Paper Review in Adversarial Robustness Theory and Algorithms

Genghis Luo KL4747@NYU.EDU
NYU Shanghai
Jackie Chen JC11815@NYU.EDU
NYU Shanghai
Daniel Jin HJ2528@NYU.EDU
NYU

Instructor: Mehryar Mohri

Abstract

By looking through recent works in adversarial robustness (e.g. Goodfellow et al. (2015); Ma (2018); Zhang et al. (2019); Awasthi et al. (2023)), we start by defining the question of what adversarial robustness is and why it is important. We then consider frameworks for training robust models, and survey theoretical results that provide insights into the fundamental trade-offs between accuracy and robustness. Specifically, Zhang et al. (2019) introduces TRADES, a theory-based algorithm for balancing this trade-off, and Awasthi et al. (2023) introduces a thorough theoretical framework for adversarial robustness theory. Overall, we examine recent advances that improve training by leveraging conditions such as classification-calibrated surrogate losses and the concept of H-consistency, thereby guiding the design of robust models that maintain strong theoretical guarantees.

Keywords: Adversarial Robustness, Robust Optimization, TRADES, H-Consistency, Classification-Calibrated Surrogate Loss

1. Introduction

Deep neural networks have achieved remarkable success in recent years, yet their vulnerability to adversarial perturbations remains a significant concern. Goodfellow et al. (2015) demonstrated that even small, imperceptible changes to input data can deceive state-of-the-art models with high confidence. This phenomenon challenges the reliability of neural networks in safety-critical applications. Further research by Ma (2018) showed that existing defenses, despite their apparent success, often fail against stronger, adaptive adversaries. These findings suggest that current methods lack the theoretical guarantees necessary for practical robustness.

A key issue in addressing adversarial robustness is the fundamental trade-off between robustness and accuracy. This trade-off was formalized by Zhang et al. (2019), who introduced TRADES, a theoretically principled framework for balancing robustness and accuracy. In parallel, work by Awasthi et al. (2023) and others provided rigorous theoretical foundations, illustrating that certain adversarial robustness guarantees can only be achieved at the expense of reduced natural accuracy.

At the same time, the concept of H-consistency has emerged as a critical property for surrogate loss functions, ensuring alignment between the surrogate loss and the true classification loss under adversarial settings. This concept guides the choice of training objectives, ensuring that improvements in surrogate loss directly translate into improved adversarial robustness.

The goal of this work is to revisit and connect key theories and results from foundational research on adversarial robustness. By analyzing these contributions, we aim to develop a deeper understanding of both the theoretical and practical aspects of adversarial defenses, ultimately bridging the gap between accuracy and robustness.

2. Methodology

2.1. Formulation of Robust Optimization

As shown in Ma (2018), the robust optimization objective can be formalized as:

$$\min_{\theta} \rho(\theta), \quad \text{where } \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in \mathcal{S}} \mathcal{L}(\theta, x+\delta, y) \right],$$

where $\mathcal{S} \subseteq \mathbb{R}^d$ represents the set of allowable perturbations. Minimizing $\rho(\theta)$ ensures that the loss remains small for all allowed adversarial perturbations. Thus, the search for robust models reduces to solving a well-defined optimization problem.

Standard optimization techniques, such as Stochastic Gradient Descent (SGD), cannot be directly applied to this saddle-point formulation because the adversarial loss involves solving the inner maximization problem at each iteration. The paper applies Danskin's Theorem to address this issue. Specifically, for a differentiable function $g(\theta, \delta)$, where $\delta \in S$, the max-function $\phi(\theta) = \max_{\delta \in S} g(\theta, \delta)$ has a well-defined gradient, given by:

$$\phi'(\theta, h) = \sup_{\delta \in \delta^*(\theta)} h^T \nabla_\theta g(\theta, \delta).$$

If $\delta^*(\theta)$ is a singleton, then $\phi(\theta)$ is differentiable, and its gradient satisfies:

$$\nabla \phi(\theta) = \nabla_\theta g(\theta, \delta^*(\theta)).$$

The paper establish the following Corollary, which states that the negative gradient $-\nabla_\theta L(\theta, x+\bar{\delta}, y)$ provides a valid descent direction for the outer optimization problem. Formally, let $\bar{\delta}$ be a maximizer of $\max_{\delta \in S} L(\theta, x+\delta, y)$. Then, as long as $\nabla_\theta L(\theta, x+\bar{\delta}, y)$ is nonzero, we have:

$$\phi'(\theta, h) = \sup_{\delta \in \delta^*(\theta)} h^T \nabla_\theta L(\theta, x+\delta, y) \geq h^T h = \|\nabla_\theta L(\theta, x+\bar{\delta}, y)\|_2^2 \geq 0.$$

This result guarantees that $-\nabla_\theta L(\theta, x+\bar{\delta}, y)$ is a descent direction for the max-function $\phi(\theta)$, and also the saddle point problem defined earlier in this section.

We then utilize all the results above to give a principled approach to compute gradients for robust optimization efficiently. Note that the ReLU activation function and max-pooling operations in neural networks may cause the objective function to be non-differentiable at certain points. However, these points are of measure zero and thus do not pose practical issues. Additionally, the inner maximization problem can be non-concave, but this can be addressed by selecting a concave subset of the domain and applying algorithms within that subset.

In summary, robust optimization reduces the challenge of adversarial robustness to solving a well-defined optimization problem. By explicitly considering worst-case perturbations and leveraging theoretical tools like Danskin's theorem, we can design algorithms that effectively optimize for adversarial robustness.

However, a deeper issue arises from the saddle point optimization formulation itself:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in S} \mathcal{L}(\theta, x+\delta, y) \right].$$

This formulation focuses exclusively on minimizing the loss at adversarially perturbed inputs while neglecting the natural data distribution. The primary consequence is that the model is inherently

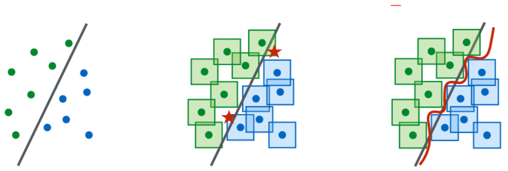


Figure 1: Increasing Complexity in Adversarial Training

forced to **overfit adversarial examples**, which introduces unnecessary complexity into the decision boundary.

To illustrate this, consider the geometric impact of adversarial examples on the decision boundary. In the absence of adversarial perturbations, the model can rely on a relatively simple decision boundary, such as a linear separator, to classify natural data points. However, adversarial examples expand the input space that must be correctly classified, particularly in regions defined by the ℓ_∞-balls around natural data points. As a result, the decision boundary becomes significantly more complex and nonlinear to account for the adversarial perturbations, shown in Figure 1.

This phenomenon further increases the dependency on model capacity. To learn these complex decision boundaries, the model requires larger architectures, such as deeper layers or wider hidden units, which come with higher computational costs. Thus, the saddle point formulation inherently drives the model toward greater capacity, not necessarily because of a fundamental need for complexity, but due to the exclusive focus on adversarial data.

2.2. Establishment of Theoretical Framework and Improvement on the Adversarial Training

The key motivation of Zhang et al. (2019)'s work is to address the overfitting issue and the lack of theoretical framework in adversarial training. All the detailed work has been done and proved in binary classification problem. Define $\mathcal{R}_{\text{rob}}$ to characterize the robustness of a score function $f : \mathcal{X} \to \mathbb{R}$ by:

$$\mathcal{R}_{\text{rob}}\left(f\right) := \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{D}}\mathbf{1}_{\{\exists \boldsymbol{X}'\in\mathbb{B}(\boldsymbol{X},\epsilon)\text{ s.t. } f(\boldsymbol{X}')Y\leq 0\}}$$

Write the natural generalization error as:

$$\mathcal{R}_{\text{nat}}\left(f\right) := \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{D}}\mathbf{1}_{\{f(\boldsymbol{X})Y\leq 0\}}$$

Note that the two errors satisfy $\mathcal{R}_{\text{rob}}\left(f\right) \geq \mathcal{R}_{\text{nat}}\left(f\right)$ for all f the robust error is equal to the natural error when $\epsilon = 0$.

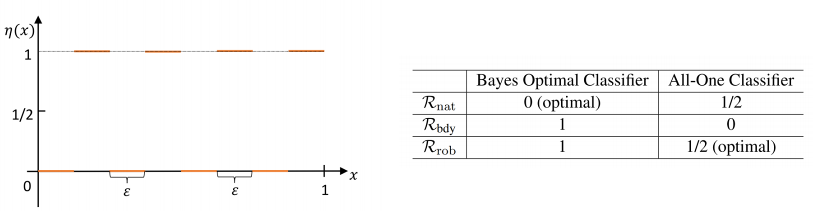
Introduce the boundary error defined as:

$$\mathcal{R}_{\text{bdy}}\left(f\right) := \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{D}}\mathbf{1}_{\{\boldsymbol{X}\in\mathbb{B}(\text{DB}(f),\epsilon), f(\boldsymbol{X})Y>0\}}$$

It can be easily seen that

$$\mathcal{R}_{\text{rob}}\left(f\right) = \mathcal{R}_{\text{nat}}\left(f\right) + \mathcal{R}_{\text{bdy}}\left(f\right)$$

as the first term $\mathcal{R}_{\text{nat}}\left(f\right)$ includes all misclassified points regarding the accuracy, and the second term $\mathcal{R}_{\text{bdy}}\left(f\right)$ includes all the points that are classified correctly but within $\mathbb{B}(\text{DB}(f), \epsilon)$, regarding the robustness.



	Bayes Optimal Classifier	All-One Classifier
$\mathcal{R}_{\text{nat}}$	0 (optimal)	1/2
$\mathcal{R}_{\text{bdy}}$	1	0
$\mathcal{R}_{\text{rob}}$	1	1/2 (optimal)

Figure 2: Trade-off Between $\mathcal{R}_{\text{nat}}(f)$ and $\mathcal{R}_{\text{bdy}}(f)$

There is in fact a trade-off between $\mathcal{R}_{\text{nat}}(f)$ and $\mathcal{R}_{\text{bdy}}(f)$, showcased by the following toy example: Consider the case $(X, Y) \sim \mathcal{D}$, where the marginal distribution over the sample space $\mathcal{X}$ is a uniform distribution over $[0, 1]$, and for $k = 0, 1, \ldots, \lceil \frac{1}{2\epsilon} - 1 \rceil$,

$$\begin{aligned} \eta(x) &:= \Pr(Y = 1 \mid X = x) \\ &= \begin{cases} 0, & x \in [2k\epsilon, (2k+1)\epsilon) \\ 1, & x \in [(2k+1)\epsilon, (2k+2)\epsilon) \end{cases} \end{aligned}$$

The results are shown in Figure 2.

Our goal is then to derive a good upper bound on $\mathcal{R}_{\text{rob}}(f)$ that we want to minimize, in the sense that a free hyper-parameter can be introduced to manipulate the trade-off between accuracy and robustness, and therefore a good algorithm can be derived to minimize this upper bound. We need several tools to achieve this goal.

Introduce the surrogate loss $\mathcal{R}_\phi(f) := \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{D}}\phi(f(\boldsymbol{X})Y)$. Formally, for $\eta \in [0, 1]$, define the conditional ϕ-risk by

$$H(\eta) := \inf_{\alpha\in\mathbb{R}} C_\eta(\alpha) := \inf_{\alpha\in\mathbb{R}} (\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)),$$

and define $H^-(\eta) := \inf_{\alpha(2\eta-1)\le 0} C_\eta(\alpha)$.

The classification-calibrated condition requires that imposing the constraint that α has an inconsistent sign with the Bayes decision rule $\text{sign}(2\eta - 1)$ leads to a strictly larger ϕ-risk:

Assumption 1 (Classification-Calibrated Condition): Assume that the surrogate loss ϕ is classification-calibrated, meaning that for any $\eta \neq 1/2$, $H^-(\eta) > H(\eta)$, i.e., Bayesian estimator is always the minimizer.

One remark is that classfication-calibrated is a weak condition on the surrogate loss, by which ensures a rich class. Surrogate losses like Hinge loss, Sigmoid loss, Exponential loss, and Logistic loss are within this class.

Define the ψ transform of classification-calibrated surrogate loss $\phi : [0, 1] \to [0, \infty)$ by

$$\psi = \widetilde{\psi}^{**}$$

where $\widetilde{\psi}(\theta) := H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right)$. In fact, the function $\psi(\theta)$ is the largest convex lower bound on $\tilde{\psi}$. The value $H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right)$ characterizes how close the surrogate loss ϕ is to the class of non-classification-calibrated losses.

Lemma 2.1 [Bartlett et al. (2006)]: Under Assumption 1, the function ψ has the following properties: ψ is non-decreasing, continuous, convex on $[0, 1]$ and $\psi(0) = 0$.

By using good properties of this ψ transform, we can derive a tight upper bound in the sense of the following two theorems:

Theorem 3.1 [Zhang et al. (2019)]: Let $\mathcal{R}_\phi(f) := \mathbb{E}_\phi[f(\mathbf{X})Y]$ and $R^*_\phi := \min_f \mathcal{R}_\phi(f)$. Under Assumption 1, for any non-negative loss function ϕ such that $\phi(0) \geq 1$, any measurable $f : \mathcal{X} \to \mathbb{R}$, any probability distribution on $\mathcal{X} \times \{\pm 1\}$, and any $\lambda > 0$, we have:

$$\begin{aligned} \mathcal{R}_{\text{rob}}(f) - \mathcal{R}^*_{\text{nat}} &= \mathcal{R}_{\text{nat}}(f) - \mathcal{R}^*_{\text{nat}} + \mathcal{R}_{\text{bdy}}(f) \\ \leq \psi^{-1}\left(\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi\right) &+ \Pr[\mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon), f(\mathbf{X})Y > 0] \\ \leq \psi^{-1}\left(\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi\right) &+ \mathbb{E}\left(\max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon)} \phi\left(f\left(\mathbf{X}'\right) f(\mathbf{X})/\lambda\right)\right) \end{aligned}$$

Theorem 3.2 [Zhang et al. (2019)]: Suppose that $|\mathcal{X}| \geq 2$. Under Assumption 1, for any non-negative loss function ϕ such that $\phi(x) \to 0$ as $x \to +\infty$, any $\xi > 0$, and any $\theta \in [0, 1]$, there exists a probability distribution on $\mathcal{X} \times \{\pm 1\}$, a function $f : \mathbb{R}^d \to \mathbb{R}$, and a regularization parameter $\lambda > 0$ such that $\mathcal{R}_{\text{rob}}(f) - \mathcal{R}^*_{\text{nat}} = \theta$ and

$$\begin{aligned} \psi\left(\theta - \mathbb{E} \max_{\boldsymbol{x}' \in \mathbb{B}(\boldsymbol{X}, \epsilon)} \phi\left(f\left(\boldsymbol{X}'\right) f(\boldsymbol{X})/\lambda\right)\right) &\leq \mathcal{R}_\phi(f) - \mathcal{R}^*_\phi \\ \leq \psi\left(\theta - \mathbb{E} \max_{\boldsymbol{x}' \in \mathbb{B}(\boldsymbol{X}, \epsilon)} \phi\left(f\left(\boldsymbol{X}'\right) f(\boldsymbol{X})/\lambda\right)\right) &+ \xi \end{aligned}$$

TRADES Algorithm [Zhang et al. (2019)]: Optimization on Upper Bound Theorems 3.1 and 3.2 shed light on algorithmic designs of adversarial defenses. In order to minimize $\mathcal{R}_{\text{rob}}(f) - \mathcal{R}^*_{\text{nat}}$, the theorems suggest minimizing [a]

$$\min_f \mathbb{E}\{\underbrace{\phi(f(\boldsymbol{X})Y)}_{\text{for accuracy}} + \underbrace{\max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X}, \boldsymbol{\epsilon})} \phi\left(f(\boldsymbol{X}) f\left(\boldsymbol{X}'\right)/\lambda\right)\}}_{\text{regularization for robustness}}$$

Heuristically, [Zhang et al. (2019)] use two heuristics to achieve more general defenses: a) extending to multi-class problems by involving multi-class calibrated loss; b) approximately solving the mini-max problem via alternating gradient descent. For multi-class problems, a surrogate loss is calibrated if minimizers of the surrogate risk are also minimizers of the $0 - 1$ risk [Pires and Szepesvári (2016)]. Examples of multi-class calibrated loss include cross-entropy loss. Algorithmically, [Zhang et al. (2019)] extend the problem to the case of multi-class classifications by replacing ϕ with a multi-class calibrated loss $\mathcal{L}(\cdot, \cdot)$:

$$\min_f \mathbb{E}\left\{\mathcal{L}(f(\boldsymbol{X}), \boldsymbol{Y}) + \max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X}, \epsilon)} \mathcal{L}\left(f(\boldsymbol{X}), f\left(\boldsymbol{X}'\right)\right)/\lambda\right\}$$

where $f(\boldsymbol{X})$ is the output vector of learning model (with soft-max operator in the top layer for the cross-entropy loss $\mathcal{L}(\cdot, \cdot)$), $\boldsymbol{Y}$ is the label-indicator vector, and $\lambda > 0$ is the regularization parameter.

2.3. H-Consistency

Adversarial training methods, such as TRADES, rely on surrogate loss functions because they are differentiable and convex, therefore, easier to optimize. While surrogate loss functions are bounded, it is essential to ensure that minimizing the surrogate loss also leads to minimizing the true target

loss. This connection is where the concept of *H-consistency* plays a pivotal role.
H-consistency is formally defined as:

$$\forall h \in H, \quad \mathcal{R}_{\text{target}}(h) - \mathcal{R}_{\text{target},H} \leq f(\mathcal{R}_{\phi}(h) - \mathcal{R}_{\phi,H}),$$

where $\mathcal{R}_{\text{target}}$ represents the true target loss, $\mathcal{R}_{\phi}$ is the surrogate loss, and H is the hypothesis space. Intuitively, this inequality ensures that the gap between the true target loss and the surrogate loss is bounded by a function of their respective differences. In simpler terms, as the surrogate loss decreases, the true target loss cannot increase within a given set of models H. This property is critical for ensuring that adversarial training methods remain effective in practice.

However, TRADES's surrogate loss has been shown to fail the H-consistency bound in certain scenarios Awasthi et al. (2023), particularly in multi-class classification tasks. In these cases, models optimized with TRADES can yield inaccurate predictions despite achieving low surrogate loss values, leading to unfavorable hypothesis being selected.

To address this limitation, a family surrogate loss function called *Smooth Adversarial Losses* was introduced in Awasthi et al. (2023). Which satisfies the H-consistency and bounded under:

$$\Phi_{smooth} \leq \Phi(yh(x)) + \nu \left| yh(x) - \inf_{x' : \|x - x'\| \leq \gamma} yh(x') \right|$$

Building on this improvement, the *Principled Smooth Adversarial Loss (PSAL)* algorithm was developed. PSAL outperforms TRADES and other state-of-the-art methods in terms of both clean accuracy and robustness to adversarial perturbations. This demonstrates that addressing the H-consistency limitation of TRADES can lead to more reliable adversarial training frameworks.

3. Future work

In TRADES, a hyperparameter λ was introduced to balance the trade-off between accuracy and robustness. However, λ does not influence the theoretical proofs of the bounds and is only constrained by the loose condition $\lambda \geq 0$. Which is a reasonable and intuitive hyperparameter for empirical tuning. However this might also suggests that there is significant room for further research and development to improve the algorithm.

4. Appendix

4.1. Proof of Theorem 3.1

Proof Claim: Classification-Calibration [Bartlett et al. (2006)]

$$\psi\left(R_{\text{nat}}(f) - R^*_{\text{nat}}\right) \leq R_\phi(f) - R^*_\phi.$$

Proof.

$$\begin{aligned} R_{\text{nat}}(f) - R^*_{\text{nat}} &= R_{\text{nat}}(f) - R\left(\eta - \frac{1}{2}\right) \\ &\overset{(1)}{=} \mathbb{E}\left[\mathbf{1}_{\left\{\text{sign}(f(x)) \neq \text{sign}\left(\eta(X) - \frac{1}{2}\right)\right\}} \cdot |2\eta(X) - 1|\right] \end{aligned}$$

where (1) is because $|(1-\eta) - \eta| = |2\eta - 1|$. Therefore,

$$\begin{aligned} \psi\left(R_{\text{nat}}(f) - R^*_{\text{nat}}\right) &\overset{(2)}{\leq} \mathbb{E}\left[\psi\left(\mathbf{1}_{\left\{\text{sign}(f(x)) \neq \text{sign}\left(g(x) - \frac{1}{2}\right)\right\}} \cdot |2g(x) - 1|\right)\right] \\ &\overset{(3)}{=} \mathbb{E}\left[\mathbf{1}_{\left\{\text{sign}(f(X)) \neq \text{sign}\left(n(x) - \frac{1}{2}\right)\right\}} \cdot \varphi(|2n(X) - 1|)\right] \\ &\overset{(4)}{\leq} \mathbb{E}\left[\mathbf{1}_{\left\{\text{sign}(f(X)) \neq \text{sign}\left(\eta(X) - \frac{1}{2}\right)\right\}} \cdot \tilde{\psi}(|2g(X) - 1|)\right] \\ &= \mathbb{E}\left[\mathbf{1}_{\{\text{sign}(f(X)) \neq \text{signn}(\eta(\text{X})\text{-}\frac{1}{2})\}} \cdot \left(\inf_{\alpha:\alpha(2\eta(X)-1)\leq 0} C_{\eta(X)}(\alpha) - H(\eta(X))\right)\right. \\ &\overset{(5)}{\leq} \mathbb{E}\left[C_{g(X)}(f(X)) - H(g(X))\right] \\ &\overset{(6)}{=} R_\phi(f) - R^*_\phi. \end{aligned}$$

where (2) is because of Jensen's Inequality, (3) is by $\psi(0) = 0$, (4) is by ψ being the convex lower bound of $\widetilde{\psi}$, (5) is because when $\text{sign}(f(X)) \neq \text{sign}\left(\eta(X) - \frac{1}{2}\right)$, $f(X)$ is a valid α for H^{-1}; otherwise clear, and (6) is because $\mathbb{E}\left[C_{\eta(X)}\right] = \mathcal{R}_\phi(f)$.

By the claim, we can directly prove Theorem 3.1:

Proof.

$$\begin{aligned} \mathcal{R}_{\text{rob}}(f) - \mathcal{R}^*_{\text{nat}} &= \mathcal{R}_{\text{nat}}(f) - \mathcal{R}^*_{\text{nat}} + \mathcal{R}_{\text{bdy}}(f) \\ &\leq \psi^{-1}\left(\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi\right) + \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{D}} \mathbf{1}_{\{\boldsymbol{X} \in \mathbb{B}(\text{DB}(f),\epsilon), f(\boldsymbol{X})Y > 0\}} \\ &= \psi^{-1}\left(\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi\right) + \Pr[\boldsymbol{X} \in \mathbb{B}(\text{DB}(f),\epsilon), f(\boldsymbol{X})Y > 0] \\ &\leq \psi^{-1}\left(\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi\right) + \Pr[\boldsymbol{X} \in \mathbb{B}(\text{DB}(f),\epsilon)] \\ &= \psi^{-1}\left(\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi\right) + \mathbb{E} \max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X},\epsilon)} \mathbf{1}\left\{f\left(\boldsymbol{X}'\right) f(\boldsymbol{X}) \leq 0\right\} \\ &= \psi^{-1}\left(\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi\right) + \mathbb{E} \max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X},\epsilon)} \mathbf{1}\left\{f\left(\boldsymbol{X}'\right) f(\boldsymbol{X})/\lambda \leq 0\right\} \\ &\leq \psi^{-1}\left(\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi\right) + \mathbb{E} \max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X},\epsilon)} \phi\left(f\left(\boldsymbol{X}'\right) f(\boldsymbol{X})/\lambda\right) \end{aligned}$$

■

References

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Theoretically Grounded Loss Functions and Algorithms for Adversarial Robustness. *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 2023.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, Classification, and Risk Bounds. *Journal of the American Statistical Association*, 101(473):138–156, March 2006. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214505000000907. URL http://www.tandfonline.com/doi/abs/10.1198/016214505000000907.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples, March 2015. URL http://arxiv.org/abs/1412.6572. arXiv:1412.6572 [stat].

Aleksander Ma. Towards Deep Learning Models Resistant to Adversarial Attacks. *International Conference on Learning Representations*, 2018.

Bernardo Ávila Pires and Csaba Szepesvári. Multiclass Classification Calibration Functions, September 2016. URL http://arxiv.org/abs/1609.06385. arXiv:1609.06385.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically Principled Trade-off between Robustness and Accuracy. *Proceedings of the 36th International Conference on Machine Learning*, 2019.