ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413587

ADVERSARIALLY ROBUST CLASSIFICATION BASED ON GLRT

Bhagyashree Puranik, Upamanyu Madhow, Ramtin Pedarsani

University of California Santa Barbara

ABSTRACT

Machine learning models are vulnerable to adversarial attacks that can often cause misclassification by introducing small but well designed perturbations. In this paper, we explore, in the setting of classical composite hypothesis testing, a defense strategy based on the generalized likelihood ratio test (GLRT), which jointly estimates the class of interest and the adversarial perturbation. We evaluate the GLRT approach for the special case of binary hypothesis testing in white Gaussian noise under ℓ_∞ norm-bounded adversarial perturbations, a setting for which a minimax strategy optimizing for the worst-case attack is known. We show that the GLRT approach yields performance competitive with that of the minimax approach under the worst-case attack, while yielding a better robustness-accuracy trade-off under weaker attacks. The GLRT defense is applicable in multi-class settings and generalizes naturally to more complex models for which optimal minimax classifiers are not known.

***Index Terms*—** Adversarial machine learning, hypothesis testing, robust classification

1. INTRODUCTION

Machine learning models such as deep neural networks and regression methods have become pervasively deployed in large-scale commercial applications that are safety-critical, such as facial recognition for surveillance, autonomous driving and virtual assistants. It has been shown that an adversary is often able to add small perturbations to signals in an intelligent way to cause misclassification with high confidence [1, 2]. In applications that demand robustness in machine learning methods, adversarial attacks are fundamental threats. There have been several defense mechanisms suggested, followed by proposal of stronger adversaries to circumvent the defenses [3, 4]. A state-of-the-art defense [5] against such attacks is to train with adversarial examples–this is purely empirical and cannot provide robustness guarantees or insights.

In this paper, we seek fundamental insight by investigating adversarial classification in the setting of classical hypothesis testing, in which the class-conditional distributions of the data is known. We propose the well-known GLRT as a general approach to defense, in which the desired class and the action of the adversary (viewed as a nuisance parameter) are estimated jointly. The GLRT approach is general, since it applies to any composite hypothesis testing problem [6], unlike minimax strategies optimizing for worst-case attacks, which are difficult to find. We compare the GLRT and minimax approaches for a simple setting, binary Gaussian hypothesis testing with ℓ_∞ bounded attacks, for which the minimax strategy has been derived [7]. We show that the proposed GLRT approach provides competitive robustness guarantees when the attacker employs the full attack budget, while providing better robustness-accuracy trade-off for weaker attacks.

Related Work: There is a growing body of research on coming up with provable robustness guarantees against adversarial attacks [8, 9, 10, 11, 12, 13, 14, 15, 16]. A recent paper [17] addresses the problem of finding optimal robust classifiers in a binary classification problem, with the class conditional distributions possessing symmetric means and white Gaussian noise. For the case when perturbations are ℓ_∞ norm bounded, they restrict attention to the class of linear classifiers and then obtain optimum robust linear classifiers for two and three-class classification problems. In general, finding robust optimal classifiers for ℓ_∞ norm bounded adversarial perturbations is not easily tractable. Analytical results have been shown only for special cases, such as in [7], where minimax optimal robust classifiers are characterized in binary classification setting under Gaussian models with symmetric means, same covariance matrices and uniform priors, using ideas from optimal transport theory. Our proposed GLRT defense can be applied to multi-class Gaussian hypothesis problems with generic means and priors. In addition, the minimax classifier in [7] is pessimistic for weaker attacks, while the GLRT scheme performs better in such regimes as it estimating the action of the attacker.

2. GLRT-BASED DEFENSE

Throughout the paper, we represent vectors in boldface letters and scalars in regular letters. The norm $||\cdot||$ denotes ℓ_2 norm unless specified otherwise. Consider the following standard classification or hypothesis testing problem: $\mathcal{H}_k : \mathbf{X} \sim p_k(\mathbf{x})$. The presence of an adversary increases the uncertainty

This work was supported by the Army Research Office under grant W911NF-19-1-0053, and by the National Science Foundation under grant CCF 1909320.

about the class-conditional densities, which can be modeled as a composite hypothesis testing problem:

$$\mathcal{H}_k : \mathbf{X} \sim p_\theta(\mathbf{x}), \theta \in \Theta_k,$$

where the size of the uncertainty sets Θ_k depends on the constraints on the adversary. The GLRT defense consists of joint maximum likelihood estimation of the class and the adversary's parameter:

$$\hat{k} = \arg\max_k \max_{\theta \in \Theta_k} p_\theta(\mathbf{x}).$$

Gaussian hypothesis testing: We now apply this framework to Gaussian hypothesis testing with an adversary which can add an ℓ_∞-bounded perturbation $\mathbf{e}$: $||\mathbf{e}||_\infty \leq \epsilon$, where we term ϵ the "attack budget" or "adversarial budget".

$$\mathcal{H}_k : \mathbf{X} = \boldsymbol{\mu}_k + \mathbf{e} + \mathbf{N},$$

where $\mathbf{X} \in \mathbb{R}^d$, $\mathbf{N} \sim \mathcal{N}(0, \sigma^2 I_d)$ is white Gaussian noise. We assume that the adversary has access to the true hypothesis and knows the distributions under each of these hypotheses.

Conditioned on the hypothesis k and the perturbation $\mathbf{e}$, the negative log likelihood is a standard quadratic expression. Applying GLRT, we first estimate $\mathbf{e}$ under each hypothesis:

$$\hat{\mathbf{e}}_k = \arg\min_{\mathbf{e}:||\mathbf{e}||_\infty \leq \epsilon} ||\mathbf{X} - \boldsymbol{\mu}_k - \mathbf{e}||^2.$$

and then plug in to obtain the cost function to be minimized over k:

$$C_k = ||\mathbf{X} - \boldsymbol{\mu}_k - \hat{\mathbf{e}}_k||^2 \tag{1}$$

This yields intuitively pleasing answers in terms of the function $g_\epsilon(x) \triangleq \text{sign}(x)\max(0, |x| - \epsilon)$, which we term as the "double-sided ReLU" and its "complement," $f_\epsilon(x) = x - g_\epsilon(x)$. The estimated perturbation under hypothesis k is obtained as $\hat{\mathbf{e}}_k = f_\epsilon(\mathbf{X} - \boldsymbol{\mu}_k)$, where the non-linearity is applied coordinate-wise. Substituting into (1), we obtain

$$C_k = ||g_\epsilon(\mathbf{X} - \boldsymbol{\mu}_k)||^2 \tag{2}$$

where $g_\epsilon(.)$ is applied coordinate-wise. Thus, the GLRT detector

$$\hat{k} = \arg\min_k C_k$$

is a modified version of the standard minimum distance rule where the coordinate-wise differences between the observation and template are passed through the double-sided ReLU.

Minimax formulation: An alternative to the GLRT defense, which treats adversarial perturbation as a "nuisance parameter," is a game-theoretic formulation. Let $\mathcal{H}$ denote the true hypothesis and $\hat{\mathcal{H}}$ be a classifier. The adversary attempts to maximize the probability of error by choosing a suitable perturbation, while the defender tries to choose a classifier such that the expected probability of error is minimized. We consider the perturbations $\mathbf{e} : ||\mathbf{e}||_\infty \leq \epsilon$. Thus the optimum adversarial risk is:

$$R^* = \min_{\hat{\mathcal{H}}} \mathbb{E}\big[\sup_{\mathbf{e}:||\mathbf{e}||_\infty \leq \epsilon} \mathbb{1}(\hat{\mathcal{H}} \neq \mathcal{H})\big].$$

Clearly, this is the best possible approach for defending against worst-case attacks. Unfortunately, such minimax games are difficult to solve, unlike the more generally applicable GLRT approach. Furthermore, the optimal minimax solution may be overly conservative, unnecessarily compromising performance against attacks that are weaker than, or different from, the worst-case attack. In such scenarios, we expect the GLRT approach, which estimates the attack parameters, to provide an advantage. In order to compare the minimax and GLRT approaches, for the remainder of this paper, we specialize to a setting where the minimax solution is known: binary Gaussian hypothesis testing with symmetric means and equal priors.

3. BINARY GAUSSIAN HYPOTHESIS TESTING

We now focus on the binary hypothesis testing problem with equal priors for which the minimax rule is known [7]:

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{X} = \boldsymbol{\mu} + \mathbf{e} + \mathbf{N} \\ \mathcal{H}_1 &: \mathbf{X} = -\boldsymbol{\mu} + \mathbf{e} + \mathbf{N} \end{aligned}$$

where $\mathbf{e}$ is chosen by an ℓ_∞ bounded adversary, with adversarial budget ϵ, who knows the true hypothesis. In the absence of attack, the optimal rule is a minimum distance rule, which can be alternatively written as a linear detector with $\hat{\mathcal{H}} = \mathcal{H}_0$ if $\mathbf{w}_{clean}^T \mathbf{X} > 0$ and $\hat{\mathcal{H}} = \mathcal{H}_1$ otherwise, where $\mathbf{w}_{clean} = \boldsymbol{\mu}$ or any positive scalar multiple of it. It is shown in [7] that the minimax decision rule is also a linear detector, with $\mathbf{w}_{minimax} = g_\epsilon(\boldsymbol{\mu})$. One of the possible worst-case attacks for the minimax classifier that achieves its worst-case error, is $\mathbf{e} = -\epsilon.\text{sign}(\boldsymbol{\mu})$ under $\mathcal{H}_0$ and $\mathbf{e} = \epsilon.\text{sign}(\boldsymbol{\mu})$ under $\mathcal{H}_1$. Our numerical experiments suggest that the same attack achieves the worst performance under the GLRT scheme.

Under this attack, it is easy to see that the "defenseless" linear detector $\mathbf{w}_{clean}$ makes errors with probability at least half whenever the attack budget satisfies $\epsilon > ||\boldsymbol{\mu}||^2/||\boldsymbol{\mu}||_1$. Thus, the system is less vulnerable (i.e., the adversary needs a large attack budget) when the ℓ_1 norm of $\boldsymbol{\mu}$ is small relative to the ℓ_2 norm. That is, signal sparsity helps in robustness, as has been observed before [18, 16].

The minimax rule derived in [7] applies the double-sided ReLU to the "signal template" $\boldsymbol{\mu}$. Thus, it simply ignores signal coordinates whose sign could be flipped using the worst-case attack budget, and shrinks the remaining coordinates to provide an optimal rule *assuming that the worst-case attack*

has been applied. Comparing with the GLRT rule

$$C_1 = ||g_\epsilon(\mathbf{X}+\boldsymbol{\mu})||^2 \underset{<}{\overset{H_0}{>}} C_0 = ||g_\epsilon(\mathbf{X}-\boldsymbol{\mu})||^2 \quad (3)$$

we see that the GLRT defense applies the (coordinate-wise) double-sided ReLU to the difference between the observation and signal templates, and hence should be better able to adapt to the attack level (as long as it is smaller than the budget ϵ). The analysis below for GLRT scheme applies, without loss of generality, to asymmetric means (say $\boldsymbol{\mu_0}$ and $\boldsymbol{\mu_1}$), by shifting of coordinates equivalently, leading to the attack of $\mathbf{e} = -\epsilon.\text{sign}(\boldsymbol{\mu_0} - \boldsymbol{\mu_1})$ under $\mathcal{H}_0$.

3.1. Analysis

Since the GLRT rule is nonlinear, its performance is more difficult to characterize than that of a linear detector. However, we are able to provide insight via a central limit theorem (CLT) based approximation (which holds for large dimension d). By the symmetry of the observation model and the resulting symmetry induced on the attack model, we may condition on $\mathcal{H}_0$ and the corresponding attack $\mathbf{e} = -\epsilon.\text{sign}(\boldsymbol{\mu})$, and consider $\mathbf{X} = \boldsymbol{\mu} - \epsilon\text{sign}(\boldsymbol{\mu}) + \mathbf{N}$. The costs are:

$$\begin{aligned} C_0 &= \sum_{i=1}^{d} (g_\epsilon(-\epsilon\text{sign}(\boldsymbol{\mu}[i]) + \mathbf{N}[i]))^2 \\ C_1 &= \sum_{i=1}^{d} (g_\epsilon(2\boldsymbol{\mu}[i] - \epsilon\text{sign}(\boldsymbol{\mu}[i]) + \mathbf{N}[i]))^2. \end{aligned}$$

and the error probability of interest is $P_e = P_{e|0} = P[C = C_1 - C_0 < 0|\mathcal{H}_0]$.

We now perform a coordinate-wise analysis of the cost difference $C[i] = C_1[i] - C_0[i]$, where $C_k[i]$ indicates the contribution in cost C_k from coordinate i. Let the mean and variance of $C[i]$ be denoted by m_i and ρ_i^2 respectively. Applying CLT on the sum across coordinates, the error probability is then estimated as:

$$P_e = P_{e|0} = P\Big(\sum_{i=1}^{d} C[i] < 0\Big) \approx Q\left(\frac{\sum_{i=1}^{d} m_i}{\sqrt{\sum_{i=1}^{d} \rho_i^2}}\right). \quad (4)$$

The approximate equality in (4) can be formalized to exact equality in the limit under the mild assumption of satisfying Lindeberg's condition for CLT to hold for independent, but not necessarily identically distributed random variables.

Consider a particular coordinate i, set $C = C[i]$, and let $\boldsymbol{\mu}[i] = \mu$. Assume $\mu > 0$ without loss of generality: we simply replace μ by $|\mu|$ after performing our analysis, since the analysis is entirely analogous for $\mu < 0$, given the symmetry of the noise and the attack. We can numerically compute the mean and variance of the cost difference for the coordinate, $C = (g_\epsilon(2\mu + N - \epsilon))^2 - (g_\epsilon(N - \epsilon))^2$, but the following lower bound yields insight:

$$C \geq Y \triangleq \mathbb{1}_{\{N \geq -t\}}(t+N)^2 - N^2 \quad (5)$$

where $t = 2(\mu - \epsilon)$. Note that $t > 0$ ($|\mu| > \epsilon$) corresponds to coordinates that the minimax detector would retain. The high-SNR (t/σ large) behavior is interesting. For $t > 0$, we can show that $Y \approx t^2 + 2Nt$; these coordinates exhibit behavior similar to the minimax detector. On the other hand, for $t < 0$, $Y \approx -N^2$; these coordinates, which would have been deleted by the minimax detector, contribute noise in favor of the incorrect hypothesis (this becomes negligible at high SNR). These observations can be used to show that, at high SNR, the performance of the GLRT detector approaches that of the minimax detector under worst-case attack.

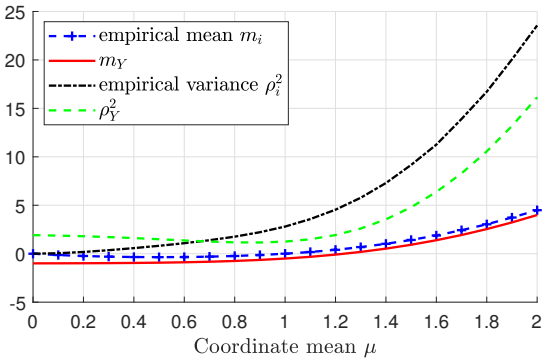

Fig. 1: Comparison of empirical mean and variance of $C[i]$ with the mean and variance of lower bounding variable Y_i.

Without loss of generality, let us redefine $t = 2(|\mu| - \epsilon)$. The mean and variance of Y, irrespective of sign(μ), can be computed in closed form as follows:

$$m_Y = Q\Big(\frac{-t}{\sigma}\Big)(t^2+\sigma^2) - \sigma^2 + \sigma t N\Big(\frac{t}{\sigma}; 0, 1\Big) \quad (6)$$

$$\begin{aligned} \rho_Y^2 &= 3\sigma^4 + Q\Big(\frac{-t}{\sigma}\Big)(t^4 + 4t^2\sigma^2 - 3\sigma^4) \\ &\quad + \sigma t N(t/\sigma; 0, 1)(t^2 + 3\sigma^2) - m_Y^2 \end{aligned} \quad (7)$$

where $N(.; 0, 1)$ denotes the density of standard Gaussian (zero-mean, unit-variance) random variable, and $Q(.)$ its complementary CDF. Figure 1 shows the empirical mean and empirical variance of $C[i]$, i.e., m_i and ρ_i^2, in comparison with m_Y and ρ_Y^2 obtained through (6) and (7). Here, the adversarial budget is set to $\epsilon = 1$ and noise variance $\sigma^2 = 1$.

The error probability in (4) can also be bounded by applying CLT on the lower bounding terms $Y_i \leq C[i]$ as follows:

$$P\Big(\sum_{i=1}^{d} C[i] < 0\Big) \leq P\Big(\sum_{i=1}^{d} Y_i < 0\Big) \approx Q\left(\frac{\sum_{i=1}^{d} m_{Y_i}}{\sqrt{\sum_{i=1}^{d} \rho_{Y_i}^2}}\right).$$

Bounding the probability of error in this fashion helps in yielding the following insight. Under low noise limit

($\sigma^2 \to 0$), the variance $\rho^2_{Y_i} = 0, \forall i$; and the mean is given by $m_{Y_i} = t^2$, if $|\boldsymbol{\mu}[i]| > \epsilon$, otherwise it is zero. Thus as long as $\exists i$ such that $|\boldsymbol{\mu}[i]| > \epsilon$, we have $P_e = 0$. Also note that since each of the means and variances are $\mathcal{O}(1)$ terms, we have $P_e \leq k_1 e^{-k_2 d}$, where k_1, k_2 are positive constants.

4. EXAMPLES AND DISCUSSION

Let a fraction p of the coordinates have means $\mu = a\epsilon_{des}$ and a fraction $(1-p)$ have $\mu = b\epsilon_{des}$, where $a > 1$ and $0 \leq b \leq 1$. Let the designed adversarial budget be ϵ_{des} and the actual attack be $\mathbf{e} = \mp\epsilon \text{sign}(\boldsymbol{\mu})$, where $\epsilon = k\epsilon_{des}$, $(k \leq 1)$. The effective signal-to-noise ratio (SNR) is:

$$\text{SNR}_{\text{minimax}} = (a-k)^2 dp\Big(\frac{\epsilon_{des}}{\sigma}\Big)^2$$

$$\text{SNR}_{\text{GLRT}} \approxeq d\frac{(pm_a + (1-p)m_b)^2}{p\rho_a^2 + (1-p)\rho_b^2}$$

where m_a and m_b are means, ρ_a^2 and ρ_b^2 are variances of a single component of $C[i]$ contributed by terms with component means $a\epsilon_{des}$ and $b\epsilon_{des}$ respectively. The probability of error is given by $Q(\sqrt{\text{SNR}})$. Note that for the GLRT detector, it is only an approximation as convergence is slow at high SNR, and we need to rely on simulations for true error probabilities.

We consider binary classification problems with symmetric means and uniform priors to draw a comparison with the minimax optimal scheme, and also a naive minimum distance classifier that is optimal under zero attack. The GLRT detector performs better than minimax for weaker attacks, and it has a significant advantage over minimax in settings where the class mean $\boldsymbol{\mu}$ has components which are smaller than ϵ_{des}, but larger than the actual attack. GLRT utilizes signal energy from these components while for minimax, such components are nulled. Figure 2 depicts the performance advantage of GLRT under weaker attacks, for a problem with parameters $\epsilon_{des} = 1$, $d = 20$, $p = 0.1$, $a = 1.1$, $b = 0.9$ and noise variance $\sigma^2 = 1$.

The naive minimum distance classifier does poorly under a large attack, specifically in settings where $\boldsymbol{\mu}$ has a large number of small components. Under strong attacks, these smaller components contribute to costs in such a way that the wrong class is favored by the naive detector. Consider a problem with parameters $d = 10$, $p = 0.1$, $\epsilon_{des} = 1$, $a = 2$, $b = 0.5$ and $\sigma^2 = 0.25$. The comparison of all three detectors under this setting is plotted in Figure 2, which clearly indicates the failure of naive scheme at high attacks, emphasizing the need for a robust detector. Figure 3 shows the variation of the error probability as a function of $(\epsilon_{des}/\sigma)^2$, under four different values of actual attack, for the same problem setting.

5. CONCLUSION

The GLRT approach to robust hypothesis testing explored in this paper can be generalized to complex models, in contrast to the difficulty of finding optimal minimax classifiers. For the simple model considered here, for which the minimax detector is known, we show that the GLRT detector has the same asymptotic performance as the minimax detector at high SNR for ℓ_∞ bounded adversarial perturbations at a designated attack level. For attack levels lower than this designated level, the GLRT detector can provide better performance, depending on the specific values of the signal components relative to the attack budget. Contrary to minimax, GLRT is a generic multi-class detector that can work with any priors.

An interesting direction for future research is to apply the GLRT approach to more complex data and attack models. It is also of interest to explore the minimax formulation in such settings: even if it is difficult to find the optimal minimax rule, a combination of insights from the minimax and GLRT formulations for simpler models might be useful.

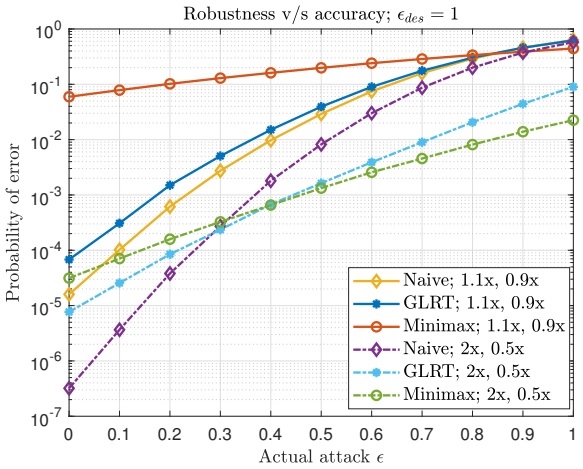

Fig. 2: Robustness v/s accuracy trade-off as the actual attack is varied, while the designed adversarial budget is fixed to $\epsilon_{des} = 1$.

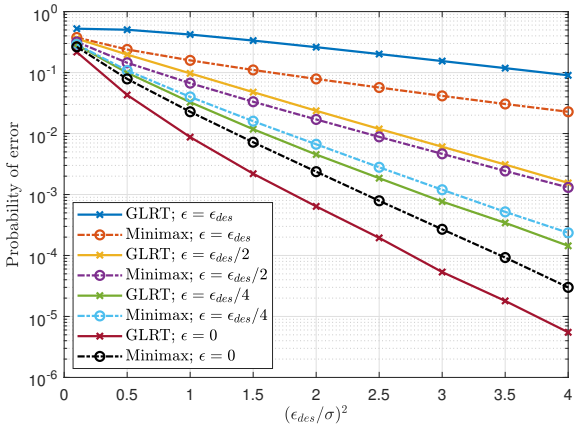

Fig. 3: Probability of error as a function of $(\epsilon_{des}/\sigma)^2$ for different values of actual attack (with $\epsilon_{des} = 1$, $a = 2$, $b = 0.5$).

6. REFERENCES

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations, 2014, Banff, AB, Canada, April 14-16*, 2014.

[2] Battista Biggio and Fabio Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition.*, vol. 84, pp. 317–331, 2018.

[3] Nicholas Carlini and David A. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy, San Jose, CA, USA, May 22-26,*, 2017, pp. 39–57.

[4] Anish Athalye, Nicholas Carlini, and David A. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning, ICML, Stockholm, Sweden, July 10-15*, 2018, pp. 274–283.

[5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3*, 2018.

[6] H Vincent Poor, *An introduction to signal detection and estimation*, Springer Science & Business Media, 2013.

[7] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal, "Lower bounds on adversarial robustness from optimal transport," in *Advances in Neural Information Processing Systems, 8-14 December 2019, Vancouver, BC, Canada*, 2019, pp. 7496–7508.

[8] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang, "Certified defenses against adversarial examples," in *6th International Conference on Learning Representations, ICLR, Vancouver, BC, Canada, April 30 - May 3, 2018*.

[9] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang, "Semidefinite relaxations for certifying robustness to adversarial examples," in *Advances in Neural Information Processing Systems, NeurIPS, 3-8 December 2018, Montréal, Canada*, pp. 10900–10910.

[10] Eric Wong and J. Zico Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *Proceedings of the 35th International Conference on Machine Learning, ICML, Stockholm, Sweden, July 10-15, 2018*.

[11] Eric Wong, Frank R. Schmidt, Jan Hendrik Metzen, and J. Zico Kolter, "Scaling provable adversarial defenses," in *Advances in Neural Information Processing Systems, NeurIPS, 3-8 December 2018, Montréal, Canada*, pp. 8410–8419.

[12] Aman Sinha, Hongseok Namkoong, and John C. Duchi, "Certifying some distributional robustness with principled adversarial training," in *6th International Conference on Learning Representations, ICLR, Vancouver, BC, Canada, April 30 - May 3, 2018*.

[13] Matthew Mirman, Timon Gehr, and Martin T. Vechev, "Differentiable abstract interpretation for provably robust neural networks," in *Proceedings of the 35th International Conference on Machine Learning, ICML, Stockholm, Sweden, July 10-15, 2018*.

[14] Matthias Hein and Maksym Andriushchenko, "Formal guarantees on the robustness of a classifier against adversarial manipulation," in *Advances in Neural Information Processing Systems, 4-9 December 2017, Long Beach, CA, USA*, pp. 2266–2276.

[15] Moustapha Cissé, Piotr Bojanowski, Edouard Grave, Yann N. Dauphin, and Nicolas Usunier, "Parseval networks: Improving robustness to adversarial examples," in *Proceedings of the 34th International Conference on Machine Learning, ICML, Sydney, NSW, Australia, 6-11 August 2017*.

[16] Zhinus Marzi, Soorya Gopalakrishnan, Upamanyu Madhow, and Ramtin Pedarsani, "Sparsity-based defense against adversarial attacks on linear classifiers," in *2018 IEEE International Symposium on Information Theory, ISIT, Vail, CO, USA, June 17-22, 2018*, pp. 31–35.

[17] Edgar Dobriban, Hamed Hassani, David Hong, and Alexander Robey, "Provable tradeoffs in adversarially robust classification," *arXiv preprint arXiv:2006.05161*, 2020.

[18] C. Bakiskan, S. Gopalakrishnan, M. Cekic, U. Madhow, and R. Pedarsani, "Polarizing Front Ends for Robust CNNs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 4257–4261.