Theoretically Grounded Loss Functions and Algorithms for Adversarial Robustness

Pranjal Awasthi
Google Research

Anqi Mao
Courant Institute

Mehryar Mohri
Google Research and Courant Institute

Yutao Zhong
Courant Institute

Abstract

Adversarial robustness is a critical property of classifiers in applications as they are increasingly deployed in complex real-world systems. Yet, achieving accurate adversarial robustness in machine learning remains a persistent challenge and the choice of the surrogate loss function used for training a key factor. We present a family of new loss functions for adversarial robustness, *smooth adversarial losses*, which we show can be derived in a general way from broad families of loss functions used in multi-class classification. We prove strong $\mathcal{H}$-consistency theoretical guarantees for these loss functions, including multi-class $\mathcal{H}$-consistency bounds for sum losses in the adversarial setting. We design new regularized algorithms based on the minimization of these principled smooth adversarial losses (PSAL). We further show through a series of extensive experiments with the CIFAR-10, CIFAR-100 and SVHN datasets that our PSAL algorithm consistently outperforms the current state-of-the-art technique, TRADES, for both robust accuracy against ℓ_∞-norm bounded perturbations and, even more significantly, for clean accuracy. Finally, we prove that, unlike PSAL, the TRADES loss in general does not admit an $\mathcal{H}$-consistency property.

1 INTRODUCTION

Adversarial robustness is a critical property of classifiers in applications as they are increasingly deployed in complex real-world systems. A classifier misclassifying a traffic sign, as a result of a minor variation, which may be the presence of a small label on the sign, may result in traffic incidents or worse, human injuries, when used for example with self-driving cars. Similar undesirable consequences may result from the lack of robustness of classifiers in medical diagnosis, speech recognition, fraud detection and many other areas.

Yet, achieving accurate adversarial robustness in machine learning remains a persistent challenge theoretically and algorithmically. 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 have been shown to be susceptible to imperceptible perturbations (Szegedy et al., 2013) and their adversarial accuracy remains substantially below their *clean accuracy*, their accuracy for the standard classification loss.

For adversarial robustness, the standard zero-one loss function used in learning is typically replaced by 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 ℓ_p distance of x (Goodfellow et al., 2014; Madry et al., 2017; Tsipras et al., 2018; Carlini and Wagner, 2017). The design of robust algorithms relies on surrogate losses since the optimization of the adversarial loss is intractable for most hypothesis sets. But, which surrogate losses should be used and which benefit from theoretical guarantees?

A key criterion for surrogate losses is their Bayes-consistency, which has been extensively studied in both binary and multi-class non-adversarial classification (Zhang, 2004; Bartlett et al., 2006; Tewari and Bartlett, 2007; Steinwart, 2007). More recently, Awasthi et al. (2021a) gave an extensive study of consistency in the adversarial setting, which they showed to be technically more complex and requiring new proofs. Bayes-consistency is a property related to the family of all measurable functions, which is much broader than the hypothesis set used by learning algorithms. But, remarkably, the authors also gave a series of results for $\mathcal{H}$-*consistency*, that is consistency restricted to the use of a specific hypothesis set $\mathcal{H}$ (Long and Servedio, 2013). These results rule out, in particular, several types of surrogates losses frequently used in applications to achieve

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

adversarial robustness. Can these results further guide the choice of effective surrogate losses for adversarial robustness?

Bayes-consistency or even $\mathcal{H}$-consistency for a specific hypothesis set $\mathcal{H}$ is only an asymptotic property, which does not provide any guarantee for approximate minimization of losses based on finite samples. More favorable guarantees called *$\mathcal{H}$-consistency bounds* were recently derived (Awasthi et al., 2022a). These are hypothesis set-specific guarantees that are stronger than $\mathcal{H}$-consistency since they do not just hold only asymptotically. Can we design surrogate losses for adversarial robustness benefiting from such strong theoretical guarantees? Can such loss functions be used to design effective algorithms?

This paper deals precisely with these questions. We present a family of new loss functions, *smooth adversarial losses*, which we show can be derived in a general way from broad families of loss functions used in multi-class classification: *max losses* (Crammer and Singer, 2001), *sum losses* (Weston and Watkins, 1998), or *constrained losses* (Lee et al., 2004). These loss functions admit a non-adversarial loss term and a smooth adversarial loss term based on the Lipschitz property of the auxiliary function in the definition of the multi-class loss.

We prove strong theoretical guarantees based on $\mathcal{H}$-consistency for these loss functions, including multi-class $\mathcal{H}$-consistency bounds for the family of sum losses in the adversarial scenario. These guarantees are more relevant to most robustness problems, which are multi-class classification tasks, than previous binary classification results given by Awasthi et al. (2022a). Their analysis is also more challenging and requires novel proof techniques. We also give new regularized algorithms based on the minimization of these principled smooth adversarial sum losses (PSAL). We show that PSAL consistently outperforms TRADES for both robust classification and clean accuracy, with an even more significant improvement of the clean accuracy, on CIFAR-10, CIFAR-100 and SVHN against ℓ_∞-norm bounded perturbations of size $\gamma = 8/255$. These results establish the new state-of-the-art benchmarks in these tasks in the scenario where no generated data, extra data or extra data augmentation is used.

The paper is structured as follows. In Section 3.1, we point out that the surrogate losses frequently used in practice in adversarial robust classification, the adversarial counterpart of the cross-entropy loss (Madry et al., 2017) and TRADES (Zhang et al., 2019) do not admit *$\mathcal{H}$-consistency*, that is, minimizing these surrogate losses over a hypothesis set $\mathcal{H}$, may not always lead to minimizing the adversarial zero-one loss over $\mathcal{H}$.

This motivates our design of a new family of surrogate losses, *smooth adversarial losses* in Section 3.2. Here, we provide a detailed derivation of smooth adversarial losses corresponding to each of the following three families of multi-class classification losses defined in the non-adversarial setting: *max losses* (Crammer and Singer, 2001), *sum losses* (Weston and Watkins, 1998), and *constrained losses* (Lee et al., 2004).

In Section 4, we show that our smooth adversarial losses benefit from $\mathcal{H}$-consistency guarantees. To obtain guarantees for our smooth adversarial loss, we first prove a multi-class adversarial $\mathcal{H}$-consistency bound for the adversarial sum loss. The guarantees based on this bound provide a strong support for our smooth loss minimization algorithm, PSAL, described in Section 5. We further discuss in that section various choices for auxiliary functions used in the objective function of PSAL. In Section 6, we further analyze the surrogate loss TRADES and prove that there exist learning problems in both the realizable and non-realizable cases for which TRADES lacks the $\mathcal{H}$-consistency guarantee while our smooth adversarial loss admits that guarantee. In Section 7, we report the results of several experiments comparing with the current state-of-the-art ones using TRADES that demonstrate the empirical significance of our PSAL algorithm. We start with some basic definitions and notation (Section 2).

2 PRELIMINARIES

We denote by $\mathcal{X}$ the input space and by $\mathcal{Y}$ the set of labels which we define by $\mathcal{Y} = \{-1, +1\}$ in binary classification, by $\mathcal{Y} = \{1, \ldots, c\}$ in multi-class classification with $c > 2$ classes. We denote by $\mathcal{H}$ a hypothesis set of functions mapping from $\mathcal{X}$ to $\mathbb{R}$ in the binary setting, from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$ in the multi-class setting. We denote by $\mathsf{h}(x)$ the label prediction made by h: in binary classification, $\mathsf{h}(x) = \operatorname{sign}(h(x))$ with the convention $\operatorname{sign}(0) = +1$; in multi-class classification, $\mathsf{h}(x) = \operatorname{argmax}_{y \in \mathcal{Y}} h(x, y)$ with an arbitrary but fixed strategy for breaking the ties.

Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ according to which samples are drawn i.i.d. We denote by $\mathcal{R}_{\ell_{0-1}}(h) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell_{0-1}(h, x, y)]$ the generalization error of a hypothesis $h \in \mathcal{H}$, where $\ell_{0-1}(h, x, y) = \mathbb{1}_{\mathsf{h}(x) \neq y}$ is the $0/1$ loss. The generalization and best-in-class errors for a surrogate loss $\ell \colon \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ are similarly defined by $\mathcal{R}_\ell(h) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell(h, x, y)]$ and $\mathcal{R}^*_{\ell, \mathcal{H}} = \inf_{h \in \mathcal{H}} \mathcal{R}_\ell(h)$.

3 SMOOTH ADVERSARIAL LOSSES

3.1 Motivation

In adversarially robust classification, the benchmark criterion is the *adversarial* $0/1$ *loss*, which is the maximum loss incurred over an adversarial perturbation of the example. Let $\gamma \in (0, 1)$ be the maximum magnitude allowed for perturbations and let $\|\cdot\|$ denote the norm adopted, which is typically an ℓ_p-norm, $p \in [1, +\infty]$. Then, the adversarial 0/1 loss ℓ_γ is defined as follows in the binary and multi-class classification settings (Goodfellow et al., 2014; Madry et al., 2017; Shafahi et al., 2019; Wong et al., 2020; Awasthi et al., 2023):

- *binary:* $\ell_\gamma(h, x, y) = \sup_{x' : \|x - x'\| \le \gamma} \mathbb{1}_{y h(x') \le 0}$;
- *multi-class:* $\ell_\gamma(h, x, y) = \sup_{x' : \|x - x'\| \le \gamma} \mathbb{1}_{\rho_h(x', y) \le 0}$,

where $\rho_h(x, y) = h(x, y) - \max_{y' \ne y} h(x, y')$ is the multi-class classification margin. As with the non-adversarial 0/1 loss, optimizing the adversarial loss ℓ_γ directly is intractable. Thus, most algorithms resort to a surrogate loss instead. But, how should this surrogate loss be defined? One commonly adopted method consists of using a surrogate loss ℓ for the standard 0/1 loss and of defining an adversarial surrogate loss $\widetilde{\ell}$ as the supremum-based version of ℓ:

$$\widetilde{\ell}(h, x, y) = \sup_{x' : \|x - x'\| \le \gamma} \ell(h, x', y). \tag{1}$$

As an example, $\widetilde{\ell}_{\text{xent}}$, the adversarial counterpart of the cross-entropy loss ℓ_{xent} is defined as follows:

$$\widetilde{\ell}_{\text{xent}}(h, x, y) = \sup_{x' : \|x - x'\| \le \gamma} \ell_{\text{xent}}(h, x', y), \tag{2}$$

where ℓ_{xent} is the cross-entropy loss (or log-loss): $\ell_{\text{xent}}(h, x, y) = -\log(h(x, y))$, subject to the requirements $h(x, y) \ge 0$ for any $y \in \mathcal{Y}$ and $\sum_{y \in \mathcal{Y}} h(x, y) = 1$, which are fulfilled for neural network hypotheses, when using the softmax activation function in the output layer.

While such surrogate losses are natural, formulation $\widetilde{\ell}_{\text{xent}}$ and other similar ones based on a convex loss ℓ suffer from a serious drawback, which may explain the persistent large empirical gap observed empirically between the natural and adversarial accuracies (Madry et al., 2017): as shown by Awasthi et al. (2021a), even in the binary scenario, no convex supremum-based loss admits the key property of *$\mathcal{H}$-consistency* (see Section 4 for a formal definition and description of this property). Thus, in general, minimizing such surrogate losses, including the adversarial cross-entropy loss $\widetilde{\ell}_{\text{xent}}$, over a hypothesis set $\mathcal{H}$, may not lead to minimizing ℓ_γ over $\mathcal{H}$.

An alternative surrogate loss adopted in the adversarial setting is TRADES (Zhang et al., 2019), which is based on the following formulation:

$$\begin{aligned} &\widetilde{\ell}_{\text{TRADES}}(h, x, y) \\ &\quad = \ell_{\text{xent}}(h, x, y) + \sup_{x' : \|x - x'\| \le \gamma} \mathcal{L}_{\text{xent}}(h, x, x')/\lambda, \end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{xent}}(h, x, x') = -\sum_{y \in \mathcal{Y}} h(x, y) \log(h(x', y))$ is the cross-entropy of $h(x, \cdot)$ and $h(x', \cdot)$, and where $\lambda > 0$ is a constant. Minimizing a regularized objective based on $\widetilde{\ell}_{\text{TRADES}}$ has been shown empirically to improve upon minimizing $\widetilde{\ell}_{\text{xent}}$ in adversarial training. In fact, this has led to the current state-of-the-art adversarial accuracy in multiple tasks (Gowal et al., 2020). We will show in Section 6, however, that, as with the adversarial cross-entropy loss $\widetilde{\ell}_{\text{xent}}$, $\widetilde{\ell}_{\text{TRADES}}$ does not benefit from $\mathcal{H}$-consistency guarantees. This suggests the need for alternative surrogate losses in the adversarially robust classification with stronger theoretical guarantees.

3.2 New Surrogate Losses

In this section, we introduce a general family of surrogate losses, *smooth adversarial losses*, which we will show benefit from an $\mathcal{H}$-consistency guarantee.

We begin with binary classification and then extend the derivation to multi-class classification. Let $\widetilde{\Phi}$ be a supremum-based margin loss based on the *auxiliary function* Φ, that is $\widetilde{\Phi}(h, x, y) = \sup_{x' : \|x - x'\| \le \gamma} \Phi(y h(x'))$, for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Assume that Φ is non-increasing and μ-Lipschitz. Then, the following decomposition and inequality hold:

$$\begin{aligned} &\widetilde{\Phi}(h, x, y) \\ &= \Phi(y h(x)) + \Phi\left(\inf_{x' : \|x - x'\| \le \gamma} y h(x')\right) - \Phi(y h(x)) \\ &\le \Phi(y h(x)) + \nu \left| y h(x) - \inf_{x' : \|x - x'\| \le \gamma} y h(x') \right|, \\ &\hspace{15em} (\Phi\ \mu\text{-Lipschitz}) \end{aligned}$$

for any $\nu \ge \mu$. We will refer to a loss function defined by the last expression as a *smooth adversarial loss* and denote it by Φ_{smooth}. The loss function admits a non-adversarial loss term and a smooth adversarial loss term based on the Lipschitz property of the auxiliary function Φ.

Let Φ be a non-increasing and Lipschitz auxiliary function. For multi-class classification, we can similarly derive the smooth adversarial loss corresponding to each of the following three families of multi-class classification losses defined in the non-adversarial setting:

- *max loss*: $\Phi^{\max}(h, x, y) = \Phi(\rho_h(x, y))$ where $\rho_h(x, y) = h(x, y) - \max_{y' \ne y} h(x, y')$, e.g. (Crammer and Singer, 2001);
- *sum loss*: $\Phi^{\text{sum}}(h, x, y) = \sum_{y' \ne y} \Phi(\Delta_h(x, y, y'))$ where $\Delta_h(x, y, y') = h(x, y) - h(x, y')$, e.g. (Weston and Watkins, 1998); and
- *constrained loss*: $\Phi^{\text{cstnd}}(h, x, y) = \sum_{y' \ne y} \Phi(-h(x, y'))$ subject to the constraint on the sum of the scores $\sum_{y \in \mathcal{Y}} h(x, y) = 0$, e.g. (Lee et al., 2004).

We give a detailed derivation in Appendix A. Table 1 gives the general form of the smooth adversarial loss for each of these three families, where $\|\cdot\|_2$ is the ℓ_2 norm, $\overline{\Delta}_h(x, y)$ denotes the $(c - 1)$-dimensional

Table 1: Multi-class classification losses and the corresponding smooth adversarial losses.

Multi-class loss	Smooth adversarial loss
Max loss	$\Phi^{\max}_{\text{smooth}} = \Phi^{\max}(h,x,y) + \nu\left\lvert\rho_h(x,y) - \inf_{x':\lVert x-x'\rVert\le\gamma} \rho_h(x',y)\right\rvert$
Sum loss	$\Phi^{\text{sum}}_{\text{smooth}} = \Phi^{\text{sum}}(h,x,y) + \nu \sup_{x':\lVert x-x'\rVert\le\gamma} \left\lVert \overline{\Delta}_h(x',y) - \overline{\Delta}_h(x,y)\right\rVert_2$
Constrained loss	$\Phi^{\text{cstnd}}_{\text{smooth}} = \Phi^{\text{cstnd}}(h,x,y) + \nu \sup_{x':\lVert x-x'\rVert\le\gamma} \left\lVert \overline{h}(x',y) - \overline{h}(x,y)\right\rVert_2$

vector $\big(\Delta_h(x,y,1),\ldots,\Delta_h(x,y,y-1),\Delta_h(x,y,y+1),\ldots,\Delta_h(x,y,c)\big)$, and $\overline{h}(x,y)$ the $(c-1)$-dimensional vector $\big(h(x,1),\ldots,h(x,y-1),h(x,y+1),\ldots,h(x,c)\big)$. As in binary classification, the loss functions in Table 1 admit an additive smooth adversarial loss term complementing the non-adversarial loss term.

A family of common auxiliary functions Ψ_ρ generalizing the ρ-margin loss $\Phi_\rho(t) = \min\Big\{\max\Big\{0, 1-\frac{t}{\rho}\Big\}, 1\Big\}$ (Mohri et al., 2018) is defined by the following:

$$\Psi_\rho(t) = \begin{cases} \Phi_\rho(t), & t<0 \text{ or } t>\rho \\ f^\mu(t), & t\in[0,\rho]. \end{cases}$$

Here, f^μ is a non-increasing and μ-Lipschitz function on $[0,\rho]$ with $f^\mu(0)=1$ and $f^\mu(\rho)=0$. Thus, by definition Ψ_ρ is continuous, non-increasing and μ-Lipschitz. Furthermore, Ψ_ρ coincides with the ρ-margin loss Φ_ρ when f^μ is the $\frac{1}{\rho}$-Lipschitz function $t\mapsto -\frac{t}{\rho}+1$. In the next section, we will show that adversarial sum losses using Ψ_ρ as auxiliary functions benefit from strong $\mathcal{H}$-consistency guarantees. It will further provide similar guarantees for smooth adversarial losses when using as auxiliary functions any convex and smooth upper bounds of Ψ_ρ.

4 $\mathcal{H}$-CONSISTENCY GUARANTEES OF SMOOTH ADVERSARIAL LOSSES

In this section, we will show that smooth adversarial losses with general auxiliary functions benefit from *$\mathcal{H}$-consistency guarantees*. We will focus on the family of sum losses, since max losses are not differentiable and since constrained losses impose a restriction that is not compatible with the standard use of the softmax function with neural network hypotheses, which make the optimization usually more difficult for those losses. Let us emphasize, however, that our results including the theoretical analysis of sum losses can be extended to the study of other families, in particular the constrained loss family.

Given a hypothesis set $\mathcal{H}$, an *$\mathcal{H}$-consistency bound* for a surrogate loss ℓ_1 of a target loss function ℓ_2 is an inequality of the form

$$\forall h\in\mathcal{H},\ \mathcal{R}_{\ell_2}(h) - \mathcal{R}^*_{\ell_2,\mathcal{H}} \le f\big(\mathcal{R}_{\ell_1}(h) - \mathcal{R}^*_{\ell_1,\mathcal{H}}\big), \quad (4)$$

where $f\colon\mathbb{R}_+\to\mathbb{R}_+$ is a non-increasing function (Awasthi et al., 2022a). Such a bound therefore relates the minimization of the estimation error for the surrogate loss ℓ_1 to that of the target loss ℓ_2 in a quantitative way.

Such guarantees are stronger than the *$\mathcal{H}$-consistency* property discussed in (Long and Servedio, 2013; Zhang and Agarwal, 2020; Awasthi et al., 2021a,b), which only requires that, *asymptotically*, the minimization of the surrogate loss estimation error results in that of the target estimation loss:

$$\lim_{n\to+\infty} \mathcal{R}_{\ell_1}(h_n) - \mathcal{R}^*_{\ell_1,\mathcal{H}} = 0 \Rightarrow \lim_{n\to+\infty} \mathcal{R}_{\ell_2}(h_n) - \mathcal{R}^*_{\ell_2,\mathcal{H}} = 0 \quad (5)$$

for all probability distributions and sequences of $\{h_n\}_{n\in\mathbb{N}}\subset\mathcal{H}$. When $\mathcal{H}$ is the family of all measurable functions $\mathcal{H}_{\text{all}}$, this coincides with the standard *Bayes-consistency*. Since they are not just asymptotic bounds, $\mathcal{H}$-consistency bounds are stronger than Bayes-consistency, $\mathcal{H}$-calibration or $\mathcal{H}$-consistency, and more informative than excess error bounds derived for $\mathcal{H}$ being the family of all measurable functions (Zhang, 2004; Bartlett et al., 2006).

To present our bounds, we first need to introduce some concepts and definitions. Given a distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$ with conditional probability $p(x,y)=\mathcal{D}(Y=y\,|\,X=x)$, the *conditional ℓ-risk* and the *minimal conditional ℓ-risk* of a loss function ℓ are defined as follows:

$$\mathcal{C}_\ell(h,x) = \sum_{y\in\mathcal{Y}} p(x,y)\ell(h,x,y) \quad \mathcal{C}^*_{\ell,\mathcal{H}}(x) = \inf_{h\in\mathcal{H}} \mathcal{C}_\ell(h,x).$$

These correspond to the error or best-in-class error, conditioned on a specific point x. For convenience, we also define the *conditional regret* $\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x) = \mathcal{C}_\ell(h,x) - \mathcal{C}^*_{\ell,\mathcal{H}}(x)$ and the *conditional ϵ-regret* $[\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x)]_\epsilon$, where we use the notation $[t]_\epsilon = t\mathbb{1}_{t>\epsilon}$.

An important quantity that appears in our bounds is the *minimizability gap*, defined by $\mathcal{M}_{\ell,\mathcal{H}} = \mathcal{R}^*_{\ell,\mathcal{H}} - \mathbb{E}_X\big[\mathcal{C}^*_{\ell,\mathcal{H}}(x)\big]$. By the super-additivity of the infimum, the minimizability gap is always non-negative. Its value depends only on the hypothesis set $\mathcal{H}$ and the loss function ℓ. As an example, for multi-class 0/1 loss functions, $\mathcal{M}_{\ell,\mathcal{H}}$ is zero for any distribution $\mathcal{D}$ and the hypothesis set of all measurable functions.

We will say that a hypothesis set $\mathcal{H}$ is *symmetric*, when there exists a family $\mathcal{F}$ of functions f mapping from $\mathcal{X}$ to $\mathbb{R}$ such that $\{[h(x,1),\ldots,h(x,c)] : h\in\mathcal{H}\} = \{[f_1(x),\ldots,f_c(x)] : f_1,\ldots,f_c\in\mathcal{F}\}$ and $|\{f(x)\colon f \in$

$\mathcal{F}\}| \geq 2$ for any $x \in \mathcal{X}$. Note that common hypothesis sets, such as the family of all measurable functions $\mathcal{H}_{\mathrm{all}} = \{(x,y) \mapsto h_y(x) \mid h_y : \mathcal{X} \to \mathbb{R} \text{ is measurable}\}$ and that of multi-layer neural networks, $\mathcal{H}_{\mathrm{NN}} = \{(x,y) \mapsto u_y \cdot \rho_n(W_{y,n}(\cdots\rho_2(W_{y,2}\rho_1(W_{y,1}x + b_{y,1}) + b_{y,2})\cdots) + b_{y,n}) \mid \|u_y\|_1 \leq \Lambda, \|W_{y,j}\| \leq W, \|b_{y,j}\|_1 \leq B, j \in [n]\}$, where ρ_j is an activation function and Λ, W, B are positive constants, are all symmetric.

We say that a hypothesis set $\mathcal{H}$ is *locally ρ-consistent* if for any $x \in \mathcal{X}$, there exists a hypothesis $h \in \mathcal{H}$ inducing the same ordering of the labels for any $x' \in \{x' : \|x - x'\| \leq \gamma\}$ and such that $\inf_{x' : \|x-x'\| \leq \gamma} |h(x', i) - h(x', j)| \geq \rho > 0$ for any $i \neq j \in \mathcal{Y}$ and $\sup_{x' : \|x-x'\| \leq \gamma} |h(x', y)| < \infty$ for any $y \in \mathcal{Y}$. The locally ρ-consistency condition only requires the existence of one such hypothesis given a point $x \in \mathcal{X}$ and thus is very general. Indeed, the family of all measurable functions, that of linear models and that of multi-layer neural networks commonly used in practice all verify the condition for a suitable choice of ρ. For example, for $\mathcal{H}_{\mathrm{NN}}$, we can consider those hypotheses such that $W_{y,j} = 0$ for all $y \in \mathcal{Y}$ and $j \in [n]$, which induce the same ordering of the labels for any x. Then, it suffices to find one such hypothesis such that $|u_i \cdot \rho_n(b_{i,n}) - u_j \cdot \rho_n(b_{j,n})| \geq \rho$ for any $i \neq j \in \mathcal{Y}$, which can be easily verified with suitable choices of ρ, u_y and $b_{y,n}$ for $y \in \mathcal{Y}$ subject to the norm constraints.

For convenience, we denote by $\sigma[h]$ the softmax output of a hypothesis h, defined as $\sigma[h](x,y) = \frac{e^{h(x,y)}}{\sum_{y' \in \mathcal{Y}} e^{h(x,y')}}$. For any hypothesis set $\mathcal{H}$, we denote by $\mathcal{H}^{\mathrm{softmax}}$ the hypothesis set that consists of all the softmax output of hypotheses in $\mathcal{H}$, defined as $\mathcal{H}^{\mathrm{softmax}} = \{\sigma[h] \mid h \in \mathcal{H}\}$. Note that if a hypothesis set $\mathcal{H}$ is symmetric, then $\mathcal{H}^{\mathrm{softmax}}$ is also symmetric. Moreover, if $\mathcal{H}$ is locally ρ-consistent for some $\rho > 0$, then there also exists $\rho' > 0$ such that $\mathcal{H}^{\mathrm{softmax}}$ is locally ρ'-consistent. Indeed, for any $x \in \mathcal{X}$, $\sigma[h]$ and h have the same ordering of labels. Let $h \in \mathcal{H}$ be the hypothesis that verifies the locally ρ-consistent condition and take $\rho' = \frac{\rho}{\sum_{y \in \mathcal{Y}} e^{\sup_{x' : \|x-x'\| \leq \gamma} |h(x', y)|}} > 0$. Then, for any $i \neq j \in \mathcal{Y}$, the following inequalities hold:

$$\inf_{x' : \|x-x'\| \leq \gamma} \left|\sigma[h](x', i) - \sigma[h](x', j)\right| \geq \frac{\inf_{x' : \|x-x'\| \leq \gamma} |h(x', i) - h(x', j)|}{\sum_{y \in \mathcal{Y}} e^{\sup_{x' : \|x-x'\| \leq \gamma} |h(x', y)|}} \geq \rho'.$$

Thus, since the hypothesis sets commonly used in practice, e.g. $\mathcal{H}_{\mathrm{all}}$ and $\mathcal{H}_{\mathrm{NN}}$, are symmetric and locally ρ-consistent for some $\rho > 0$, their counterparts with the softmax operator, e.g. $\mathcal{H}_{\mathrm{all}}^{\mathrm{softmax}}$ and $\mathcal{H}_{\mathrm{NN}}^{\mathrm{softmax}}$, are also symmetric and locally ρ-consistent for some $\rho > 0$.

To obtain guarantees for our smooth adversarial loss, we first give a multi-class adversarial $\mathcal{H}$-consistency bound for the adversarial sum loss $\widetilde{\Psi}_\rho^{\mathrm{sum}}(h, x, y) = \sup_{x' : \|x-x'\| \leq \gamma} \Psi_\rho^{\mathrm{sum}}(h, x', y)$ with symmetric and locally ρ-consistent hypothesis sets. Our multi-class $\mathcal{H}$-consistency bound is new and more significant than previous results given in the special case of binary classification (Awasthi et al., 2022a).

Theorem 1 **($\mathcal{H}$-consistency bound of $\widetilde{\Psi}_\rho^{\mathrm{sum}}$).** *Assume that $\mathcal{H}$ is symmetric and locally ρ-consistent. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D}$, the following inequality holds:*

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma, \mathcal{H}} \leq \mathcal{R}_{\widetilde{\Psi}_\rho^{\mathrm{sum}}}(h) - \mathcal{R}^*_{\widetilde{\Psi}_\rho^{\mathrm{sum}}, \mathcal{H}} + \mathcal{M}_{\widetilde{\Psi}_\rho^{\mathrm{sum}}, \mathcal{H}} - \mathcal{M}_{\ell_\gamma, \mathcal{H}}. \quad (6)$$

The proof is presented in Appendix B. The difficulty in the multi-class setting is that, because there are multiple scores, the conditional regret cannot be characterized explicitly, as in the binary classification, by using a tool such as $\mathcal{H}$-estimation error transformation in (Awasthi et al., 2022a). Instead, we use novel proof techniques that avoid directly characterizing the conditional regret. We upper bound the minimal conditional $\widetilde{\Psi}_\rho^{\mathrm{sum}}$-risk by carefully choosing a hypothesis in the class that shares the same ordering of the labels with the conditional probability. Then, we lower bound its conditional regret by applying the rearrangement inequality regarding the conditional probabilities.

As mentioned earlier, the condition of Theorem 1 is verified by a broad range of hypothesis sets commonly used in practice including the family of all measurable functions, that of linear models and that of neural networks with or without softmax operator for a suitable choice of ρ. Note that the values of ρ verifying the condition depend on the hypothesis set, e.g. ρ depends on B and Λ for $\mathcal{H}_{\mathrm{NN}}$ as shown in the previous example. Furthermore, our bound holds for a broad class of auxiliary functions Ψ_ρ, which generalizes the results of Awasthi et al. (2022b) on the ρ-margin loss for adversarial robustness.

The locally ρ-consistency condition is easier to verify for a smaller ρ. One the other hand, Ψ_ρ with a smaller ρ will be closer to the $0/1$ loss and thus typically harder to optimize. Therefore, the choice of the most suitable value of ρ is subject to a trade-off. In practice, the hyper-parameter ρ can be selected via cross-validation.

Assume that Ψ_ρ is μ-Lipschitz in Theorem 1. Then, using the inequality $\Phi^{\mathrm{sum}}_{\mathrm{smooth}} \geq \widetilde{\Psi}_\rho^{\mathrm{sum}}$ which holds for $\Phi \geq \Psi_\rho$ and $\nu \geq \mu$, we obtain the following guarantee for our proposed smooth adversarial loss under the same condition on hypothesis set $\mathcal{H}$.

Corollary 2 **(Guarantees for smooth adversarial sum losses).** *Assume that $\mathcal{H}$ is symmetric and locally ρ-consistent, and Ψ_ρ is μ-Lipschitz. Then, for any auxiliary function $\Phi \geq \Psi_\rho$ and hyper-parameter $\nu \geq \mu$, any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D}$, the following inequality* *holds:*

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} \leq \mathcal{R}_{\Phi^{\text{sum}}_{\text{smooth}}}(h) - \mathcal{R}^*_{\widetilde{\Psi}^{\text{sum}}_\rho,\mathcal{H}} + \mathcal{M}_{\widetilde{\Psi}^{\text{sum}}_\rho,\mathcal{H}} - \mathcal{M}_{\ell_\gamma,\mathcal{H}}. \quad (7)$$

This guarantee is based on the $\mathcal{H}$-consistency bound of the adversarial sum loss. Corollary 2 theoretically motivates our PSAL algorithm presented in Section 5, which is based on minimization of surrogate smooth adversarial loss error.

In practice, the minimizability gaps appearing in Theorem 1 and Corollary 2 are equal to zero, in particular, when the learning problem is *realizable* along with a natural condition on the surrogate loss (Awasthi et al., 2021a). Thus, Theorem 1 guarantees $\mathcal{H}$-consistency for all these common cases since the inequality can then be rewritten as

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} \leq \mathcal{R}_{\widetilde{\Psi}^{\text{sum}}_\rho}(h) - \mathcal{R}^*_{\widetilde{\Psi}^{\text{sum}}_\rho,\mathcal{H}},$$

which shows that minimizing the surrogate estimation error minimizes the adversarial loss estimation error. Corollary 2 implies a similar guarantee for smooth adversarial losses.

Definition 3 (realizability). *A learning problem in the adversarial scenario is* realizable *for a hypothesis $\mathcal{H}$ if there exists a best-in-class hypothesis $h^* \in \mathcal{H}$ such that $\mathcal{R}_{\ell_\gamma,\mathcal{H}}(h^*) = \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} = 0$.*

Under the realizability assumption, we have $\mathcal{M}_{\ell_\gamma,\mathcal{H}} \leq \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} = 0$. Note that, as shown in (Awasthi et al., 2021a), $\mathcal{H}$-consistency for all distributions should not be anticipated in the adversarial scenario. Even for linear models, they proved that there are no continuous surrogate losses that can be $\mathcal{H}$-consistent for all distributions. The realizability assumption can be verified empirically for real datasets and neural networks used in practice: it is not hard to reach 100% adversarial accuracy when training on the union of the training and test datasets. In contrast, TRADES does not benefit from $\mathcal{H}$-consistency guarantees even in the realizable case, as shown in Section 6.

5 ALGORITHM

In this section, we will present our PSAL algorithm, which benefits from strong guarantees, as shown in the previous sections. Given an auxiliary function Φ and a constant $\nu \geq 0$, we define the corresponding objective function $\mathcal{F}_\Phi$ as follows:

$$\mathcal{F}_\Phi(h) = \frac{1}{m}\sum_{i=1}^m \left[\sum_{y' \neq y_i} \Phi(\Delta_h(x_i, y_i, y')) + \nu \sup_{x':\|x_i - x'\| \leq \gamma} \left\|\overline{\Delta}_h(x', y_i) - \overline{\Delta}_h(x_i, y_i)\right\|_2\right], \quad (8)$$

where m is the sample size. Thus, for any $h \in \mathcal{H}$, the objective $\mathcal{F}_\Phi(h)$ can be expressed as follows in terms of the smooth adversarial loss corresponding to the sum loss (Table 1):

$$\mathcal{F}_\Phi(h) = \frac{1}{m}\sum_{i=1}^m \Phi^{\text{sum}}_{\text{smooth}}(h, x_i, y_i). \quad (9)$$

The $\mathcal{H}$-consistency guarantee of Corollary 2 suggests minimizing $\frac{1}{m}\sum_{i=1}^m \Phi^{\text{sum}}_{\text{smooth}}(h, x_i, y_i)$ for some auxiliary function $\Phi \geq \Psi_\rho$ with Lipschitz constant μ, and hyper-parameter $\nu \geq \mu$ plus a regularization term $\mathcal{R}(h)$, as suggested by standard generalization bounds. This gives rise to the following minimization problem:

$$\min_{h \in \mathcal{H}} \mathcal{F}_\Phi(h) + \tau \mathcal{R}(h), \quad (10)$$

for some regularization parameter $\tau > 0$, auxiliary function $\Phi \geq \Psi_\rho$ with Lipschitz constant μ, hyper-parameters $\rho > 0$ and $\nu \geq \mu$. One natural choice of Ψ_ρ is the ρ-margin loss Φ_ρ, which is $\frac{1}{\rho}$-Lipschitz, and natural convex and differentiable upper bounds for Φ_ρ are for example, the ρ-logistic loss used in logistic regression, defined by $\Phi_{\rho\text{-log}}(t) = \log_2\left(1 + e^{-\frac{t}{\rho}}\right)$ and the ρ-exponential loss used in AdaBoost (Freund and Schapire, 1997), defined by $\Phi_{\rho\text{-exp}}(t) = e^{-\frac{t}{\rho}}$, which are adopted in our experiments in Section 7. One can also use alternative auxiliary functions. The algorithm defined by (10), which we will call **PSAL** (Principled Smooth Adversarial Loss algorithm), benefits from the $\mathcal{H}$-consistency guarantee with respect to the adversarial $0/1$ loss.

Note that when Φ is a convex function of h, by the standard Lagrange method, (10) can be equivalently and more efficiently solved with the replacement of the ℓ_2 norm by its square in (8), since the regularization term can be moved to a constraint and then be squared. In Section 7, we employed the squared ℓ_2 norm for the experiments.

For the inner maximization problem appearing in $\mathcal{F}_\Phi(h)$, we approximately solve it by Projected Gradient-Descent (PGD) method, which is widely used in adversarial training (Madry et al., 2017; Zhang et al., 2019). For the regularization term $\mathcal{R}(h)$, we adopt the L_2 regularization, which is often referred to as weight decay.

6 ANALYSIS OF TRADES

In this section, we will show that there exists a learning problem in the realizable case such that the surrogate loss TRADES does not benefit from $\mathcal{H}$-consistency guarantees, while our smooth adversarial loss indeed does. We will also see that even in the non-realizable case, our smooth adversarial loss can be $\mathcal{H}$-consistent while TRADES remains not.

Let $\mathcal{X} = B^d_2(1) := \left\{x \in \mathbb{R}^d \mid \|x\|_2 \leq 1\right\}$, where $\|\cdot\|_2$ is the ℓ_2 norm. We consider an adversarial binary classification problem with a family of linear models $\mathcal{H}_{\text{lin}} = \{x \to w \cdot x \mid \|w\|_2 = 1\}$ under ℓ_2 perturbations. In this case, the adversarial $0/1$ loss $\ell_\gamma(h,x,y)$ can be expressed as:

$$\sup_{x':\|x-x'\|_2\le\gamma} \mathbb{1}_{yh(x')\le 0} = \mathbb{1}_{\inf_{x':\|x-x'\|_2\le\gamma} yw\cdot x'\le 0} = \mathbb{1}_{yw\cdot x\le\gamma}. \quad (11)$$

In binary classification, the TRADES loss is expressed as follows (Zhang et al., 2019, eq. (3)):

$$\widetilde{\ell}_{\text{trades}}(h,x,y) = \Phi_{\log}(yh(x)) + \sup_{x':\|x-x'\|\le\gamma} \Phi_{\log}(h(x)h(x')/\lambda), \quad (12)$$

where $\Phi_{\log} = \log(1+e^{-t})$ is the logistic loss, the binary counterpart of the cross-entropy loss used in the multi-class formulation (3). Note that in (3), the softmax operator is included in the hypothesis set, while in (12), the softmax operator is inherently included in the logistic loss function. Therefore, the composition of the binary formulation with $\mathcal{H}$ corresponds to that of the multi-class formulation with $\mathcal{H}^{\text{softmax}}$. Also, note that in their multi-class formulation (3) of TRADES, λ is outside the loss function while in their binary formulation (12), λ is inside the loss function. This is in fact one of the issues with the TRADES analysis that we will mention below: the authors only provide a theoretical analysis for the binary case, while the multi-class formulation is only given a heuristic extension. For comparison, the guarantees for our smooth adversarial loss (Corollary 2) apply to the multi-class setting, which is based on a new and more informative multi-class $\mathcal{H}$-consistency bound (Theorem 1). Nevertheless, the negative results for TRADES, Theorem 4 and 5, hold whether λ is inside or outside the loss function in (12), with basically the same proof.

By the definition in Section 3.2, using as an auxiliary function the ρ-margin loss Φ_ρ which is $\frac{1}{\rho}$-Lipschitz, our smooth adversarial loss in binary classification is expressed as:

$$\Phi_{\text{smooth}} = \Phi_\rho(yh(x)) + \frac{1}{\rho}\left[yh(x) - \inf_{x':\|x-x'\|\le\gamma} yh(x')\right]. \quad (13)$$

The next result shows that there exists a realizable learning problem for which $\widetilde{\ell}_{\text{trades}}$ does not admit the $\mathcal{H}$-consistency guarantee while Φ_{smooth} does.

Theorem 4 (**Negative results for TRADES: realizable case**). *There exists a learning problem that is realizable for $\mathcal{H}_{\text{lin}}$, such that $\widetilde{\ell}_{\text{trades}}$ with any $\lambda > 0$ is not $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ, while there exists $\rho > 0$ such that Φ_{smooth} with the auxiliary function Φ_ρ is $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ.*

The proof is presented in Appendix C. Figure 1 gives an illustration of that realizable example, where the best-in-class hypothesis for ℓ_γ coincides with that for Φ_{smooth} and achieves zero generalization error for ℓ_γ, but deviates far from that for $\widetilde{\ell}_{\text{trades}}$.

Theorem 4 rules out the $\mathcal{H}$-consistency for TRADES in the realizable case, let alone the non-realizable case where it is harder to achieve such guarantees. In contrast, our smooth adversarial loss can benefit from $\mathcal{H}$-consistency guarantees even in some non-realizable case, as shown in the below.

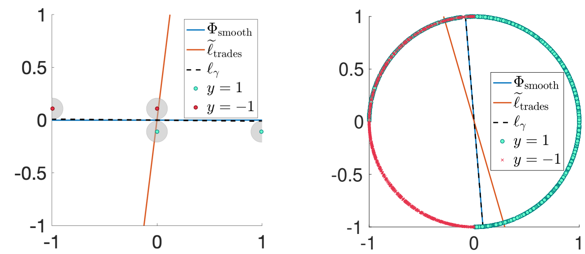

Figure 1: Left: example in the realizable case used in Theorem 4. Right: example in the non-realizable case used in Theorem 5. The best-in-class hypothesis for ℓ_γ coincides with that for Φ_{smooth}, but not for $\widetilde{\ell}_{\text{trades}}$ in both cases.

Theorem 5 (**Negative results for TRADES: non-realizable case**). *There exists a learning problem that is non-realizable for $\mathcal{H}_{\text{lin}}$, such that Φ_{smooth} with the auxiliary function Φ_ρ and a suitable $\rho > 0$ is $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ, while $\widetilde{\ell}_{\text{trades}}$ with any $\lambda > 0$ is not $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ.*

The proof is presented in Appendix D. Figure 1 also gives an illustration of that non-realizable example. Theorem 4 and 5 suggest that TRADES does not admit consistency guarantees for the adversarial $0/1$ loss while our smooth adversarial loss does. Zhang et al. (2019) showed that for any surrogate loss that is Bayes consistent with respect to the standard binary $0/1$ loss, the difference of the robust accuracy and the natural accuracy of the classifier obtained by optimizing the surrogate can be upper bounded by a term that captures the vulnerability of the surrogate near the boundary. However, this does not provide any theoretical guarantees with respect to the generalization error of the adversarial $0/1$ loss itself. Moreover, their guarantees (Zhang et al., 2019, Theorem 3.1) only apply to the binary case and the hypothesis set of all measurable functions. In contrast, our guarantee is based on a multi-class $\mathcal{H}$-consistency bound in the adversarial setting, which is directly relevant to the adversarial $0/1$ loss, applies to the multi-class setting, and holds for general hypothesis sets. Additionally, while TRADES is based on the trade-off between clean accuracy and robust accuracy, the experiments in Section 7 show empirically that our smooth adversarial losses can achieve a better performance for both robust and clean accuracy, with an even more significant improvement of the clean accuracy (Table 2).

Table 2: Clean accuracy and robust accuracy under $\mathrm{PGD}^{40}_{\mathrm{margin}}$ and AutoAttack, with reported mean and standard deviation over three runs in each setting for both PSAL and the state-of-the-art TRADES in (Gowal et al., 2020). Results of other well-known adversarial defense models are included for completeness. PSAL consistently outperforms TRADES for both robust and clean accuracy, with an even more significant improvement of the clean accuracy in all the cases.

Method	Dataset	Norm	Maximum magnitude	Clean	$\mathrm{PGD}^{40}_{\mathrm{margin}}$	AutoAttack
Gowal et al. (2020) (WRN-70-16)	CIFAR-10	ℓ_∞	$\gamma = 8/255$	85.34 ± 0.04%	57.90 ± 0.13%	57.05 ± 0.17%
PSAL (WRN-70-16)				**86.63** ± **0.24%**	**59.01** ± **0.13%**	**57.46** ± **0.12%**
Gowal et al. (2020) (WRN-34-20)				85.21 ± 0.16%	57.54 ± 0.18%	56.70 ± 0.14%
PSAL (WRN-34-20)				**86.71** ± **0.08%**	**58.68** ± **0.16%**	**57.13** ± **0.18%**
Gowal et al. (2020) (WRN-28-10)				84.33 ± 0.18%	55.92 ± 0.20%	55.19 ± 0.23%
PSAL (WRN-28-10)				**86.07** ± **0.14%**	**57.12** ± **0.19%**	**55.66** ± **0.16%**
Pang et al. (2020) (WRN-34-20)				86.43%	—	54.39%
Rice et al. (2020) (WRN-34-20)				85.34%	—	53.42%
Wu et al. (2020) (WRN-34-10)				85.36%	—	56.17%
Qin et al. (2019) (WRN-40-8)				86.28%	—	52.84%
Xu et al. (2022) (ResNet-32)				80.43%	—	44.15%
Gowal et al. (2020) (WRN-70-16)	CIFAR-100	ℓ_∞	$\gamma = 8/255$	60.56 ± 0.31%	31.39 ± 0.19%	29.93 ± 0.14%
PSAL (WRN-70-16)				**62.25** ± **0.26%**	**34.11** ± **0.17%**	**30.63** ± **0.10%**
Gowal et al. (2020) (WRN-34-20)	SVHN	ℓ_∞	$\gamma = 8/255$	93.03 ± 0.13%	61.01 ± 0.16%	57.84 ± 0.19%
PSAL (WRN-34-20)				**94.31** ± **0.17%**	**63.12** ± **0.14%**	**58.08** ± **0.15%**

7 EXPERIMENTS

In this section, we present experimental results on CIFAR-10 (Krizhevsky, 2009), CIFAR-100 (Krizhevsky, 2009) and SVHN (Netzer et al., 2011) datasets to demonstrate the effectiveness of our algorithm PSAL.

Experimental Settings We follow the settings of Gowal et al. (2020) and apply WideResNet (WRN) (Zagoruyko and Komodakis, 2016) with SiLU activations (Hendrycks and Gimpel, 2016; He et al., 2016), where WRN-n-k denotes a residual network that has a total number of convolutional layers n and a widening factor k (for example, network with 76 layers and $k = 16$ times wider than original would be denoted as WRN-70-16). In training, we use Stochastic Gradient Descent (SGD) with Nesterov momentum (Nesterov, 1983) with a batch size of 1,024 and weight decay 5×10^{-4}. The training runs for 800 epochs with the cosine decay learning rate schedule (Loshchilov and Hutter, 2016), using an initial learning rate of 0.4 for CIFAR-10 and SVHN, and an initial learning rate of 0.1 for CIFAR-100 without restarts. For CIFAR-10 and CIFAR-100, the commonly used data augmentations, 32×32 random crops after padding by 4 pixels and random horizontal flips, are applied. The training attacks are generated by a 10-step PGD adversary as mentioned in Section 5, with random starts. We adopt model weight averaging (Izmailov et al., 2018) with a decay rate of 0.9975. For our smooth adversarial losses, we set both ρ and ν to 1.0 for CIFAR-10 and SVHN, whereas we set $\rho = 0.3$ and $\nu = 6.0$ for CIFAR-100. For TRADES, we use the same setup as Gowal et al. (2020).

Evaluation We mitigate robust overfitting with early stopping (Rice et al., 2020). Throughout training, we measure the robust accuracy on a held-out validation set of 1,024 samples using 40-step PGD on the margin loss, denoted by $\mathrm{PGD}^{40}_{\mathrm{margin}}$, to select the best check-point. We report the mean and standard deviation over three runs in each setting for both PSAL and the state-of-the-art TRADES in (Gowal et al., 2020). We evaluate the robustness of the trained models via AutoAttack (Croce and Hein, 2020b),[1] a widely recognized benchmark with an ensemble of three white-box attacks, that are Auto-PGD (APGD) on the cross-entropy, APGD on the DLR-loss and FAB (Croce and Hein, 2020a), and one black-box attack, that is the Square Attack (Andriushchenko et al., 2020). We also report the clean accuracy and the robust accuracy measured by $\mathrm{PGD}^{40}_{\mathrm{margin}}$ on the full test sets. Here, the clean accuracy refers to the standard classification accuracy, as opposed to the adversarial accuracy. For SVHN, the accuracy is measured on 5,000 points randomly chosen from the test set. The results for TRADES reproduced by us match those reported in (Gowal et al., 2020).

Comparison with TRADES Gowal et al. (2020) achieved state-of-the-art results by adopting TRADES with a combination of early stopping, model weight averaging and a well-tuned hyperparameter configuration. On CIFAR-10, CIFAR-100 and SVHN, we consider ℓ_∞-norm bounded perturbations of size $\gamma = 8/255$. Table 2 shows that PSAL consistently outperforms TRADES for both robust and clean accuracy, with an even more significant improvement of the clean accuracy. Here, PSAL is implemented with

[1] https://github.com/fra31/auto-attack.

$\Phi = \Phi_{\rho-\log}$ for CIFAR-10 and SVHN, and $\Phi = \Phi_{\rho-\exp}$ for CIFAR-100. For a fair comparison, the same neural network architecture, WRN-70-16, WRN-34-20 or WRN-28-10, is adopted for both methods.

In Table 2, we include results of some other well-known adversarial defense models for completeness, among which Pang et al. (2020) studied a series of tricks for adversarial training, Rice et al. (2020) advocated the use of early stopping in adversarially robust deep learning, Wu et al. (2020) proposed Adversarial Weight Perturbation, Qin et al. (2019) introduced a regularizer to avoid gradient obfuscation (Athalye et al., 2018) through local linearization, and Xu et al. (2022) constructed a special type of dense orthogonal weights. Our algorithm PSAL with WRN-34-20 surpasses all of them.

It is worth pointing out that we are in the setting where no additional generated data or extra data is used. Let us also emphasize that, while improving the robust accuracy, we further achieve a very significant improvement in clean accuracy as compared to (Gowal et al., 2020).

8 CONCLUSION

We presented a series of theoretical, algorithmic, and empirical results for adversarial robustness. Our theoretical analysis, including our proofs of multi-class $\mathcal{H}$-consistency for sum losses, provides new tools for the analysis of other similar loss functions in adversarial multi-class classification. Our PSAL algorithms provide effective solutions for adversarial robustness in multiple tasks. Our extensive empirical analysis demonstrates their effectiveness of these algorithms and establishes the new state-of-the-art for multiple problems. The family of smooth losses we introduced can potentially be useful for the design of similar algorithms in other scenarios beyond adversarial robustness for classification. We hope that these results will provide new tools for the study of adversarial robustness, which remains a challenging question, in spite of the improvements reported.

Acknowledgements

We warmly thank Sven Gowal for the generous help on the experimental details.

References

M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pages 484–501, 2020.

A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283, 2018.

P. Awasthi, N. Frank, A. Mao, M. Mohri, and Y. Zhong. Calibration and consistency of adversarial surrogate losses. In *Advances in Neural Information Processing Systems*, pages 9804–9815, 2021a.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. A finer calibration analysis for adversarial robustness. *arXiv preprint arXiv:2105.01550*, 2021b.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. $\mathcal{H}$-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*, pages 1117–1174, 2022a.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. Multi-class $\mathcal{H}$-consistency bounds. In *Advances in Neural Information Processing Systems*, 2022b.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. DC-programming for neural network optimizations. *Journal of Global Optimization*, 2023.

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

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

K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001.

F. Croce and M. Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning*, pages 2196–2205, 2020a.

F. Croce and M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216, 2020b.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1): 119–139, 1997.

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

S. Gowal, C. Qin, J. Uesato, T. Mann, and P. Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

P. Izmailov, D. Podoprikhin, T. Garipov, D. P. Vetrov, and A. G. Wilson. Averaging weights leads to wider optima and better generalization. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 876–885, 2018.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Toronto University, 2009.

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

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

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

I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

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

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

Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Dokl. akad. nauk Sssr*, 269:543–547, 1983.

Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *Advances in Neural Information Processing Systems*, 2011.

T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu. Bag of tricks for adversarial training. *arXiv preprint arXiv:2010.00467*, 2020.

C. Qin, J. Martens, S. Gowal, D. Krishnan, K. Dvijotham, A. Fawzi, S. De, R. Stanforth, and P. Kohli. Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems*, 2019.

L. Rice, E. Wong, and J. Z. Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104, 2020.

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

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

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

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

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

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

J. Weston and C. Watkins. Multi-class support vector machines. Technical report, Citeseer, 1998.

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

D. Wu, S.-T. Xia, and Y. Wang. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems*, pages 2958–2969, 2020.

C. Xu, X. Li, and M. Yang. An orthogonal classifier for improving the adversarial robustness of neural networks. *Information Sciences*, 591:251–262, 2022.

S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

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

M. Zhang and S. Agarwal. 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, 2020.

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

Contents of Appendix

A DERIVATION OF SMOOTH ADVERSARIAL LOSSES

A.1 Max Loss

The adversarial max loss is defined by

$$\widetilde{\Phi}^{\max}(h,x,y) = \sup_{x' : \|x - x'\| \leq \gamma} \Phi^{\max}(h,x',y) = \sup_{x' : \|x - x'\| \leq \gamma} \Phi(\rho_h(x',y)).$$

If Φ is non-increasing and μ-Lipschitz, then the following decomposition and inequality hold for any $\nu \geq \mu$:

$$\begin{aligned}
\widetilde{\Phi}^{\max}(h,x,y) &= \Phi^{\max}(h,x,y) + \widetilde{\Phi}^{\max}(h,x,y) - \Phi^{\max}(h,x,y)\\
&= \Phi^{\max}(h,x,y) + \sup_{x' : \|x - x'\| \leq \gamma} \Phi(\rho_h(x',y)) - \Phi(\rho_h(x,y))\\
&= \Phi^{\max}(h,x,y) + \Phi\left(\inf_{x' : \|x - x'\| \leq \gamma} \rho_h(x',y)\right) - \Phi(\rho_h(x,y)) && (\Phi \text{ is non-increasing})\\
&\leq \Phi^{\max}(h,x,y) + \nu \left|\rho_h(x,y) - \inf_{x' : \|x - x'\| \leq \gamma} \rho_h(x',y)\right| && (\Phi\ \mu\text{-Lipschitz and } \nu \geq \mu)\\
&= \Phi^{\max}_{\text{smooth}}
\end{aligned}$$

A.2 Sum Loss

The adversarial max loss is defined by

$$\widetilde{\Phi}^{\text{sum}}(h,x,y) = \sup_{x' : \|x - x'\| \leq \gamma} \Phi^{\text{sum}}(h,x',y) = \sup_{x' : \|x - x'\| \leq \gamma} \sum_{y' \neq y} \Phi(h(x',y) - h(x',y')).$$

Let $\Delta_h(x,y,y') = h(x,y) - h(x,y')$ and $\overline{\Delta}_h(x,y)$ denote the $c-1$ dimensional vector $\big(\Delta_h(x,y,1), \ldots, \Delta_h(x,y,y-1), \Delta_h(x,y,y+1), \ldots, \Delta_h(x,y,c)\big)$. If Φ is non-increasing and μ-Lipschitz, then the following decomposition and inequality hold for any $\nu \geq \sqrt{c-1}\mu$:

$$\begin{aligned}
\widetilde{\Phi}^{\text{sum}}(h,x,y) &= \Phi^{\text{sum}}(h,x,y) + \widetilde{\Phi}^{\text{sum}}(h,x,y) - \Phi^{\text{sum}}(h,x,y)\\
&= \Phi^{\text{sum}}(h,x,y) + \sup_{x' : \|x - x'\| \leq \gamma} \sum_{y' \neq y} \Phi(\Delta_h(x',y,y')) - \sum_{y' \neq y} \Phi(\Delta_h(x,y,y'))\\
&= \Phi^{\text{sum}}(h,x,y) + \sup_{x' : \|x - x'\| \leq \gamma} \sum_{y' \neq y} \left(\Phi(\Delta_h(x',y,y')) - \Phi(\Delta_h(x,y,y'))\right)\\
&\leq \Phi^{\text{sum}}(h,x,y) + \mu \sup_{x' : \|x - x'\| \leq \gamma} \left\|\overline{\Delta}_h(x',y) - \overline{\Delta}_h(x,y)\right\|_1 && (\Phi\ \mu\text{-Lipschitz})\\
&\leq \Phi^{\text{sum}}(h,x,y) + \mu\sqrt{c-1} \sup_{x' : \|x - x'\| \leq \gamma} \left\|\overline{\Delta}_h(x',y) - \overline{\Delta}_h(x,y)\right\|_2 && (\text{Cauchy-Schwarz ineq.})\\
&\leq \Phi^{\text{sum}}(h,x,y) + \nu \sup_{x' : \|x - x'\| \leq \gamma} \left\|\overline{\Delta}_h(x',y) - \overline{\Delta}_h(x,y)\right\|_2 && (\nu \geq \mu\sqrt{c-1})\\
&= \Phi^{\text{sum}}_{\text{smooth}}.
\end{aligned}$$

A.3 Constrained Loss

The adversarial constrained loss $\widetilde{\Phi}^{\text{cstnd}}$ is defined by

$$\widetilde{\Phi}^{\text{cstnd}}(h,x,y) = \sup_{x' : \|x - x'\| \leq \gamma} \Phi^{\text{cstnd}}(h,x',y) = \sup_{x' : \|x - x'\| \leq \gamma} \sum_{y' \neq y} \Phi(-h(x',y')).$$

with the constraint that $\sum_{y \in \mathcal{Y}} h(x,y) = 0$. Let $\overline{h}(x,y)$ denote the $(c-1)$-dimensional vector $\big(h(x,1), \ldots, h(x,y-1), h(x,y+1), \ldots, h(x,c)\big)$. If Φ is non-increasing and μ-Lipschitz, then, the following decomposition and inequality hold for any $\nu \geq \sqrt{c-1}\mu$:

$$\begin{aligned}
\widetilde{\Phi}^{\mathrm{cstnd}}(h,x,y) &= \Phi^{\mathrm{cstnd}}(h,x,y) + \widetilde{\Phi}^{\mathrm{cstnd}}(h,x,y) - \Phi^{\mathrm{cstnd}}(h,x,y)\\
&= \Phi^{\mathrm{cstnd}}(h,x,y) + \sup_{x':\|x-x'\|\leq\gamma} \sum_{y'\neq y} \Phi(-h(x',y')) - \sum_{y'\neq y} \Phi(-h(x,y'))\\
&= \Phi^{\mathrm{cstnd}}(h,x,y) + \sup_{x':\|x-x'\|\leq\gamma} \sum_{y'\neq y} \left(\Phi(-h(x',y')) - \Phi(-h(x,y'))\right)\\
&\leq \Phi^{\mathrm{cstnd}}(h,x,y) + \mu \sup_{x':\|x-x'\|\leq\gamma} \left\|\overline{h}(x',y) - \overline{h}(x,y)\right\|_1 && (\Phi\ \mu\text{-Lipschitz})\\
&\leq \Phi^{\mathrm{cstnd}}(h,x,y) + \mu\sqrt{c-1} \sup_{x':\|x-x'\|\leq\gamma} \left\|\overline{h}(x',y) - \overline{h}(x,y)\right\|_2 && (\text{Cauchy-Schwarz ineq.})\\
&\leq \Phi^{\mathrm{cstnd}}(h,x,y) + \nu \sup_{x':\|x-x'\|\leq\gamma} \left\|\overline{h}(x',y) - \overline{h}(x,y)\right\|_2 && (\nu \geq \mu\sqrt{c-1})\\
&= \Phi^{\mathrm{cstnd}}_{\mathrm{smooth}}.
\end{aligned}$$

B PROOF OF THEOREM 1

Theorem 1 **($\mathcal{H}$-consistency bound of $\widetilde{\Psi}^{\mathrm{sum}}_{\rho}$).** *Assume that $\mathcal{H}$ is symmetric and locally ρ-consistent. Then, for any hypothesis $h \in \mathcal{H}$ and any distribution $\mathcal{D}$, the following inequality holds:*

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} \leq \mathcal{R}_{\widetilde{\Psi}^{\mathrm{sum}}_{\rho}}(h) - \mathcal{R}^*_{\widetilde{\Psi}^{\mathrm{sum}}_{\rho},\mathcal{H}} + \mathcal{M}_{\widetilde{\Psi}^{\mathrm{sum}}_{\rho},\mathcal{H}} - \mathcal{M}_{\ell_\gamma,\mathcal{H}}. \tag{6}$$

Proof. Let $\overline{\mathcal{H}}_\gamma(x) = \left\{h \in \mathcal{H} : \inf_{x':\|x-x'\|\leq\gamma} \rho_h(x', \mathsf{h}(x)) > 0\right\}$ and $p(x) = (p(x,1),\ldots,p(x,c))$. For any $x \in \mathcal{X}$ and $h \in \mathcal{H}$, we define $h\Big(x,\{1\}^h_x\Big), h\Big(x,\{2\}^h_x\Big),\ldots,h\Big(x,\{c\}^h_x\Big)$ by sorting the scores $\{h(x,y) : y \in \mathcal{Y}\}$ in increasing order, and $p_{[1]}(x), p_{[2]}(x),\ldots,p_{[c]}(x)$ by sorting the probabilities $\{p(x,y) : y \in \mathcal{Y}\}$ in increasing order. Note $\{c\}^h_x = \mathsf{h}(x)$. Since $\mathcal{H}$ is symmetric and locally ρ-consistent, for any $x \in \mathcal{X}$, there exists a hypothesis $h^* \in \mathcal{H}$ such that

$$\begin{aligned}
&\inf_{x':\|x-x'\|\leq\gamma} |h^*(x',i) - h^*(x',j)| \geq \rho, \forall i \neq j \in \mathcal{Y}\\
&p(x,\{k\}^{h^*}_{x'}) = p_{[k]}(x), \forall x' \in \{x' : \|x-x'\| \leq \gamma\}, \forall k \in \mathcal{Y}.
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\mathcal{C}^*_{\widetilde{\Psi}^{\mathrm{sum}}_{\rho},\mathcal{H}}(x)\\
&\leq \mathcal{C}_{\widetilde{\Psi}^{\mathrm{sum}}_{\rho}}(h^*,x)\\
&= \sum_{y\in\mathcal{Y}} \sup_{x':\|x-x'\|\leq\gamma} p(x,y) \sum_{y'\neq y} \Psi_\rho(h^*(x',y) - h^*(x',y'))\\
&= \sum_{i=1}^{c} \sup_{x':\|x-x'\|\leq\gamma} p(x,\{i\}^{h^*}_{x'}) \left[\sum_{j=1}^{i-1} \Psi_\rho\Big(h^*(x',\{i\}^{h^*}_{x'}) - h^*(x',\{j\}^{h^*}_{x'})\Big) + \sum_{j=i+1}^{c} \Psi_\rho\Big(h^*(x',\{i\}^{h^*}_{x'}) - h^*(x',\{j\}^{h^*}_{x'})\Big)\right]\\
&= \sum_{i=1}^{c} \sup_{x':\|x-x'\|\leq\gamma} p(x,\{i\}^{h^*}_{x'}) \left[\sum_{j=1}^{i-1} \Psi_\rho\Big(h^*(x',\{i\}^{h^*}_{x'}) - h^*(x',\{j\}^{h^*}_{x'})\Big) + c - i\right] && (\Psi_\rho(t) = 1,\ \forall t \leq 0)\\
&= \sum_{i=1}^{c} \sup_{x':\|x-x'\|\leq\gamma} p(x,\{i\}^{h^*}_{x'})(c-i) && (\inf_{x':\|x-x'\|\leq\gamma}|h^*(x',i) - h^*(x',j)| \geq \rho \text{ for any } i \neq j \text{ and } \Psi_\rho(t) = 0,\ \forall t \geq \rho)\\
&= \sum_{i=1}^{c} p_{[i]}(x)(c-i) && (p(x,\{k\}^{h^*}_{x'}) = p_{[k]}(x), \forall x' \in \{x' : \|x-x'\| \leq \gamma\}, \forall k \in \mathcal{Y})\\
&= c - \sum_{i=1}^{c} i\, p_{[i]}(x) && \left(\textstyle\sum_{i=1}^{c} p_{[i]}(x) = 1\right)
\end{aligned}$$

Note $\overline{\mathcal{H}}_\gamma(x) \neq \emptyset$ under the assumption. Then, use the derivation above, we obtain

$$\begin{aligned}
&\Delta\mathcal{C}_{\widetilde{\Psi}^{\mathrm{sum}}_\rho,\mathcal{H}}(h,x)\\
&= \sum_{i=1}^{c} \sup_{x':\|x-x'\|\leq\gamma} p(x,\{i\}^h_{x'})\left[\sum_{j=1}^{i-1}\Psi_\rho\Big(h(x',\{i\}^h_{x'}) - h(x',\{j\}^h_{x'})\Big) + c - i\right] - \left(c - \sum_{i=1}^{c} i\, p_{[i]}(x)\right)\\
&\geq p(x,\mathsf{h}(x))\mathbb{1}_{h\notin\overline{\mathcal{H}}_\gamma(x)} + \sum_{i=1}^{c}\sup_{x':\|x-x'\|\leq\gamma} p(x,\{i\}^h_{x'})(c-i) - \left(c - \sum_{i=1}^{c} i\, p_{[i]}(x)\right) && (\Psi_\rho \text{ is non-negative})\\
&\geq p(x,\mathsf{h}(x))\mathbb{1}_{h\notin\overline{\mathcal{H}}_\gamma(x)} + \sum_{i=1}^{c} i\, p_{[i]}(x) - \sum_{i=1}^{c} i\, p(x,\{i\}^h_x) && (\textstyle\sup_{x':\|x-x'\|\leq\gamma} p(x,\{i\}^h_{x'}) \geq p(x,\{i\}^h_x))\\
&= p(x,\mathsf{h}(x))\mathbb{1}_{h\notin\overline{\mathcal{H}}_\gamma(x)} + \max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x)) + \begin{bmatrix} c-1\\ c-1\\ c-2\\ \vdots\\ 1\end{bmatrix}\cdot\begin{bmatrix} p_{[c]}(x)\\ p_{[c-1]}(x)\\ p_{[c-2]}(x)\\ \vdots\\ p_{[1]}(x)\end{bmatrix} - \begin{bmatrix} c-1\\ c-1\\ c-2\\ \vdots\\ 1\end{bmatrix}\cdot\begin{bmatrix} p(x,\{c\}^h_x)\\ p(x,\{c-1\}^h_x)\\ p(x,\{c-2\}^h_x)\\ \vdots\\ p(x,\{1\}^h_x)\end{bmatrix}\\
& && (p_{[c]}(x) = \textstyle\max_{y\in\mathcal{Y}} p(x,y) \text{ and } \{c\}^h_x = \mathsf{h}(x))\\
&\geq p(x,\mathsf{h}(x))\mathbb{1}_{h\notin\overline{\mathcal{H}}_\gamma(x)} + \max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x))\\
& && (\text{ rearrangement inequality for } c-1\geq c-1\geq c-2\geq\cdots\geq 1 \text{ and } p_{[c]}(x)\geq\cdots\geq p_{[1]}(x))\\
&= \max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x))\mathbb{1}_{h\in\overline{\mathcal{H}}_\gamma(x)}
\end{aligned}$$

for any $h \in \mathcal{H}$. Since $\mathcal{H}$ is symmetric and $\overline{\mathcal{H}}_\gamma(x) \neq \emptyset$, we have

$$\begin{aligned}
\Delta\mathcal{C}_{\ell_\gamma,\mathcal{H}}(h,x) &= \mathcal{C}_{\ell_\gamma}(h,x) - \mathcal{C}^*_{\ell_\gamma,\mathcal{H}}(x)\\
&= \sum_{y\in\mathcal{Y}} p(x,y) \sup_{x':\|x-x'\|\leq\gamma}\mathbb{1}_{\rho_h(x',y)\leq 0} - \inf_{h\in\mathcal{H}}\sum_{y\in\mathcal{Y}} p(x,y)\sup_{x':\|x-x'\|\leq\gamma}\mathbb{1}_{\rho_h(x',y)\leq 0}\\
&= (1-p(x,\mathsf{h}(x)))\mathbb{1}_{h\in\overline{\mathcal{H}}_\gamma(x)} + \mathbb{1}_{h\notin\overline{\mathcal{H}}_\gamma(x)} - \inf_{h\in\mathcal{H}}\Big[(1-p(x,\mathsf{h}(x)))\mathbb{1}_{h\in\overline{\mathcal{H}}_\gamma(x)} + \mathbb{1}_{h\notin\overline{\mathcal{H}}_\gamma(x)}\Big]\\
&= (1-p(x,\mathsf{h}(x)))\mathbb{1}_{h\in\overline{\mathcal{H}}_\gamma(x)} + \mathbb{1}_{h\notin\overline{\mathcal{H}}_\gamma(x)} - \left(1 - \max_{y\in\mathcal{Y}} p(x,y)\right) \qquad (\mathcal{H} \text{ is symmetric and } \overline{\mathcal{H}}_\gamma(x)\neq\emptyset)\\
&= \max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x))\mathbb{1}_{h\in\overline{\mathcal{H}}_\gamma(x)}.
\end{aligned}$$

Therefore, by the definition, we obtain

$$\begin{aligned}
\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} + \mathcal{M}_{\ell_\gamma,\mathcal{H}} &= \mathbb{E}_X\big[\Delta\mathcal{C}_{\ell_\gamma}(h,x)\big]\\
&= \mathbb{E}_X\left[\max_{y\in\mathcal{Y}} p(x,y) - p(x,\mathsf{h}(x))\mathbb{1}_{h\in\overline{\mathcal{H}}_\gamma(x)}\right]\\
&\leq \mathbb{E}_X\left[\Delta\mathcal{C}_{\widetilde{\Psi}^{\mathrm{sum}}_\rho,\mathcal{H}}(h,x)\right]\\
&= \mathcal{R}_{\widetilde{\Psi}^{\mathrm{sum}}_\rho}(h) - \mathcal{R}^*_{\widetilde{\Psi}^{\mathrm{sum}}_\rho,\mathcal{H}} + \mathcal{M}_{\widetilde{\Psi}^{\mathrm{sum}}_\rho,\mathcal{H}},
\end{aligned}$$

which implies that

$$\mathcal{R}_{\ell_\gamma}(h) - \mathcal{R}^*_{\ell_\gamma,\mathcal{H}} \leq \mathcal{R}_{\widetilde{\Psi}^{\mathrm{sum}}_\rho}(h) - \mathcal{R}^*_{\widetilde{\Psi}^{\mathrm{sum}}_\rho,\mathcal{H}} + \mathcal{M}_{\widetilde{\Psi}^{\mathrm{sum}}_\rho,\mathcal{H}} - \mathcal{M}_{\ell_\gamma,\mathcal{H}}.$$

□

C PROOF OF THEOREM 4

Theorem 4 (Negative results for TRADES: realizable case). *There exists a learning problem that is realizable for $\mathcal{H}_{\mathrm{lin}}$, such that $\widetilde{\ell}_{\mathrm{trades}}$ with any $\lambda > 0$ is not $\mathcal{H}_{\mathrm{lin}}$-consistent with respect to ℓ_γ, while there exists $\rho > 0$ such that Φ_{smooth} with the auxiliary function Φ_ρ is $\mathcal{H}_{\mathrm{lin}}$-consistent with respect to ℓ_γ.*

Proof. We consider the following distribution. Let $x = (x_1, x_2)$, $x_1^2 + x_2^2 \le 1$ follow the distribution concentrated on the four points $(0, -\hat{\gamma})$, $(0, \hat{\gamma})$, $(I_{\hat{\gamma}}, -\hat{\gamma})$ and $(-I_{\hat{\gamma}}, \hat{\gamma})$ with the marginal distribution $\mathbb{P}[x = (0, -\hat{\gamma})] = \mathbb{P}[x = (0, \hat{\gamma})] = \frac{1-\beta}{2}$, $\mathbb{P}[x = (I_{\hat{\gamma}}, -\hat{\gamma})] = \mathbb{P}[x = (-I_{\hat{\gamma}}, \hat{\gamma})] = \frac{\beta}{2}$ and the conditional distribution

$$p(x, +1) = \begin{cases} 1, & x_2 < 0 \\ 0, & x_2 > 0 \end{cases} \qquad p(x, -1) = 1 - p(x, +1)$$

where $\beta \in (0, 1)$, $\hat{\gamma} = \gamma + \frac{1-\gamma}{100} = \frac{1+99\gamma}{100}$ and $I_{\hat{\gamma}} = \sqrt{1 - \hat{\gamma}^2}$. Let $\gamma = 0.1$ and $w = (\cos t, \sin t)$, $t \in [-\frac{\pi}{2}, \frac{3\pi}{2})$. By (11), for any $h \in \mathcal{H}_{\mathrm{lin}}$, the generalization error of ℓ_γ can be expressed as

$$\mathcal{R}_{\ell_\gamma}(h) = (1-\beta)\mathbb{1}_{-\hat{\gamma}\sin t \le \gamma} + \beta \mathbb{1}_{I_{\hat{\gamma}}\cos t - \hat{\gamma}\sin t \le \gamma}.$$

Therefore, the best-in-class hypotheses for adversarial 0/1 loss ℓ_γ are $w^*_{\ell_\gamma, \mathcal{H}_{\mathrm{lin}}} = \left(\cos t^*_{\ell_\gamma}, \sin t^*_{\ell_\gamma}\right)$, where $t^*_{\ell_\gamma} \in \left[-\frac{\pi}{2}, -\arcsin\frac{\gamma}{\hat{\gamma}}\right)$ and the best-in-class error for ℓ_γ is $\mathcal{R}^*_{\ell_\gamma, \mathcal{H}_{\mathrm{lin}}} = 0$.

For the linear hypothesis set $\mathcal{H}_{\mathrm{lin}}$, $\widetilde{\ell}_{\mathrm{trades}}$ can be written as

$$\begin{aligned} \widetilde{\ell}_{\mathrm{trades}}(h, x, y) &= \Phi_{\log}(yh(x)) + \sup_{x' : \|x - x'\| \le \gamma} \Phi_{\log}(h(x)h(x')/\lambda) \\ &= \Phi_{\log}(yw \cdot x) + \Phi_{\log}\left(\inf_{x' : \|x - x'\| \le \gamma} (w \cdot x)(w \cdot x')/\lambda\right) \\ &= \Phi_{\log}(yw \cdot x) + \Phi_{\log}\left(\left(|w \cdot x|^2 - \gamma |w \cdot x|\right)/\lambda\right). \end{aligned} \tag{14}$$

Thus, for any $h \in \mathcal{H}_{\mathrm{lin}}$, the generalization error of $\widetilde{\ell}_{\mathrm{trades}}$ can be expressed as

$$\begin{aligned} \mathcal{R}_{\widetilde{\ell}_{\mathrm{trades}}}(h) &= (1-\beta)\left[\Phi_{\log}(-\hat{\gamma}\sin t) + \Phi_{\log}\left(\left(|\hat{\gamma}\sin t|^2 - \gamma|\hat{\gamma}\sin t|\right)/\lambda\right)\right] \\ &\quad + \beta\left[\Phi_{\log}(I_{\hat{\gamma}}\cos t - \hat{\gamma}\sin t) + \Phi_{\log}\left(\left(|I_{\hat{\gamma}}\cos t - \hat{\gamma}\sin t|^2 - \gamma|I_{\hat{\gamma}}\cos t - \hat{\gamma}\sin t|\right)/\lambda\right)\right]. \end{aligned}$$

Therefore, as $\beta \to 1$, the best-in-class hypothesis for $\widetilde{\ell}_{\mathrm{trades}}$ tends to be $w^*_{\widetilde{\ell}_{\mathrm{trades}}, \mathcal{H}_{\mathrm{lin}}} = \left(\cos t^*_{\widetilde{\ell}_{\mathrm{trades}}}, \sin t^*_{\widetilde{\ell}_{\mathrm{trades}}}\right)$ with $t^*_{\widetilde{\ell}_{\mathrm{trades}}} = -\arcsin\hat{\gamma} \notin \left[-\frac{\pi}{2}, -\arcsin\frac{\gamma}{\hat{\gamma}}\right)$ since $\hat{\gamma}^2 < \gamma$. Therefore, $\widetilde{\ell}_{\mathrm{trades}}$ with any $\lambda > 0$ is not $\mathcal{H}_{\mathrm{lin}}$-consistent with respect to ℓ_γ.

On the other hand, for the linear hypothesis set $\mathcal{H}_{\mathrm{lin}}$, Φ_{smooth} can be written as

$$\begin{aligned} \Phi_{\mathrm{smooth}} &= \Phi_\rho(yh(x)) + \frac{1}{\rho}\left(yh(x) - \inf_{x' : \|x - x'\| \le \gamma} yh(x')\right) \\ &= \Phi_\rho(yw \cdot x) + \frac{1}{\rho}\left(yw \cdot x - \inf_{x' : \|x - x'\| \le \gamma} yw \cdot x'\right) \\ &= \Phi_\rho(yw \cdot x) + \frac{\gamma}{\rho}. \end{aligned} \tag{15}$$

Then, the generalization error of Φ_{smooth} can be expressed as

$$\mathcal{R}_{\Phi_{\mathrm{smooth}}}(h) = (1-\beta)\Phi_\rho(-\hat{\gamma}\sin t) + \beta\Phi_\rho(I_{\hat{\gamma}}\cos t - \hat{\gamma}\sin t) + \frac{\gamma}{\rho}.$$

Let $\rho = \hat{\gamma}$. Thus, the unique best-in-class hypothesis for Φ_{smooth} is $w^*_{\Phi_{\mathrm{smooth}}, \mathcal{H}_{\mathrm{lin}}} = \left(\cos t^*_{\Phi_{\mathrm{smooth}}}, \sin t^*_{\Phi_{\mathrm{smooth}}}\right)$, where $t^*_{\Phi_{\mathrm{smooth}}} = -\frac{\pi}{2} \in \left[-\frac{\pi}{2}, -\arcsin\frac{\gamma}{\hat{\gamma}}\right)$. Therefore, Φ_{smooth} with $\rho = \hat{\gamma}$ is $\mathcal{H}_{\mathrm{lin}}$-consistent with respect to ℓ_γ on this distribution.

□

D PROOF OF THEOREM 5

Theorem 5 (**Negative results for TRADES: non-realizable case**). *There exists a learning problem that is non-realizable for $\mathcal{H}_{\text{lin}}$, such that Φ_{smooth} with the auxiliary function Φ_ρ and a suitable $\rho > 0$ is $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ, while $\widetilde{\ell}_{\text{trades}}$ with any $\lambda > 0$ is not $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ.*

Proof. We consider the following distribution. Let $x = (\cos\theta, \sin\theta)$, $\theta \in [0, 2\pi)$ follow the uniform distribution on the unit circle with the conditional distribution defined by

$$p(x, +1) = \begin{cases} \frac{1}{2}, & \theta \in [\frac{\pi}{2}, \pi) \\ 1, & \theta \in [0, \frac{\pi}{2}) \text{ or } [\frac{3\pi}{2}, 2\pi) \\ 0, & \theta \in [\pi, \frac{3\pi}{2}) \end{cases} \qquad p(x, -1) = 1 - p(x, +1).$$

Let $\gamma = \cos\beta = 0.1$, that is $\beta = \arccos(\gamma) = \arccos(0.1) \in \left(\frac{\pi}{4}, \frac{\pi}{2}\right)$ and $w = (\cos t, \sin t)$, $t \in [0, 2\pi)$. Note that we have $w \cdot x = \cos(\theta - t)$. By (11), for any $h \in \mathcal{H}_{\text{lin}}$, the generalization error of ℓ_γ can be expressed as

$$\begin{aligned}
&\mathcal{R}_{\ell_\gamma}(h) \\
&= \frac{1}{2\pi}\left(\int_{\frac{\pi}{2}}^{\pi} \frac{1}{2}\mathbb{1}_{\cos(\theta-t)\le\cos\beta} + \frac{1}{2}\mathbb{1}_{-\cos(\theta-t)\le\cos\beta}\, d\theta + \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \mathbb{1}_{\cos(\theta-t)\le\cos\beta}\, d\theta + \int_{-\pi}^{-\frac{\pi}{2}} \mathbb{1}_{-\cos(\theta-t)\le\cos\beta}\, d\theta\right) \\
&= \frac{1}{2\pi}\left(\int_{\frac{\pi}{2}}^{\pi} \frac{1}{2}\mathbb{1}_{\cos(\theta-t)\le\cos\beta}\, d\theta + \int_{-\frac{\pi}{2}}^{0} \frac{1}{2}\mathbb{1}_{\cos(\theta-t)\le\cos\beta}\, d\theta + \int_{-\frac{\pi}{2}}^{0} \mathbb{1}_{\cos(\theta-t)\le\cos\beta}\, d\theta + \int_{0}^{\frac{\pi}{2}} 2\mathbb{1}_{\cos(\theta-t)\le\cos\beta}\, d\theta\right) \\
&\qquad \text{(change of variables)} \\
&= \frac{1}{2\pi}\left(\int_{\frac{\pi}{2}}^{\pi} \frac{1}{2}\mathbb{1}_{\cos(\theta-t)\le\cos\beta}\, d\theta + \int_{-\frac{\pi}{2}}^{0} \frac{3}{2}\mathbb{1}_{\cos(\theta-t)\le\cos\beta}\, d\theta + \int_{0}^{\frac{\pi}{2}} \left(\frac{1}{2} + 2 - \frac{1}{2}\right)\mathbb{1}_{\cos(\theta-t)\le\cos\beta}\, d\theta\right) \\
&= \frac{1}{2\pi}\left(\int_{0}^{\pi} \frac{1}{2}\mathbb{1}_{\cos(\theta-t)\le\cos\beta}\, d\theta + \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{3}{2}\mathbb{1}_{\cos(\theta-t)\le\cos\beta}\, d\theta\right) \\
&= \frac{1}{2\pi}\left(\int_{0}^{\pi} \frac{1}{2}\mathbb{1}_{\cos(\theta-t)\le\cos\beta}\, d\theta + \int_{0}^{\pi} \frac{3}{2}\mathbb{1}_{\sin(\theta-t)\le\cos\beta}\, d\theta\right) \qquad \text{(change of variables)} \\
&= \frac{1}{2\pi}\int_{-t}^{\pi-t} \frac{1}{2}\mathbb{1}_{\cos\theta\le\cos\beta} + \frac{3}{2}\mathbb{1}_{\sin\theta\le\cos\beta}\, d\theta. \qquad \text{(change of variables)}
\end{aligned}$$

Next, we analyze eight cases:

- When $-t \in \left[-\beta, \beta - \frac{\pi}{2}\right]$,

$$\begin{aligned}
\mathcal{R}_{\ell_\gamma}(h) &= \frac{1}{2\pi}\left[0 \times \left(2\beta - \frac{\pi}{2}\right) + \frac{1}{2} \times (\pi - \beta - t) + \frac{3}{2} \times \left(\frac{\pi}{2} - \beta + t\right) + 2 \times 0\right] \\
&\ge \frac{7}{8} - \frac{3\beta}{2\pi},
\end{aligned}$$

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

- When $-t \in \left[\beta - \frac{\pi}{2}, \frac{\pi}{2} - \beta\right]$,

$$\begin{aligned}
\mathcal{R}_{\ell_\gamma}(h) &= \frac{1}{2\pi}\left[0 \times \left(2\beta - \frac{\pi}{2}\right) + \frac{1}{2} \times \left(\frac{\pi}{2}\right) + \frac{3}{2} \times \left(\frac{\pi}{2} - \beta + t\right) + 2 \times \left(\frac{\pi}{2} - \beta - t\right)\right] \\
&\ge \frac{7}{8} - \frac{3\beta}{2\pi},
\end{aligned}$$

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

- When $-t \in \left[\frac{\pi}{2} - \beta, \beta\right]$,

$$\begin{aligned}
\mathcal{R}_{\ell_\gamma}(h) &= \frac{1}{2\pi}\left[0 \times (\beta + t) + \frac{1}{2} \times \left(\frac{\pi}{2}\right) + \frac{3}{2} \times 0 + 2 \times \left(\frac{\pi}{2} - \beta - t\right)\right] \\
&\ge \frac{9}{8} - \frac{2\beta}{\pi},
\end{aligned}$$

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

- When $-t \in [\beta, \pi - \beta]$,

$$\begin{aligned}\mathcal{R}_{\ell_\gamma}(h) &= \frac{1}{2\pi}\left[0 \times 0 + \frac{1}{2} \times \left(\beta + \frac{\pi}{2} + t\right) + \frac{3}{2} \times 0 + 2 \times \left(\frac{\pi}{2} - \beta - t\right)\right]\\ &\geq \frac{5}{8},\end{aligned}$$

  where the equality is achieved when $t = -\beta$.

- When $-t \in \left[\pi - \beta, \beta + \frac{\pi}{2}\right]$,

$$\begin{aligned}\mathcal{R}_{\ell_\gamma}(h) &= \frac{1}{2\pi}\left[0 \times 0 + \frac{1}{2} \times \left(\beta + \frac{\pi}{2} + t\right) + \frac{3}{2} \times (\beta - \pi - t) + 2 \times \left(\frac{3\pi}{2} - 2\beta\right)\right]\\ &\geq \frac{9}{8} - \frac{\beta}{2\pi},\end{aligned}$$

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

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

$$\begin{aligned}\mathcal{R}_{\ell_\gamma}(h) &= \frac{1}{2\pi}\left[0 \times 0 + \frac{1}{2} \times 0 + \frac{3}{2} \times (\beta - \pi - t) + 2 \times (2\pi - \beta + t)\right]\\ &\geq \frac{7}{8},\end{aligned}$$

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

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

$$\begin{aligned}\mathcal{R}_{\ell_\gamma}(h) &= \frac{1}{2\pi}\left[0 \times \left(-\frac{3\pi}{2} + \beta - t\right) + \frac{1}{2} \times 0 + \frac{3}{2} \times \left(\frac{\pi}{2}\right) + 2 \times (2\pi - \beta + t)\right]\\ &\geq \frac{11}{8} - \frac{2\beta}{\pi},\end{aligned}$$

  where the equality is achieved when $t = -\beta - \pi$.

- When $-t \in [\beta - \pi, -\beta]$,

$$\begin{aligned}\mathcal{R}_{\ell_\gamma}(h) &= \frac{1}{2\pi}\left[0 \times \left(-\frac{\pi}{2} + 2\beta\right) + \frac{1}{2} \times (\pi - \beta - t) + \frac{3}{2} \times \left(\frac{\pi}{2}\right) + 2 \times (-\beta + t)\right]\\ &\geq \frac{5}{8} - \frac{\beta}{2\pi},\end{aligned}$$

  where the equality is achieved when $t = \beta$.

Therefore, the unique best-in-class hypothesis for adversarial 0/1 loss ℓ_γ is $w^*_{\ell_\gamma, \mathcal{H}_{\mathrm{lin}}} = \left(\cos t^*_{\ell_\gamma}, \sin t^*_{\ell_\gamma}\right)$, where $t^*_{\ell_\gamma} = \frac{\pi}{2} - \beta$ and the best-in-class error for ℓ_γ is $\mathcal{R}^*_{\ell_\gamma, \mathcal{H}_{\mathrm{lin}}} = \frac{7}{8} - \frac{3\beta}{2\pi}$.

For the linear hypothesis set $\mathcal{H}_{\mathrm{lin}}$, $\widetilde{\ell}_{\mathrm{trades}}$ can be written as

$$\begin{aligned}\widetilde{\ell}_{\mathrm{trades}}(h, x, y) &= \Phi_{\log}(yh(x)) + \sup_{x' : \|x - x'\| \leq \gamma} \Phi_{\log}(h(x)h(x')/\lambda)\\ &= \Phi_{\log}(yw \cdot x) + \Phi_{\log}\left(\inf_{x' : \|x - x'\| \leq \gamma} (w \cdot x)(w \cdot x')/\lambda\right) \qquad (16)\\ &= \Phi_{\log}(yw \cdot x) + \Phi_{\log}\left(\left(|w \cdot x|^2 - \gamma |w \cdot x|\right)/\lambda\right).\end{aligned}$$

Thus, for any $h \in \mathcal{H}_{\text{lin}}$, the generalization error of $\widetilde{\ell}_{\text{trades}}$ can be expressed as

$$\begin{aligned}
&\mathcal{R}_{\widetilde{\ell}_{\text{trades}}}(h)\\
&= \frac{1}{2\pi}\bigg(\int_0^{\pi} \frac{1}{2}\Phi_{\log}(\cos(\theta-t)) + \frac{3}{2}\Phi_{\log}(\sin(\theta-t))\,d\theta + \int_0^{2\pi} \Phi_{\log}\Big(\cos(\theta-t)^2 - \gamma|\cos(\theta-t)|/\lambda\Big)\bigg)\,d\theta\\
&= \frac{1}{2\pi}\int_{-t}^{\pi-t} \frac{1}{2}\Phi_{\log}(\cos\theta) + \frac{3}{2}\Phi_{\log}(\sin\theta)\,d\theta + \frac{1}{2\pi}\int_0^{2\pi} \Phi_{\log}\Big(\cos(\theta-t)^2 - \gamma|\cos(\theta-t)|/\lambda\Big)\,d\theta \quad \text{(change of variables)}\\
&= \frac{1}{2\pi}\int_{-t}^{\pi-t} \frac{1}{2}\Phi_{\log}(\cos\theta) + \frac{3}{2}\Phi_{\log}(\sin\theta)\,d\theta + \text{constant} \quad \Big(\text{constant} = \tfrac{1}{2\pi}\textstyle\int_0^{2\pi} \Phi_{\log}\Big(\cos(\theta)^2 - \gamma|\cos(\theta)|/\lambda\Big)\,d\theta\Big)\\
&= \mathcal{U}_{\text{trades}}(t) + \text{constant}. \quad \Big(\mathcal{U}_{\text{trades}}(t) = \tfrac{1}{2\pi}\textstyle\int_{-t}^{\pi-t} \tfrac{1}{2}\Phi_{\log}(\cos\theta) + \tfrac{3}{2}\Phi_{\log}(\sin\theta)\,d\theta\Big)
\end{aligned}$$

Since $\Phi_{\log}$ is continuous, by Leibniz Integral Rule, the best-in-class hypothesis $w^*_{\widetilde{\ell}_{\text{trades}},\mathcal{H}_{\text{lin}}} = \Big(\cos t^*_{\widetilde{\ell}_{\text{trades}}}, \sin t^*_{\widetilde{\ell}_{\text{trades}}}\Big)$ for $\widetilde{\ell}_{\text{trades}}$ satisfies that $\mathcal{U}'_{\text{trades}}\Big(t^*_{\widetilde{\ell}_{\text{trades}}}\Big) = 0$, that is,

$$\begin{aligned}
&-\frac{1}{2}\Phi_{\log}\Big(\cos\Big(\pi - t^*_{\widetilde{\ell}_{\text{trades}}}\Big)\Big) + \frac{1}{2}\Phi_{\log}\Big(\cos\Big(-t^*_{\widetilde{\ell}_{\text{trades}}}\Big)\Big) - \frac{3}{2}\Phi_{\log}\Big(\sin\Big(\pi - t^*_{\widetilde{\ell}_{\text{trades}}}\Big)\Big) + \frac{3}{2}\Phi_{\log}\Big(\sin\Big(-t^*_{\widetilde{\ell}_{\text{trades}}}\Big)\Big) = 0\\
&\implies \frac{1}{2}\Big[\Phi_{\log}\Big(\cos t^*_{\widetilde{\ell}_{\text{trades}}}\Big) - \Phi_{\log}\Big(-\cos t^*_{\widetilde{\ell}_{\text{trades}}}\Big)\Big] - \frac{3}{2}\Big[\Phi_{\log}\Big(\sin t^*_{\widetilde{\ell}_{\text{trades}}}\Big) - \Phi_{\log}\Big(-\sin t^*_{\widetilde{\ell}_{\text{trades}}}\Big)\Big] = 0\\
&\implies t^*_{\widetilde{\ell}_{\text{trades}}} \neq \frac{\pi}{2} - \beta = t^*_{\ell_\gamma}, \text{ where } \beta = \arccos(0.1).
\end{aligned}$$

Therefore, $\widetilde{\ell}_{\text{trades}}$ with any $\lambda > 0$ is not $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ.

On the other hand, for the linear hypothesis set $\mathcal{H}_{\text{lin}}$, Φ_{smooth} can be written as

$$\begin{aligned}
\Phi_{\text{smooth}} &= \Phi_\rho(yh(x)) + \frac{1}{\rho}\bigg(yh(x) - \inf_{x':\|x-x'\|\le\gamma} yh(x')\bigg)\\
&= \Phi_\rho(yw\cdot x) + \frac{1}{\rho}\bigg(yw\cdot x - \inf_{x':\|x-x'\|\le\gamma} yw\cdot x'\bigg)\\
&= \Phi_\rho(yw\cdot x) + \frac{\gamma}{\rho}.
\end{aligned} \tag{17}$$

Then, the generalization error of Φ_{smooth} can be expressed as

$$\begin{aligned}
\mathcal{R}_{\Phi_{\text{smooth}}}(h) &= \frac{\gamma}{\rho} + \frac{1}{2\pi}\int_0^{\pi} \frac{1}{2}\Phi_\rho(\cos(\theta-t)) + \frac{3}{2}\Phi_\rho(\sin(\theta-t))\,d\theta\\
&= \frac{\gamma}{\rho} + \frac{1}{2\pi}\int_{-t}^{\pi-t} \frac{1}{2}\Phi_\rho(\cos\theta) + \frac{3}{2}\Phi_\rho(\sin\theta)\,d\theta \quad \text{(change of variables)}\\
&= \frac{\gamma}{\rho} + \mathcal{U}_{\text{smooth}}(t). \quad \Big(\mathcal{U}_{\text{smooth}}(t) = \tfrac{1}{2\pi}\textstyle\int_{-t}^{\pi-t} \tfrac{1}{2}\Phi_\rho(\cos\theta) + \tfrac{3}{2}\Phi_\rho(\sin\theta)\,d\theta\Big)
\end{aligned}$$

Let $\rho = 0.3 \in (0.1, \sqrt{0.99}) = (\cos\beta, \sin\beta)$. Note that $\arccos(\rho) = \arccos(0.3) \in \big(\frac{\pi}{4}, \frac{\pi}{2}\big)$. Since Φ_ρ is continuous, by Leibniz Integral Rule, the best-in-class hypothesis $w^*_{\Phi_{\text{smooth}},\mathcal{H}_{\text{lin}}} = \Big(\cos t^*_{\Phi_{\text{smooth}}}, \sin t^*_{\Phi_{\text{smooth}}}\Big)$ for Φ_{smooth} satisfies that $\mathcal{U}'_{\text{smooth}}\Big(t^*_{\Phi_{\text{smooth}}}\Big) = 0$, that is,

$$\frac{1}{2}\Big[\Phi_\rho\Big(\cos t^*_{\Phi_{\text{smooth}}}\Big) - \Phi_\rho\Big(-\cos t^*_{\Phi_{\text{smooth}}}\Big)\Big] - \frac{3}{2}\Big[\Phi_\rho\Big(\sin t^*_{\Phi_{\text{smooth}}}\Big) - \Phi_\rho\Big(-\sin t^*_{\Phi_{\text{smooth}}}\Big)\Big] = 0. \tag{18}$$

By solving (18) and plugging the solutions in $\mathcal{U}_{\text{smooth}}(t)$, we obtain $t^*_{\Phi_{\text{smooth}}} = \frac{\pi}{2} - \beta$, which is consistent with $t^*_{\ell_\gamma}$. Therefore, Φ_{smooth} with $\rho = 0.3$ is $\mathcal{H}_{\text{lin}}$-consistent with respect to ℓ_γ on this distribution. □