \epsilon \le \mathcal{C}^*_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + \epsilon$$

where we used the fact that $m_{f^*,\mathbf{x}}$ satisfies $\mathbb{P}(Y = 1|X = m_{f^*,\mathbf{x}}) = 1$ by Lemma 34 and ϕ is non-increasing. Similarly, when $\mathbb{P}(Y = 1|X = \mathbf{x}) = 0$, we obtain

$$\mathcal{C}_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f^*) = \phi(-f^*(M_{f^*,\mathbf{x}})) = \mathcal{C}_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|M_{f^*,\mathbf{x}}),M_{f^*,\mathbf{x}}}(f^*) < \mathcal{C}^*_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|M_{f^*,\mathbf{x}}),M_{f^*,\mathbf{x}}} + \epsilon \le \mathcal{C}^*_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + \epsilon$$

where we used the fact that $M_{f^*,\mathbf{x}}$ satisfies $\mathbb{P}(Y = 1|X = M_{f^*,\mathbf{x}}) = 0$ by Lemma 34 and ϕ is non-increasing. Above all, for all $\epsilon > 0$, there exists an $f^* \in \mathcal{H}$ such that for almost all $\mathbf{x} \in \mathcal{X}$ we have

$$\mathcal{C}_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f^*) < \mathcal{C}^*_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + \epsilon.$$

□

We modify Theorem 2.8 of (Steinwart, 2007), whose proof is very similar.

Theorem 36. *Given a distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$ and a hypothesis set $\mathcal{H}$. Let $\ell_1 : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0, +\infty]$, $\ell_2 : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0, +\infty]$ be two loss functions such that $\mathcal{R}^*_{\ell_1,\mathcal{P}} = \int_{\mathcal{X}} \mathcal{C}^*_{\ell_1,\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}\, d\mathcal{P}_X(\mathbf{x}) < +\infty$ and $\int_{\mathcal{X}} \mathcal{C}^*_{\ell_2,\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}\, d\mathcal{P}_X(\mathbf{x}) \le \mathcal{R}^*_{\ell_2,\mathcal{P}} \le \int_{\mathcal{X}} \mathcal{C}^*_{\ell_2,\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}\, d\mathcal{P}_X(\mathbf{x}) + \eta < +\infty$ for $\eta \ge 0$. Furthermore, assume that there exist a function $b \in \mathcal{L}_1(\mathcal{P}_X)$ and measurable functions $\delta(\epsilon, \cdot) : X \to (0, +\infty)$, $\epsilon > 0$, such that*

$$\mathcal{C}_{\ell_1,\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) \le \mathcal{C}^*_{\ell_1,\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + b(\mathbf{x})$$

and

$$\mathcal{C}_{\ell_2,\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) < \mathcal{C}^*_{\ell_2,\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + \delta(\epsilon, \mathbf{x}) \implies \mathcal{C}_{\ell_1,\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) < \mathcal{C}^*_{\ell_1,\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + \epsilon$$

for all $\mathbf{x} \in \mathcal{X}$, $\epsilon > 0$ and $f \in \mathcal{H}$. Then, for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $f \in \mathcal{H}$ we have

$$\mathcal{R}_{\ell_2,\mathcal{P}}(f) + \eta < \mathcal{R}^*_{\ell_2,\mathcal{P}} + \delta \implies \mathcal{R}_{\ell_1,\mathcal{P}}(f) < \mathcal{R}^*_{\ell_1,\mathcal{P}} + \epsilon.$$

Proof. Define $\mathcal{C}_{1,\mathbf{x}}(f) = \mathcal{C}_{\ell_1,\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) - \mathcal{C}^*_{\ell_1,\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}$ and $\mathcal{C}_{2,\mathbf{x}}(f) = \mathcal{C}_{\ell_2,\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) - \mathcal{C}^*_{\ell_2,\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}$ for $\mathbf{x} \in \mathcal{X}$, $f \in \mathcal{H}$. For a fixed $\epsilon > 0$, define $h(\mathbf{x}) = \delta(\epsilon, \mathbf{x})$, $\mathbf{x} \in \mathcal{X}$. Then for all $\mathbf{x} \in \mathcal{X}$ and $f \in \mathcal{H}$ such that $\mathcal{C}_{1,\mathbf{x}}(f) \geq \epsilon$, we have $\mathcal{C}_{2,\mathbf{x}}(f) \geq h(\mathbf{x})$. Therefore,

$$\begin{aligned}\mathcal{R}_{\ell_2,\mathcal{P}}(f) - \mathcal{R}^*_{\ell_2,\mathcal{P}} + \eta &\geq \mathcal{R}_{\ell_2,\mathcal{P}}(f) - \int_{\mathcal{X}} \mathcal{C}^*_{\ell_2,\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}\, d\mathcal{P}_X(\mathbf{x}) \\ &= \int_{\mathcal{X}} \mathcal{C}_{2,\mathbf{x}}(f)\, d\mathcal{P}_X(\mathbf{x}) \geq \int_{\mathcal{C}_{1,\mathbf{x}}(f)\geq\epsilon} h(\mathbf{x})\, d\mathcal{P}_X(\mathbf{x}),\end{aligned}$$

for all $f \in \mathcal{H}$. Furthermore, since $h(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}$, the measure $\nu := b\mathcal{P}_X$ is absolutely continuous with respect to $\mu := h\mathcal{P}_X$, and thus there exists $\delta > 0$ such that $\nu(A) < \epsilon$ for all measurable $A \subset X$ with $\mu(A) < \delta$. Therefore, for $f \in \mathcal{H}$ with $\mathcal{R}_{\ell_2,\mathcal{P}}(f) - \mathcal{R}^*_{\ell_2,\mathcal{P}} + \eta < \delta$ and $A := \{\mathbf{x} \in \mathcal{X}, \mathcal{C}_{1,\mathbf{x}}(f) \geq \epsilon\}$, we obtain

$$\begin{aligned}\mathcal{R}_{\ell_1,\mathcal{P}}(f) - \mathcal{R}^*_{\ell_1,\mathcal{P}} &= \int_{\mathcal{C}_{1,\mathbf{x}}(f)\geq\epsilon} \mathcal{C}_{1,\mathbf{x}}(f)\, d\mathcal{P}_X(\mathbf{x}) + \int_{\mathcal{C}_{1,\mathbf{x}}(f)<\epsilon} \mathcal{C}_{1,\mathbf{x}}(f)\, d\mathcal{P}_X(\mathbf{x}) \\ &\leq \int_A b(\mathbf{x})\, d\mathcal{P}_X(\mathbf{x}) + \epsilon < 2\epsilon.\end{aligned}$$

□

Theorem 20. *Let $\mathcal{P}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{H}$ a hypothesis set for which $\mathcal{R}^*_{\ell_\gamma,\mathcal{H}} = 0$. Let ϕ be a margin-based loss. If for $\eta \geq 0$, there exists $f^* \in \mathcal{H} \subset \mathcal{H}_{\text{all}}$ such that $\mathcal{R}_\phi(f^*) \leq \mathcal{R}^*_{\phi,\mathcal{H}_{\text{all}}} + \eta < +\infty$ and ϕ is $\mathcal{H}$-calibrated with respect to ℓ_γ, then for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $f \in \mathcal{H}$,*

$$\mathcal{R}_\phi(f) + \eta < \mathcal{R}^*_{\phi,\mathcal{H}} + \delta \implies \mathcal{R}_{\ell_\gamma}(f) < \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} + \epsilon.$$

Proof. Since $\mathcal{R}^*_{\ell_{\gamma_{\mathcal{H}}},\mathcal{P}} = \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} = 0$, we obtain

$$0 \leq \int_{\mathcal{X}} \mathcal{C}^*_{\ell_{\gamma_{\mathcal{H}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}\, d\mathcal{P}_X(\mathbf{x}) \leq \mathcal{R}^*_{\ell_{\gamma_{\mathcal{H}}},\mathcal{P}} = 0.$$

By Lemma 33, $\phi_{\mathcal{H}}$ satisfies

$$\int_X \mathcal{C}^*_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}\, d\mathcal{P}_X(\mathbf{x}) \leq \mathcal{R}^*_{\phi_{\mathcal{H}},\mathcal{P}} \leq \int_{\mathcal{X}} \mathcal{C}^*_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}\, d\mathcal{P}_X(\mathbf{x}) + \eta < +\infty.$$

Since for all $\mathbf{x} \in \mathcal{X}$ and $f \in \mathcal{H}$, $\mathcal{C}_{\ell_{\gamma_{\mathcal{H}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) \leq 1$, we obtain

$$\mathcal{C}_{\ell_{\gamma_{\mathcal{H}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) \leq \mathcal{C}^*_{\ell_{\gamma_{\mathcal{H}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + 1.$$

Also, since ϕ is $\mathcal{H}$-calibrated with respect to ℓ_γ, for all $x \in \mathcal{X}$, $\epsilon > 0$ and $f \in \mathcal{H}$, there exists $\delta > 0$ such that

$$\mathcal{C}_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),x}(f) < \mathcal{C}^*_{\phi_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + \delta \implies \mathcal{C}_{\ell_{\gamma_{\mathcal{H}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) < \mathcal{C}^*_{\ell_{\gamma_{\mathcal{H}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + \epsilon.$$

Therefore by Theorem 36, for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $f \in \mathcal{H}$ we have

$$\mathcal{R}_{\phi_{\mathcal{H}},\mathcal{P}}(f) + \eta < \mathcal{R}^*_{\phi_{\mathcal{H}},\mathcal{P}} + \delta \implies \mathcal{R}_{\ell_{\gamma_{\mathcal{H}}},\mathcal{P}}(f) < \mathcal{R}^*_{\ell_{\gamma_{\mathcal{H}}},\mathcal{P}} + \epsilon. \tag{45}$$

Using the notation in Section 2, we can rewrite (45) as

$$\mathcal{R}_\phi(f) + \eta < \mathcal{R}^*_{\phi,\mathcal{H}} + \delta \implies \mathcal{R}_{\ell_\gamma}(f) < \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} + \epsilon.$$

□

Theorem 21. *Let $\mathcal{P}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. Assume that there exists $g^* \in \mathcal{H}_{\text{lin}}$ such that $\mathcal{R}_{\ell_\gamma}(g^*) = \mathcal{R}^*_{\ell_\gamma,\mathcal{H}_{\text{all}}}$. Let ϕ be a margin-based loss. If for $\eta \geq 0$, there exists $f^* \in \mathcal{H}_{\text{lin}} \subset \mathcal{H}_{\text{all}}$ such that $\mathcal{R}_\phi(f^*) \leq \mathcal{R}^*_{\phi,\mathcal{H}_{\text{all}}} + \eta < +\infty$ and ϕ is $\mathcal{H}_{\text{lin}}$-calibrated with respect to ℓ_γ, then for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $f \in \mathcal{H}_{\text{lin}}$ we have*

$$\mathcal{R}_\phi(f) + \eta < \mathcal{R}^*_{\phi,\mathcal{H}_{\text{lin}}} + \delta \implies \mathcal{R}_{\ell_\gamma}(f) < \mathcal{R}^*_{\ell_\gamma,\mathcal{H}_{\text{lin}}} + \epsilon.$$

Proof. As shown by Bao et al. (2020a), the adversarial $0/1$ loss $\ell_\gamma = \mathbb{1}_{yf(\mathbf{x})\le\gamma}$ is a margin-based loss when $f \in \mathcal{H}_{\text{lin}}$. By Lemma 33, $\ell_{\gamma\mathcal{H}_{\text{lin}}}$ and $\phi_{\mathcal{H}_{\text{lin}}}$ satisfy

$$\int_{\mathcal{X}} \mathcal{C}^*_{\ell_{\gamma\mathcal{H}_{\text{lin}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}\, d\mathcal{P}_X(\mathbf{x}) = \mathcal{R}^*_{\ell_{\gamma\mathcal{H}_{\text{lin}}},\mathcal{P}},$$
$$\int_{X} \mathcal{C}^*_{\phi_{\mathcal{H}_{\text{lin}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}\, d\mathcal{P}_X(\mathbf{x}) \le \mathcal{R}^*_{\phi_{\mathcal{H}_{\text{lin}}},\mathcal{P}} \le \int_{\mathcal{X}} \mathcal{C}^*_{\phi_{\mathcal{H}_{\text{lin}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}\, d\mathcal{P}_X(\mathbf{x}) + \eta < +\infty.$$

Since for all $\mathbf{x} \in \mathcal{X}$ and $f \in \mathcal{H}_{\text{lin}}$, $\mathcal{C}_{\ell_{\gamma\mathcal{H}_{\text{lin}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) \le 1$, we obtain

$$\mathcal{C}_{\ell_{\gamma\mathcal{H}_{\text{lin}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) \le \mathcal{C}^*_{\ell_{\gamma\mathcal{H}_{\text{lin}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + 1.$$

Also, since ϕ is $\mathcal{H}_{\text{lin}}$-calibrated with respect to ℓ_γ, for all $x \in \mathcal{X}$, $\epsilon > 0$ and $f \in \mathcal{H}_{\text{lin}}$, there exists $\delta > 0$ such that

$$\mathcal{C}_{\phi_{\mathcal{H}_{\text{lin}}},\mathcal{P}(\cdot|\mathbf{x}),x}(f) < \mathcal{C}^*_{\phi_{\mathcal{H}_{\text{lin}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + \delta \implies \mathcal{C}_{\ell_{\gamma\mathcal{H}_{\text{lin}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) < \mathcal{C}^*_{\ell_{\gamma\mathcal{H}_{\text{lin}}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + \epsilon.$$

Therefore by Theorem 36, for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $f \in \mathcal{H}_{\text{lin}}$ we have

$$\mathcal{R}_{\phi_{\mathcal{H}_{\text{lin}}},\mathcal{P}}(f) + \eta < \mathcal{R}^*_{\phi_{\mathcal{H}_{\text{lin}}},\mathcal{P}} + \delta \implies \mathcal{R}_{\ell_{\gamma\mathcal{H}_{\text{lin}}},\mathcal{P}}(f) < \mathcal{R}^*_{\ell_{\gamma\mathcal{H}_{\text{lin}}},\mathcal{P}} + \epsilon. \tag{46}$$

Using the notation in Section 2, we can rewrite (46) as

$$\mathcal{R}_\phi(f) + \eta < \mathcal{R}^*_{\phi,\mathcal{H}_{\text{lin}}} + \delta \implies \mathcal{R}_{\ell_\gamma}(f) < \mathcal{R}^*_{\ell_\gamma,\mathcal{H}_{\text{lin}}} + \epsilon.$$

□

Theorem 23. *Given a distribution $\mathcal{P}$ over $\mathcal{X}\times\mathcal{Y}$ and a hypothesis set $\mathcal{H}$ such that $\mathcal{R}^*_{\ell_\gamma,\mathcal{H}} = 0$. Let ϕ be a non-increasing margin-based loss. If there exists $f^* \in \mathcal{H} \subset \mathcal{H}_{\text{all}}$ such that $\mathcal{R}_\phi(f^*) = \mathcal{R}^*_{\phi,\mathcal{H}_{\text{all}}} < +\infty$ and $\tilde{\phi}(f,\mathbf{x},y) = \sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|\le\gamma} \phi(yf(\mathbf{x}'))$ is $\mathcal{H}$-calibrated with respect to ℓ_γ, then for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $f \in \mathcal{H}$ we have*

$$\mathcal{R}_{\tilde{\phi}}(f) < \mathcal{R}^*_{\tilde{\phi},\mathcal{H}} + \delta \implies \mathcal{R}_{\ell_\gamma}(f) < \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} + \epsilon.$$

Proof. By Lemma 32 and Lemma 35, $\tilde{\phi}_{\mathcal{H}}$ is $\mathcal{P}$-minimizable in the almost surely sense. Then for any $n \in \mathbb{N}$, there exists an $f^*_n \in \mathcal{H}$ such that for almost all $\mathbf{x} \in \mathcal{X}$ we have

$$\mathcal{C}_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f^*_n) < \mathcal{C}^*_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + \frac{1}{n}.$$

Therefore

$$\begin{aligned}\mathcal{R}^*_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}} &\le \int_{\mathcal{X}} \mathcal{C}_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f^*_n)\, d\mathcal{P}_X(\mathbf{x}) \le \int_{\mathcal{X}} \mathcal{C}^*_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}\, d\mathcal{P}_X(\mathbf{x}) + \frac{1}{n}\\ &\le \inf_{f\in\mathcal{H}} \int_{\mathcal{X}} \mathcal{C}_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f)\, d\mathcal{P}_X(\mathbf{x}) + \frac{1}{n} \le \mathcal{R}^*_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}} + \frac{1}{n}.\end{aligned}$$

By taking $n \to +\infty$, we obtain

$$\mathcal{R}^*_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}} = \int_{\mathcal{X}} \mathcal{C}^*_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}\, d\mathcal{P}_X(\mathbf{x}).$$

Since $\mathcal{R}^*_{\ell_{\gamma\mathcal{H}},\mathcal{P}} = \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} = 0$, we obtain

$$0 \le \int_{\mathcal{X}} \mathcal{C}^*_{\ell_{\gamma\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}\, d\mathcal{P}_X(\mathbf{x}) \le \mathcal{R}^*_{\ell_{\gamma\mathcal{H}},\mathcal{P}} = 0.$$

Since for all $x \in \mathcal{X}$ and $f \in \mathcal{H}$, $\mathcal{C}_{\ell_{\gamma\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) \le 1$, we obtain

$$\mathcal{C}_{\ell_{\gamma\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) \le \mathcal{C}^*_{\ell_{\gamma\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + 1.$$

Also, since $\tilde{\phi}$ is $\mathcal{H}$-calibrated with respect to ℓ_γ, for all $x \in \mathcal{X}$, $\epsilon > 0$ and $f \in \mathcal{H}$, there exists $\delta > 0$ such that

$$\mathcal{C}_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) < \mathcal{C}^*_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + \delta \implies \mathcal{C}_{\ell_{\gamma\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}}(f) < \mathcal{C}^*_{\ell_{\gamma\mathcal{H}},\mathcal{P}(\cdot|\mathbf{x}),\mathbf{x}} + \epsilon.$$

Therefore by Theorem 36 ($\eta = 0$ here), for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $f \in \mathcal{H}$ we have

$$\mathcal{R}_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}}(f) < \mathcal{R}^*_{\tilde{\phi}_{\mathcal{H}},\mathcal{P}} + \delta \implies \mathcal{R}_{\ell_{\gamma\mathcal{H}},\mathcal{P}}(f) < \mathcal{R}^*_{\ell_{\gamma\mathcal{H}},\mathcal{P}} + \epsilon. \tag{47}$$

Using the notation in Section 2, we can rewrite (47) as

$$\mathcal{R}_{\tilde{\phi}}(f) < \mathcal{R}^*_{\tilde{\phi},\mathcal{H}} + \delta \implies \mathcal{R}_{\ell_\gamma}(f) < \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} + \epsilon.$$

□