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  
f1066@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 $H$-calibration and $H$-consistency of adversarial surrogate losses. We show that convex loss functions, or the supremum-based convex losses often used in applications, are not $H$-calibrated for common hypothesis sets used in machine learning. We then give a characterization of $H$-calibration and prove that some surrogate losses are indeed $H$-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 $H$-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 $H$-consistent. We also report a series of empirical results which show that many $H$-calibrated surrogate losses are indeed not $H$-consistent, and validate our theoretical assumptions. Our adversarial $H$-consistency results are novel, even for the case where $H$ 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 $\ell_p$ 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

then replaced with a more stringent adversarial loss, which requires a predictor to correctly classify an input point \(x\) and also to maintain the same classification for all points at a small \(\ell_p\) distance of \(x\). 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 \(x\) is inherently not just a function of \(f(x)\) but also of its values around a neighborhood of \(x\). 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 \(\phi\) 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. (2020) 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 \(\mathcal{H} = \mathcal{H}_{\text{lin}}\) 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., 2020) 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 \(\rho\)-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 \(\rho\)-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
We will denote vectors as lowercase bold letters (e.g. $\mathbf{x}$). The $d$-dimensional $l_2$-ball with radius $r$ is denoted by $B_d^2(r):=\{\mathbf{z}\in\mathbb{R}^d\mid \|\mathbf{z}\|_2\leq r\}$. 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 $P$ 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 $P$. The generalization error of a classifier $f\in\mathcal{H}$ is defined by $\mathcal{R}_{\ell_0}(f) = \mathbb{E}_{(x,y)\sim P}[\ell_0(f,x,y)]$, where $\ell_0(f,x,y) = \mathbb{1}_{yf(x)\leq 0}$ is the standard $0/1$ loss. More generally, the $\ell$-risk of a classifier $f$ for a surrogate loss $\ell(f,x,y)$ is defined by

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

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

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

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 $\ell_1$ is $\mathcal{H}$-consistent with respect to a loss function $\ell_2$, if the following holds:

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

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

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

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

Moreover, the minimal inner $\ell$-risk on $\mathcal{H}$ is denoted by $\mathcal{C}_\ell^{\star}(\mathcal{X},\eta) = \inf_{f\in\mathcal{H}}\mathcal{C}_\ell(f,x,\eta)$. For a margin-based loss $\phi$, the generic conditional $\phi$-risk is $\mathcal{C}_\phi(t,\eta) = \eta\phi(t) + (1-\eta)\phi(-t)$ 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 $\ell_1$ is $\mathcal{H}$-calibrated with respect to a loss function $\ell_2$ if, for any $\epsilon > 0$, $\eta \in [0,1]$, and $x \in \mathcal{X}$, there exists $\delta > 0$ such that for all $f \in \mathcal{H}$ we have

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

Steinwart (2007) points out that if $\ell_1$ is $\mathcal{H}$-calibrated wrt $\ell_2$, 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 $(\ell_1,\ell_2)$ as follows: for all $x\in\mathcal{X}$, $\eta\in[0,1]$ and $\epsilon > 0$,

$$\delta_{\max}(\epsilon,x,\eta) = \inf_{f\in\mathcal{H}} \left\{ \mathcal{C}_{\ell_1}(f,x,\eta) - \mathcal{C}_{\ell_1,\mathcal{H}}(x,\eta) \mid \mathcal{C}_{\ell_2}(f,x,\eta) - \mathcal{C}_{\ell_2,\mathcal{H}}(x,\eta) \geq \epsilon \right\}.$$
will consider perturbations in the
\( \ell \)

(2007) further establishes a sufficient condition called minimizability under which
\( H \)

(\( H \)et al., 2006). In this special case, calibration implies consistency. Steinwart
\( H \)

(\( H \)ao et al., 2020) are not
\( H \)

fact for any
\( \ell \)
tolerant perturbations. Given
\( x \)
In adversarially robust classification, the loss at
\( 3 \)
Adversarially Robust Classification

We will denote by
\( \gamma \)
∈ \[0, 1\]
and
\( \epsilon > 0 \), the calibration function gives the maximal \( \delta \) satisfying the calibration condition (5). The following proposition is an important result from (Steinwart, 2007).

**Proposition 4** (Lemma 2.9 in \( \text{Steinwart, 2007} \)). Given a hypothesis set \( \mathcal{H} \), loss \( \ell_1 \) is \( \mathcal{H} \)-calibrated with respect to \( \ell_2 \) if and only if its calibration function \( \delta_{\text{max}} \) satisfies \( \delta_{\text{max}} (\epsilon, x, \eta) > 0 \) for all \( x \in \mathcal{X} \), \( \eta \in [0, 1] \) 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 \( \ell_2 \)-risk is close to the minimal inner \( \ell_2 \)-risk within \( \mathcal{H} \) for each \( x \in \mathcal{X} \). \( \mathcal{H} \)-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 (\( \text{Bao et al., 2020} \)) 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 (\( \text{Zhang, 2004; Bartlett et al., 2006} \)). 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 (\( \text{Zhang, 2004; Bartlett et al., 2006} \)). 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 \( \mathcal{H} = \mathcal{H}_{\text{full}} \), 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 \((x, y)\) is measured in terms of the worst loss incurred over an adversarial perturbation of \( x \) within a ball of a certain radius in a norm. For simplicity, we will consider perturbations in the \( \ell_2 \) norm \(|·|\).\(^1\) We will denote by \( \gamma \) the maximum magnitude of the allowed perturbations. Given \( \gamma > 0 \), a data point \((x, y)\), a function \( f \in \mathcal{H} \), and a margin-based loss \( \phi : \mathbb{R} \rightarrow \mathbb{R}^+ \), we define the adversarial loss of \( f \) at \((x, y)\) as

\[
\hat{\phi}(f, x, y) = \sup_{x' \in \mathcal{X}} \phi(y f(x')), \tag{7}
\]

The above naturally motivates \textit{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 \( \ell_1 \) ball or \( \ell_\infty \) ball, and in fact for any \( \ell_p \) 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 \( \ell_2 \) and work for other norms too.

\[\text{Figure 1: Illustration of } \mathcal{H}\text{-calibration and } \mathcal{H}\text{-consistency. Left: } \mathcal{H}\text{-calibration, for any } x \in \mathcal{X}, \text{ minimization of } C_0(f, x) \text{ can lead to that of } C_0(f, x). \text{ Right: } \mathcal{H}\text{-consistency, minimization of } R_{\ell_2}(f) \text{ can lead to that of } R_{\ell_2}(f). \text{ } \mathcal{H}\text{-consistency reduces to } \mathcal{H}\text{-calibration when the support of underlying distribution } P \text{ is the single point set } \{x\} \subset \mathcal{X}; \text{ Under the minimizability condition, } \mathcal{H}\text{-calibration would imply } \mathcal{H}\text{-consistency.}\]
We first study convex losses, which are often used for standard binary classification problems. We say that the supremum-based surrogate is convex if the function \( \phi \) in (7) is convex. When \( \phi \) is non-increasing, the following equality holds (Yin et al., 2019):

\[
\sup_{x' \in \mathcal{X}} \phi(yf(x')) = \phi\left( \inf_{x' \in \mathcal{X}} yf(x') \right).
\]

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

\[
\ell_{\gamma}(f, x, y) = \sup_{x' \in \mathcal{X}} \mathbb{I}_{yf(x') \leq 0} = \mathbb{I}_{yf(x') \leq 0} \inf_{x' \in \mathcal{X}} yf(x') \leq 0.
\]

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 \( \ell_1 \) 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 \( x \) in \( \mathcal{X} \), that is, if there exist \( f, g \in \mathcal{H} \) such that \( \inf_{|x'-x| \leq \gamma} f(x') > 0 \) and \( \sup_{|x'-x| \leq \gamma} g(x') < 0 \).

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 \( \ell \) is \( \mathcal{H} \)-calibrated with respect to \( \ell_{\gamma} \).

Moreover, we specifically study the following hypothesis sets that are regular for adversarial calibration: linear models: \( \mathcal{H}_{\text{lin}} = \{ x \rightarrow w \cdot x \mid \|w\| = 1 \} \), as in (Bao et al., 2020); generalized linear models: \( \mathcal{H}_g = \{ x \rightarrow g(w \cdot x) + b \mid \|w\| = 1, |b| \leq G \} \) where \( g \) is a non-decreasing function; the family of all measurable functions: \( \mathcal{H}_{\text{all}} \); and multi-layer neural networks: \( \mathcal{H}_{\text{NN}} = \{ x \rightarrow u \cdot \rho_1(W_n(\ldots\rho_2(W_1x + b_1) + b_2)\ldots) + b_n) \mid \|u\|_1 \leq \Lambda, \|W_1\| \leq W_1, \|b\|_1 \leq B \} \), where \( \rho_1 \) is an activation function; In the special case of \( g = (\cdot)_+ = \max(\cdot, 0) \), we denote the corresponding ReLU-based hypothesis set by \( \mathcal{H}_{\text{relu}} = \{ x \rightarrow (w \cdot x)_+ + b \mid \|w\| = 1, |b| \leq G \} \).

## 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 \( x_0 \in \mathcal{X} \) and \( f_0 \in \mathcal{H} \) such that \( f_0(x_0) = 0 \). If a margin-based loss \( \phi: \mathbb{R} \rightarrow \mathbb{R}_+ \) is convex, then it is not \( \mathcal{H} \)-calibrated with respect to \( \ell_{\gamma} \).

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. (2020) 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 \( x_0 \in \mathcal{X} \) and \( f_0 \in \mathcal{H} \) such that \( f_0(x_0) = 0 \). When \( g(\gamma) + G > 0 \) and \( g(\gamma) - G < 0 \), \( \mathcal{H}_{\phi} \) also satisfies this assumption. Verifying this condition on \( \mathcal{H}_{\phi} \) is straightforward for \( G \) sufficiently large.
We study losses of the type \( \phi : \mathbb{R} \to \mathbb{R} \), is convex, then \( \phi \) is not \( \mathcal{H} \)-calibrated with respect to \( \ell_\gamma \), 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 \( \tilde{\phi}(f, x, y) = \sup_{x' : \|x - x'\| \leq \gamma} \phi(y f(x')) \), with \( \phi \) convex, which are often used in practice as surrogates for the adversarial 0/1 loss. The following theorems present negative results for supremum-based convex surrogate losses for the common hypothesis sets \( \mathcal{H} \).

**Theorem 9.** Let \( \phi \) be a convex and non-increasing margin-based loss. Consider the surrogate loss defined by \( \hat{\phi}(f, x, y) = \sup_{x' : \|x - x'\| \leq \gamma} \phi(y f(x')) \). Then \( \hat{\phi} \) is not \( \mathcal{H} \)-calibrated with respect to \( \ell_\gamma \), 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 \( \phi \) is convex and non-increasing, then the surrogate loss defined by \( \hat{\phi}(f, x, y) = \sup_{x' : \|x - x'\| \leq \gamma} \phi(y f(x')) \) is not \( \mathcal{H} \)-calibrated with respect to \( \ell_\gamma \).

The theorems above provide 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 \( \phi \) be a convex and non-increasing margin-based loss. Consider the surrogate loss defined by \( \hat{\phi}(f, x, y) = \sup_{x' : \|x - x'\| \leq \gamma} \phi(y f(x')) \). Then \( \phi \) is not \( \mathcal{H} \)-calibrated with respect to \( \ell_\gamma \), for \( \mathcal{H} = \mathcal{H}_{\text{NN}}, \mathcal{H}_{\text{all}} \).

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 \( \delta_{max}(\epsilon, x, \eta) \) as defined in (6) of losses \( (\ell, \ell_\gamma) \) at \( \eta = \frac{1}{2}, \epsilon = \frac{1}{2} \) and distinguishing \( x_0 \in X \). Naturally, this requires us to understand the inner risk \( C_\ell(f, x, \eta) \) that in turn depends on the worst case perturbation of a given data point according to \( \ell \). Our key insight (Lemma 25) is that \( \delta_{max}(\epsilon, x_0, \eta) \) can be characterized by two quantities \( \Delta(f, x_0, \gamma) = \inf_{x' : \|x_0 - x'\| \leq \gamma} f(x') \), \( \overline{M}(f, x_0, \gamma) = \sup_{x' : \|x_0 - x'\| \leq \gamma} f(x') \). Requiring \( \delta_{max}(\frac{1}{2}, x_0, \frac{1}{2}) > 0 \) 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 \( \ell_\gamma \).

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. (2020) that satisfy the property that the generic conditional \( \phi \)-risk \( C_\gamma(t, \eta) \) 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., 2020), where we prove the theorem under the correct calibration definition.

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

\[
\phi(\gamma) + \phi(-\gamma) > \phi(t) + \phi(-t). \tag{10}
\]
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 \( \phi \) be bounded, continuous, non-increasing, and satisfy the property that \( C_\phi(t, \eta) \) 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 \( \phi \) is \( \mathcal{H}_g \)-calibrated with respect to \( \ell_1 \) if and only if for any \( 0 \leq t \leq 1 \),

\[
\phi(G - g(t)) + \phi(g(-t) - G) = \phi(g(t) + G) + \phi(-g(t) - G)
\]

and

\[
\min \{ \phi(A(t)) + \phi(-A(t)), \phi(A(t)) + \phi(-A(t)) \} > \phi(G - g(t)) + \phi(g(-t) - G),
\]

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

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 \( \phi \) be bounded, continuous, non-increasing, and satisfy the property that \( C_\phi(t, \eta) \) is quasi-concave in \( t \in \mathbb{R} \) for all \( \eta \in [0, 1] \). Assume that \( \phi(-G) > \phi(G) \). Then \( \phi \) is \( \mathcal{H}_{\text{relu}} \)-calibrated with respect to \( \ell_1 \) 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 \( \rho \)-margin loss \( \phi_\rho(t) := \min \{ 1, \max \{ 0, 1 - \frac{t}{\rho} \} \} \), \( \rho > 0 \), which is a generalization of the ramp loss (see, Mohri et al., 2018). We also define its supremum-based counterpart as \( \tilde{\phi}_\rho(f, x, y) := \sup_{t \in \mathbb{R}^d} \rho(yf(x)) \). Using Theorem 12, Theorem 13 and Corollary 14 in Section 4.3.1, we can conclude that the \( \rho \)-margin loss is calibrated under reasonable conditions for linear hypothesis sets and non-decreasing \( \gamma \)-based hypothesis sets, since \( \phi_\rho(t) \) is bounded, continuous, non-increasing, and satisfies \( C_{\phi_\rho}(t, \eta) \) is quasi-concave in \( t \in \mathbb{R} \) for all \( \eta \in [0, 1] \). This is stated formally below.

Theorem 15. The surrogate \( \tilde{\phi}_\rho \) is \( \mathcal{H}_{\text{lin}} \)-calibrated with respect to \( \ell_1 \) 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 \( \tilde{\phi}_\rho \) is \( \mathcal{H}_{\gamma} \)-calibrated with respect to \( \ell_1 \) if and only if for any \( 0 \leq t \leq 1 \),

\[
\tilde{\phi}_\rho(G - g(-t)) = \phi_\rho(g(t) + G) + \min \{ \phi_\rho(A(t)), \phi_\rho(-A(t)) \} > \phi_\rho(G - g(-t)),
\]

where \( A(t) = \max_{s \in [-t,t]} g(s) - g(s - \gamma) \) and \( \overline{A}(t) = \min_{s \in [-t,t]} g(s) - g(s + \gamma) \). Assume that \( G > 1 + \gamma \), then \( \tilde{\phi}_\rho \) is \( \mathcal{H}_{\text{relu}} \)-calibrated with respect to \( \ell_1 \) 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 \( \rho \)-margin loss is indeed \( \mathcal{H}_\gamma \)-calibrated, where \( \mathcal{H} \) is any symmetric hypothesis set.

Theorem 16. Let \( \mathcal{H} \) be a symmetric hypothesis set, then \( \tilde{\phi}_\rho \) is \( \mathcal{H}_\gamma \)-calibrated with respect to \( \ell_1 \).

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 \( \mathcal{H}_\gamma \)-calibrated with respect to \( \ell_1 \), for \( \mathcal{H} = \mathcal{H}_{\text{lin}}, \mathcal{H}_{\text{NN}}, \) and \( \mathcal{H}_{\text{all}} \).

The results of this section suggest that the \( \rho \)-margin loss and supremum-based \( \rho \)-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 surrogate losses to be \( \mathcal{H}_\gamma \)-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 $\rho$-margin loss and supremum-based $\rho$-margin loss present an intriguing possibility. Bao et al. (2020) 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 $\phi$ is $\mathcal{H}_{\text{lin}}$-consistent with respect to $\ell_\gamma$. Furthermore, for any continuous and non-increasing margin-based loss $\phi$, surrogates of the form

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

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

The above theorem is proven in Appendix E.6. In particular, Theorem 18 contradicts the $\mathcal{H}$-consistency claim of Bao et al. (2020) 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 which requires the Bayes ($\ell_0$, $\mathcal{H}$)-risk to be zero.

**Definition 19 ($\mathcal{H}$-realizability).** A distribution $\mathcal{P}$ over $X \times Y$ 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 $\mathcal{P}(x,y) \cdot \text{sgn}(f(x)) = y = 1$.

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 $X \times Y$ and $\mathcal{H}$ a hypothesis set for which $\mathcal{R}_{\ell_\gamma, \mathcal{H}}^* = 0$. Let $\phi$ be a margin-based loss. If for $\eta \geq 0$, there exists $f^* \in \mathcal{H} \subset \mathcal{H}_{\text{all}}$ such that $\mathcal{R}_\phi(f^*) \leq \mathcal{R}_{\phi, \mathcal{H}_{\text{all}}}(\eta) + \eta < +\infty$ and $\phi$ is $\mathcal{H}$-calibrated with respect to $\ell_\gamma$, 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 $\mathcal{R}_{\ell_\gamma, \mathcal{H}_{\text{lin}}}(f) = 0$ can be further relaxed.

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

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

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 $\rho$-margin loss is consistent with respect to $\ell_\gamma$ for all distributions that satisfy our realizability assumptions.
We generate data points \( x \). We can also extend the above to obtain
with respect to \( \gamma \). Given a distribution \( \mathcal{P} \) over \( \mathcal{X} \times \mathcal{Y} \) that satisfies \( \mathcal{R}_{\ell_{\gamma}}^* = 0 \) and there exists \( f^* \in \mathcal{H}_g \) such that \( \mathcal{R}_{\phi_{\gamma}}(f^*) = \mathcal{R}_{\phi_{\gamma}}^* \). If \( G > 1 + \gamma \) and \( G \geq \gamma \), then \( \phi_{\gamma} \) is \( \mathcal{H}_{\phi_{\gamma}} \)-consistent wrt \( \ell_{\gamma} \) for all distribution \( \mathcal{P} \) over \( \mathcal{X} \times \mathcal{Y} \) that satisfies \( \mathcal{R}_{\ell_{\gamma}}^* = 0 \) and there exists \( f^* \in \mathcal{H}_{\phi_{\gamma}} \) such that \( \mathcal{R}_{\phi_{\gamma}}(f^*) = \mathcal{R}_{\phi_{\gamma}}^* \).

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 \( \mathcal{R}_{\ell_{\gamma}}^* = 0 \). Let \( \phi \) be a non-increasing margin-based loss. If there exists \( f^* \in \mathcal{H} \) such that \( \mathcal{R}_{\phi}(f^*) = \mathcal{R}_{\phi}^* \) and \( \tilde{\phi}(f, x, y) = \sup_{x' \in \mathcal{X}} \phi(x', y) \) is \( \mathcal{H} \)-calibrated with respect to \( \ell_{\gamma} \), then for all \( \epsilon > 0 \) there exists \( \delta > 0 \) such that for all \( f \in \mathcal{H} \) we have

\[
\mathcal{R}_{\phi}(f) < \mathcal{R}_{\phi}^* + \delta \implies \mathcal{R}_{\ell_{\gamma}}(f) < \mathcal{R}_{\ell_{\gamma}}^* + \epsilon.
\]

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 \( \rho \)-margin loss is also \( \mathcal{H} \)-consistent with respect to \( \ell_{\gamma} \) for all distributions that satisfy our realizability assumptions.

**Theorem 24.** Let \( \mathcal{H} \) be a symmetric hypothesis set, then \( \phi_{\gamma} \) is \( \mathcal{H} \)-consistent with respect to \( \ell_{\gamma} \) for all distributions \( \mathcal{P} \) over \( \mathcal{X} \times \mathcal{Y} \) that satisfy: \( \mathcal{R}_{\ell_{\gamma}}^* = 0 \) and there exists \( f^* \in \mathcal{H} \) such that \( \mathcal{R}_{\phi_{\gamma}}(f^*) = \mathcal{R}_{\phi_{\gamma}}^* \) and is achieved by \( w_{\phi_{\gamma}} = (\cos(\theta), \sin(\theta))^T \) with \( \theta \approx 0.7855 \). The results obtained by optimizing the different surrogate losses are reported in Table 1(a) and the plots for

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

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

<table>
<thead>
<tr>
<th>φ</th>
<th>R_0*(f)</th>
<th>θ_0</th>
<th>R_{lin-cal}</th>
<th>R_{lin-cons}</th>
<th>φ</th>
<th>R_0*(f)</th>
<th>θ_0</th>
<th>R_{lin-cal}</th>
<th>R_{lin-cons}</th>
</tr>
</thead>
<tbody>
<tr>
<td>hinge</td>
<td>0.5257</td>
<td>0.1420</td>
<td>X</td>
<td>X</td>
<td>hinge</td>
<td>0.0781</td>
<td>0.6907</td>
<td>1.3548</td>
<td>X</td>
</tr>
<tr>
<td>ramp</td>
<td>0.5263</td>
<td>0.1288</td>
<td>✓</td>
<td>X</td>
<td>ramp</td>
<td>0.0781</td>
<td>0.3454</td>
<td>1.3548</td>
<td>✓</td>
</tr>
<tr>
<td>sig</td>
<td>0.5261</td>
<td>0.1320</td>
<td>✓</td>
<td>X</td>
<td>sig</td>
<td>0.0777</td>
<td>0.4247</td>
<td>1.3498</td>
<td>✓</td>
</tr>
<tr>
<td>log</td>
<td>0.5258</td>
<td>0.1414</td>
<td>X</td>
<td>X</td>
<td>log</td>
<td>0.0763</td>
<td>0.8078</td>
<td>1.3341</td>
<td>X</td>
</tr>
<tr>
<td>γ</td>
<td>0.0111</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>γ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>δ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>δ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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 $H_{lin}$-consistent with respect to $\ell_0$ 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 $\ell_0$. This shows that indeed calibrated surrogates may not be consistent and contradicts Figure 12 of (Bao et al., 2020). The discrepancy results from an incorrect calculation of the adversarial Bayes risk in (Bao et al., 2020).

Next, we justify the realizability assumptions made in Section 5.2 for obtaining $H$-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 $ρ > γ$. 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 $H_{lin}$-calibrated while $φ_{ramp}$, $φ_{sig}$ and $φ_2$ are $H_{lin}$-calibrated with respect to $\ell_γ$.

Let $I_γ = \sqrt{1 - γ^2}$ and consider: $P(Y = 1) = P(Y = -1) = \frac{1}{2}$, and $X | Y = 1$ is the uniform distribution on the line segment $\{(γ, z) | z \in [0, I_γ]\}$ and $X | Y = -1$ is the uniform distribution on the line segment $\{(-γ, z) | z \in [-I_γ, 0]\}$ where $γ = γ + \frac{1 - γ}{100} = \frac{1 + 99γ}{100}, γ ∈ (0, 1)$. We choose $γ = 0.1$ and set $w^* = (1, 0)'$. It is easy to check that $w^*$ achieves the Bayes ($\ell_γ, H_{lin}$)-risk $R_{\ell_γ,H_{lin}} = 0$. The results for the six different surrogate losses are indicated in Table 1(b) and the plot for 5,000 samples are shown in Figure 2. For $φ_{hinge}$, $φ_{ramp}$, $φ_{sig}$ and $φ_{log}$, the Bayes ($φ_{sur}$, $H_{lin}$)-risk $R_{φ_{sur},H_{lin}} = 0$. Table 1(b) shows that they are not $H_{lin}$-consistent with respect to $\ell_γ$. For $φ_1$ and $φ_2$, the Bayes ($φ_{sur}$, $H_{lin}$)-risk $R_{φ_{sur},H_{lin}} = 0$. Table 1(b) shows that $φ_1$ is not $H_{lin}$-consistent (recall that $φ_1$ is not calibrated) but $φ_2$ is $H_{lin}$-consistent for this distribution. Hence, even when $R_{φ_{sur},H_{lin}} = 0$, unless a condition is also imposed on $R_{φ_{sur},H_{lin}}$, one cannot expect consistency, thereby justifying our realizability assumption. Note that $R_{φ_{sur},H_{lin}} = R_{φ_{sur},H_{lin}} = 0$ is a special case verifying the conditions of Theorem 20 for $η = 0$. For this distribution, $φ_{ramp}$ is not $H_{lin}$-consistent while $φ_2$ is $H_{lin}$-consistent, although both are $H_{lin}$-calibrated. We compare them in Figure 3, showing that minimizing $H_{lin}$-consistent surrogate $φ_2$ minimizes the adversarial generalization error for large sample sizes but the same does not hold for non $H_{lin}$-consistent surrogate $φ_{ramp}$.

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


