Detecting Adversarial Examples Is (Nearly) As Hard As Classifying Them

Florian Tramèr [1] [2]

Abstract

Making classifiers robust to adversarial examples is challenging. Thus, many works tackle the seemingly easier task of *detecting* perturbed inputs.

We show a barrier towards this goal. We prove a *hardness reduction* between detection and classification of adversarial examples: given a robust detector for attacks at distance ϵ (in some metric), we show how to build a similarly robust (but computationally inefficient) *classifier* for attacks at distance $\epsilon/2$.

Our reduction is *computationally inefficient*, but preserves the *sample complexity* of the original detector. The reduction thus cannot be directly used to build practical classifiers. Instead, it is a useful sanity check to test whether empirical detection results imply something much stronger than the authors presumably anticipated (namely a highly robust and data-efficient *classifier*).

To illustrate, we revisit 14 empirical detector defenses published over the past years. For $12/14$ defenses, we show that the claimed detection results imply an inefficient classifier with robustness far beyond the state-of-the-art.

1. Introduction

Building models that are robust to adversarial examples (Szegedy et al., 2014; Biggio et al., 2013) is a major challenge and open problem in machine learning. Due to the inherent difficulty in building robust *classifiers*, researchers have attempted to build techniques to at least *detect* adversarial examples, a weaker task that is largely considered easier than robust classification (Xu et al., 2018; Pang et al., 2021; Sheikholeslami et al., 2021).

Yet, evaluating the robustness of detector defenses is challenging. This is in part due to a lack of strong evaluation guidelines and benchmarks—akin to those developed for robust classifiers (Carlini et al., 2019; Croce et al., 2020)—as well as to a lack of long-standing comparative baselines such as adversarial training (Madry et al., 2018). As a result, when a new detection defense is proposed, it can be hard to judge if the claimed empirical results are credible.

To illustrate this challenge, consider the following (fictitious) claims about two defenses against adversarial examples on CIFAR-10:

- defense A is a classifier that claims robust accuracy of 90% under ℓ_∞-perturbations bounded by $\epsilon = {}^{4}/_{255}$;
- defense B also has a "rejection" option, and claims robust accuracy of 90% under ℓ_∞-perturbations bounded by $\epsilon = {}^{8}/_{255}$ (we say that defense B is robust for some example if it classifies that example correctly, and either rejects/detects or correctly classifies all perturbed examples at distance ϵ.)

Which of these two (empirical) claims are you more likely to believe to be correct?

Defense A claims much higher robustness than the current best results achieved with adversarial training (Madry et al., 2018; Rebuffi et al., 2021), the only empirical defense against adversarial examples that has stood the test of time. Indeed, the state-of-the-art ℓ_∞ robustness for $\epsilon = {}^{4}/_{255}$ on CIFAR-10 (without external data) is $\approx 79\%$ (Rebuffi et al., 2021). Thus, the claim of defense A is likely to be met with initial skepticism and heightened scrutiny, as could be expected for such a claimed breakthrough result.

The claim of defense B is harder to assess, due to a lack of long-standing baselines for robust detectors (many detection defenses have been shown to be broken (Carlini & Wagner, 2017; Tramèr et al., 2020)). On one hand, detection of adversarial examples has largely been considered to be an easier task than classification (Xu et al., 2018; Pang et al., 2021; Sheikholeslami et al., 2021). On the other hand, defense B claims robustness to perturbations that are twice as large as defense A ($\epsilon = {}^{8}/_{255}$ vs. $\epsilon = {}^{4}/_{255}$).

In this paper, we show that the claims of defenses A and B are, in fact, equivalent! (up to computational efficiency.)

[1] Google Research [2] Work done while the author was at Stanford University. Correspondence to: Florian Tramèr <tramer@cs.stanford.edu>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

We prove a general *hardness reduction* between classification and detection of adversarial examples. Given a detector defense that achieves robust risk α for attacks at distance ϵ (under any metric), we show how to build an *explicit but inefficient* classifier that achieves robust risk α for classifying attacks at distance $\epsilon/2$. The reverse implication also holds: a classifier robust at distance $\epsilon/2$ implies an explicit but inefficient robust detector at distance ϵ. Thus, we prove that robust classification and detection are *equivalent*, up to computational complexity and a factor 2 in the robustness radius.

The reader might argue that a computationally inefficient reduction may be of limited use. That is, maybe we should not be surprised by the possibility of a breakthrough in robust classification that leverages inefficient methods. Yet, we do not yet know of any way of leveraging computational *inefficiency* to build more robust models. To the contrary, there is growing evidence that the robustness of classifiers is limited by the availability of *data* (Schmidt et al., 2018; Carmon et al., 2019; Uesato et al., 2019; Zhai et al., 2019) rather than by *compute*. Since our reduction does preserve the *sample complexity* of the original detector model, we thus argue that—given our current knowledge—we should be as "surprised" by the claim made by defense B above as by the claim made by defense A.

Our reduction thus provides a way of assessing the "plausibility" of new robust detection claims, by contrasting them with results from the more mature literature on robust classification. To illustrate, we revisit 14 published detection defenses across three datasets, and show that in 12/14 cases the defense's robust detection claims actually imply a classifier with robustness far superior to the current state-of-the-art. Yet, none of these detection papers make the claim that their techniques should imply such a breakthrough in robust *classification*.

This situation is reminiscent of an informal principle laid out by Scott Aaronson to recognize likely-to-be-incorrect mathematical breakthroughs:

> "*The approach seems to yield something much stronger and maybe even false (but the authors never discuss that).*" (Aaronson, 2008)

Using our reduction, it is obvious that many detector defenses are implicitly claiming much stronger robustness than we believe feasible with current techniques. And indeed, subsequent re-evaluations of many of these defenses resulted in complete breaks (Carlini & Wagner, 2017; Tramèr et al., 2020).

Remarkably, we find that for *certified* defenses, the state-of-the-art results for provable robust classification and detection perfectly match the results implied by our reduction. For example, Sheikholeslami et al. (2021) recently proposed a certified detector on CIFAR-10 with provable robust error that is within 3% of the provable error of the *inefficient* detector obtained by combining our result with the state-of-the-art robust classifier of Zhang et al. (2020a).

In summary, we prove that giving classifiers access to a detection option does not help robustness (or at least, not much). Our work provides, to our knowledge, the first example of a hardness reduction between different approaches for robust machine learning. As in the case of computational complexity, we believe that such reductions can be useful for identifying research questions or areas that are unlikely to bear fruit (bar a significant breakthrough)—so that the majority of the community's efforts can be redirected elsewhere.

On a technical level, our reduction exposes a natural connection between robustness and error correcting codes, which may be of independent interest.

2. Hardness Reductions Between Robust Classifiers and Detectors

In this section, we prove our main result: a reduction between robust detectors and robust classifiers. We first introduce some useful notation and define the (robust) risk of classifiers with and without a detection option.

2.1. Preliminaries

We consider a classification task with a distribution $\mathcal{D}$ over examples $x \in \mathbb{R}^n$ with labels $y \in [C]$. A classifier is a function $f_{\texttt{clf}} : \mathbb{R}^n \to [C]$. A detector $f_{\texttt{det}}$ is a classifier with an extra "rejection" or "detection" option $\bot$, that indicates the absence of a classification. We assume for simplicity that classifiers and detectors are deterministic. Our results can easily be extended to randomized functions as well. The binary indicator function $\mathbb{1}_{\{A\}}$ is 1 if and only if the predicate A is true.

We first define a classifier's *risk*, i.e., its classification error on unperturbed samples.

Definition 1 (Risk). *Let f be either a classifier $f_{\texttt{clf}} : \mathbb{R}^n \to [C]$ or a classifier with detection $f_{\texttt{det}} : \mathbb{R}^n \to [C] \cup \{\bot\}$. The risk of f is the expected rate at which f fails to correctly classify a sample:*

$$R(f) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\mathbb{1}_{\{f(x)\neq y\}}\right] \tag{1}$$

Note that for a detector, rejecting an unperturbed example sampled from the distribution $\mathcal{D}$ is counted as an error.

For classifiers without a rejection option, we define the *robust risk* as the risk on worst-case adversarial inputs (Madry et al., 2018). For an input x sampled from

$\mathcal{D}$, an adversarial example $\hat{x}$ is constrained to being within distance $d(x, \hat{x}) \leq \epsilon$ from x, where d is some metric.

Definition 2 (Robust risk)**.** *Let $f_{\texttt{clf}} : \mathbb{R}^n \to [C]$ be a classifier. The robust risk at distance ϵ is:*

$$R^{\epsilon}_{adv}(f_{\texttt{clf}}) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\max_{d(x,\hat{x})\leq\epsilon} \mathbb{1}_{\{f_{\texttt{clf}}(\hat{x})\neq y\}} \right] \tag{2}$$

Thus, a sample (x, y) is robustly classified if and only if every point within distance ϵ of x (including x itself) is correctly classified as y.

For a detector (a classifier with an extra detection/rejection output), we similarly define the robust risk with detection. The classifier is now allowed to reject adversarial inputs.

Definition 3 (Robust risk with detection)**.** *Let $f_{\texttt{det}} : \mathbb{R}^n \to [C] \cup \{\perp\}$ be a classifier with an extra detection output $\perp$. The robust risk with detection at distance ϵ is:*

$$R^{\epsilon}_{adv\text{-}det}(f_{\texttt{det}}) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\max_{d(x,\hat{x})\leq\epsilon} \mathbb{1}_{\{f_{\texttt{det}}(x)\neq y \,\vee\, f_{\texttt{det}}(\hat{x})\notin\{y,\perp\}\}} \right] \tag{3}$$

That is, a detector defense $f_{\texttt{det}}$ is robust on a sample x from the distribution if and only if the defense classifies the clean input x correctly, and the defense either rejects or correctly classifies every perturbed input $\hat{x}$ within distance ϵ from x. The requirement that the defense correctly classify unperturbed examples eliminates pathological defenses that reject all inputs.

2.2. Robust Detection Implies Inefficient Robust Classification

We are now ready to introduce our main result, a reduction from a robust detector for adversarial examples at distance ϵ, to an inefficient robust classifier at distance $\epsilon/2$. We later prove that this reduction also holds in the reverse direction, thereby demonstrating the equivalence between robust detection and classification—up to computational hardness.

Theorem 4 (ϵ-robust detection implies inefficient $\epsilon/2$-robust classification)**.** *Let $d(\cdot, \cdot)$ be an arbitrary metric. Let $f_{\texttt{det}}$ be a detector that achieves risk $R(f_{\texttt{det}}) = \alpha$, and robust risk with detection $R^{\epsilon}_{adv\text{-}det}(f_{\texttt{det}}) = \beta$. Then, we can construct an explicit (but inefficient) classifier $f_{\texttt{clf}}$ that achieves risk $R(f_{\texttt{clf}}) \leq \alpha$ and robust risk $R^{\epsilon/2}_{adv}(f_{\texttt{clf}}) \leq \beta$.*

The classifier $f_{\texttt{clf}}$ is constructed as follows on a (possibly perturbed) input $\hat{x}$:

- *Run the detector model $y \leftarrow f_{\texttt{det}}(\hat{x})$. If the input is not rejected, i.e., $y \neq \perp$, then output the label y that was predicted by the detector.*

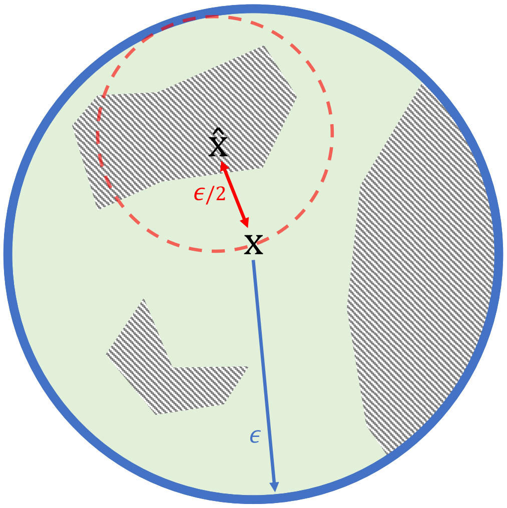

Figure 1. Illustration of the construction of a robust classifier from a robust detector in Theorem 4. The outer blue circle represents all inputs at distance ϵ from the input x. For a detector $f_{\texttt{det}}$, the areas in green correspond to correctly classified inputs, and hatched gray areas correspond to rejected inputs. The detector $f_{\texttt{det}}$ is thus robust on x up to distance ϵ. The classifier $f_{\texttt{clf}}$ classifies a perturbed input $\hat{x}$, at distance $\epsilon/2$ from x, by finding any input within distance $\epsilon/2$ from $\hat{x}$ (the red dashed circle) that is not rejected by $f_{\texttt{det}}$. Such an input necessarily exists and is correctly labeled by $f_{\texttt{det}}$. The classifier $f_{\texttt{clf}}$ is thus robust on x up to distance $\epsilon/2$.

- *Otherwise, find an input x' within distance $\epsilon/2$ of $\hat{x}$ that is not rejected, i.e., $d(\hat{x}, x') \leq \epsilon/2$ and $f_{\texttt{det}}(x') \neq \perp$. If such an input x' exists, output the label $y \leftarrow f_{\texttt{det}}(x')$. Else, output a uniformly random label $y \in [C]$.*

An intuitive illustration for our construction, and for the proof of the theorem (see below) is in Figure 1.

Our construction can be viewed as an analog of *minimum distance decoding* in coding theory. We can view a clean data point x sampled from $\mathcal{D}$ as a codeword, and an adversarial example $\hat{x}$ as a noisy message with a certain number of errors (where the error magnitude is measured using an arbitrary metric on $\mathbb{R}^n$ rather than the Hamming distance that is typically used for error correcting codes). A standard result in coding theory states that if a code can *detect* α errors, then it can *correct* $\alpha/2$ errors. This result follows from a "ball-packing" argument: if α errors can be detected, then any two valid codewords must be at least at distance α from each other, and therefore $\alpha/2$ errors can be corrected via minimum distance decoding.

Proof of Theorem 4. First, note that the natural accuracy of our constructed classifier $f_{\texttt{clf}}$ is at least as high as that of the detector $f_{\texttt{det}}$, since $f_{\texttt{clf}}$ mimics the output of $f_{\texttt{det}}$

whenever $f_{\texttt{det}}$ does not reject an input sampled from $\mathcal{D}$. Thus, $R(f_{\texttt{clf}}) \leq R(f_{\texttt{det}}) = \alpha$.

Now, for the sake of contradiction, consider an input $(x, y) \sim \mathcal{D}$ for which the constructed classifier $f_{\texttt{clf}}$ is not robust at distance $\epsilon/2$. By construction, this means that there exists some input $\hat{x}$ at distance $\epsilon/2$ from x such that $\hat{x}$ is misclassified, i.e., $f_{\texttt{clf}}(\hat{x}) = \hat{y} \neq y$. We will show that the detector $f_{\texttt{det}}$ is not robust with detection for x either (for attacks at distance up to ϵ).

By definition of the classifier $f_{\texttt{clf}}$, if $f_{\texttt{clf}}(\hat{x}) = \hat{y} \neq y$ then either:

- The detector $f_{\texttt{det}}$ also misclassifies $\hat{x}$, i.e., $f_{\texttt{det}}(\hat{x}) = \hat{y}$. So $f_{\texttt{det}}$ is not robust with detection for x at distance ϵ.
- There exists an input x' within distance $\epsilon/2$ of $\hat{x}$, such that the detector $f_{\texttt{det}}$ misclassifies x', i.e. $f_{\texttt{det}}(x') = \hat{y}$.

  Note that by the triangular inequality, $d(x, x') \leq d(x, \hat{x}) + d(\hat{x}, x') \leq \epsilon/2 + \epsilon/2 = \epsilon$, and thus $f_{\texttt{det}}$ is not robust with detection for x at distance ϵ.
- The detector $f_{\texttt{det}}$ rejects all inputs x' within distance $\epsilon/2$ of $\hat{x}$ (and thus $f_{\texttt{clf}}$ sampled an output label at random).

  Since $d(x, \hat{x}) \leq \epsilon/2$, this implies that the detector also rejects the clean input x, i.e., $f_{\texttt{det}}(x) = \bot$, and thus $f_{\texttt{det}}$ is not robust with detection for x.

In summary, whenever the constructed classifier $f_{\texttt{clf}}$ fails to robustly classify an input x up to distance $\epsilon/2$, the detector $f_{\texttt{det}}$ also fails to robustly classify x with detection up to distance ϵ. Taking expectations over the entire distribution $\mathcal{D}$ concludes the proof. □

The classifier $f_{\texttt{clf}}$ constructed in Theorem 4 is computationally inefficient. Indeed, the second step of the defense consists in finding a non-rejected input within some metric ball. If the original detector $f_{\texttt{det}}$ is a non-convex function (e.g., a deep neural network), then this step consists in solving an intractable non-convex optimization problem. Our reduction is thus typically not suitable for building a practical robust classifier. Instead, it demonstrates the existence of an inefficient but *explicitly instantiable* robust classifier. We discuss the implications of this result more thoroughly in Section 3.

A corollary to our reduction is that many "information theoretic" results about robust classifiers can be directly extended to robust detectors. For example, Tsipras et al. (2019) prove that there exists a formal tradeoff between a classifier's clean accuracy and robust accuracy for certain natural tasks. Since their result applies to *any* classifier (including inefficient ones), combining their result with our reduction implies that a similar accuracy-robustness tradeoff exists for detectors. More precisely, Tsipras et al. (2019) show that for certain classification tasks and suitable choices of parameters α, β, ϵ, any classifier $f_{\texttt{clf}}$ which achieves risk $R(f_{\texttt{clf}}) \leq \alpha$ must have robust risk at least $R_{\text{adv}}^{\epsilon}(f_{\texttt{clf}}) \geq \beta$ against ℓ_∞-perturbations bounded by ϵ. By our reduction, this implies that any detector $f_{\texttt{det}}$ with risk at most $R(f_{\texttt{det}}) \leq \alpha$ must also have robust risk with detection at least $R_{\text{adv-det}}^{\epsilon/2}(f_{\texttt{det}}) \geq \beta$ against ℓ_∞-perturbations bounded by $\epsilon/2$.

Similar arguments can be applied to show, for instance, that the increased sample complexity of robust generalization from Schmidt et al. (2018), or the tradeoff between robustness to multiple perturbation types from Tramèr & Boneh (2019), also apply to robust detectors.

Our reduction does not apply to "computational" hardness results that have been shown for robust classification. For example, Garg et al. (2020) and Bubeck et al. (2018) show ("unnatural") distributions where learning a robust classifier is computationally hard—under standard cryptographic assumptions. We cannot use Theorem 4 to conclude that learning a robust *detector* is hard for these distributions, since the existence of such a detector would only imply an inefficient robust classifier which does not contradict the results of Garg et al. (2020) or Bubeck et al. (2018).

2.3. Robust Classification Implies Inefficient Robust Detection

A similar argument to the proof of Theorem 4 can be used in the opposite direction, to show that a robust classifier at distance $\epsilon/2$ implies an inefficient robust detector at distance ϵ. Taken together, Theorem 4 and Theorem 5 show that robust detection and classification are *equivalent*, up to a factor 2 in the norm bound and up to computational constraints.

Theorem 5 ($\epsilon/2$ robust-classification implies inefficient ϵ-robust detection). *Let $d(\cdot, \cdot)$ be an arbitrary metric. Let $f_{\texttt{clf}}$ be a classifier that achieves robust risk $R_{\textit{adv}}^{\epsilon/2}(f_{\texttt{clf}}) = \beta$. Then, we can construct an explicit (but inefficient) detector $f_{\texttt{det}}$ that achieves risk $R(f_{\texttt{det}}) \leq \beta$ and robust risk with detection $R_{\textit{adv-det}}^{\epsilon}(f_{\texttt{det}}) \leq \beta$.*

The detector $f_{\texttt{det}}$ is constructed as follows on a (possibly perturbed) input $\hat{x}$:

- *Run the classifier $y \leftarrow f_{\texttt{clf}}(\hat{x})$.*
- *Find a perturbed input x' withing distance $\epsilon/2$ of $\hat{x}$ that is classified differently, i.e., $d(\hat{x}, x') \leq \epsilon/2$ and $f_{\texttt{clf}}(x') \neq y$. If such an input x' exists, reject the input and output $\bot$. Else, output the class y.*

Proof of Theorem 5. Note that for any input (x, y) for which the classifier $f_{\texttt{clf}}$ is robust at distance $\epsilon/2$, no input x' above exists and so $f_{\texttt{det}}(x) = y$. Thus, the risk of $f_{\texttt{det}}$ is at most the robust risk of $f_{\texttt{clf}}$, so $R(f_{\texttt{det}}) \leq \beta$.

Now, consider an input $(x, y) \sim \mathcal{D}$ for which $f_{\texttt{det}}$ is not robust with detection at distance ϵ. That is, either $f_{\texttt{det}}(x) \neq y$, or there exists an input $\hat{x}$ at distance $d(x, \hat{x}) \leq \epsilon$ such that $f_{\texttt{det}}(\hat{x}) = \hat{y} \notin \{y, \perp\}$. We will show that the classifier $f_{\texttt{clf}}$ is not robust for x either (for attacks at distance up to $\epsilon/2$.)

If $f_{\texttt{det}}(x) \neq y$, then by the same argument as above it cannot be the case that the classifier $f_{\texttt{clf}}$ is robust at distance $\epsilon/2$ for x.

So let us consider the case where $f_{\texttt{det}}(\hat{x}) = \hat{y} \notin \{y, \perp\}$. By the definition of $f_{\texttt{det}}$, this means that for all x' at distance at most $\epsilon/2$ from $\hat{x}$, we have $f_{\texttt{clf}}(x') = \hat{y}$. But, note that there exists a point x^* that is at distance at most $\epsilon/2$ from both $\hat{x}$ and x. Since we must have $f_{\texttt{clf}}(x^*) = \hat{y}$, we conclude that $f_{\texttt{clf}}$ is not robust at distance $\epsilon/2$ for x.

Taking expectations over the distribution $\mathcal{D}$ concludes the proof. □

A main distinction between Theorem 4 and Theorem 5 is that the construction in Theorem 4 preserves clean accuracy, but the construction in Theorem 5 does not. The constructed robust detector in Theorem 5 has *clean* accuracy that is equal to the robust classifier's *robust* accuracy.

The construction in Theorem 5 can be efficiently (but approximately) instantiated by a *certifiably robust* classifier (Wong & Kolter, 2018; Raghunathan et al., 2018). These defenses can certify that a classifier's output is constant for all points within some distance ϵ of the input. For an adversarial example $\hat{x}$ for $f_{\texttt{clf}}$, the certification always fails and thus the constructed detector $f_{\texttt{det}}$ will reject $\hat{x}$. If $f_{\texttt{clf}}$ is robust and the certification succeeds, the detector $f_{\texttt{det}}$ copies the output of $f_{\texttt{clf}}$. However, a certified defense may fail to certify a robust input (a false negative), and thus the detector $f_{\texttt{det}}$ may reject more inputs than with the construction in Theorem 5. This reduction from a certified classifier to a detector is implicit in (Wong & Kolter, 2018, Section 3.1).

2.4. On Computational Complexity and Sample Complexity

In statistical learning theory, it is customary to consider two types of *complexity* for a learning algorithm: its *sample complexity*, and its *computational complexity*.

The sample complexity of a learning algorithm corresponds to the number of training samples that are necessary for attaining a given test error. In turn, the computational complexity represents the runtime of the learning algorithm.

Many classical results in statistical learning theory—such as standard generalization bounds—are only concerned with the sample complexity of a learning problem, and make no explicit assumptions on a learner's computational complexity.

Note that our reductions in Theorem 4 and Theorem 5 preserve the sample complexity of the original detector or classifier that the reduction is applied to. Indeed, the reduction uses a pre-trained detector/classifier as a black-box and only modifies the algorithm's inference phase. Thus, from a sample complexity perspective, our results show that robust classification and detection are equally hard (up to a factor 2 in the perturbation bound).

3. What Are Detection Defenses Claiming?

We now survey 14 detection defenses, and consider the robust *classification* performance that these defenses implicitly claim (via Theorem 4). As we will see, in 12/14 cases, the defenses' detection results imply a computationally inefficient classifier with far better robust accuracy than the state-of-the-art.

Before presenting our experimental setup and the explicit results from the reduction, we first discuss how we believe these results should be interpreted.

3.1. Interpreting our Reduction

Suppose that some detector defense claims a robust accuracy that implies—via our reduction—an inefficient classifier with much higher robustness than the state-of-the-art (e.g., the defense A described in the introduction of this paper).

A first possible interpretation of our reduction is that this robust detector implies the *existence* of a robust classifier. This interpretation is rather weak however, since it is typically presumed that robust classification is possible, and that human perception gives an existence proof of a robust classifier. The mere existence of a robust classifier is thus typically already assumed to be true.

*Our reduction yields a stronger result. It gives an **explicit construction** of a sample efficient but computationally inefficient robust classifier from a robust detector.*

The question then is whether we should expect computationally inefficient robust classifiers to be easier to construct than efficient ones (at a fixed sample complexity). That is, do we expect that we can leverage computational inefficiency to build more robust classifiers than the current state-of-the-art, without having to collect additional training data?

We do not know of an answer to this question, but there is evidence to suggest that the answer may be negative.[1] For example, the work of Schmidt et al. (2018) proves that for a synthetic classification task between Gaussian distributions, building more robust classifiers requires additional *data*—regardless of the amount of computation power. Their results are corroborated by current state-of-the-art robust classifiers based on adversarial training (Madry et al., 2018), which do not appear to be limited by computational constraints. On CIFAR-10 for example, adversarial training achieves 100% robust *training* accuracy (Schmidt et al., 2018). Thus, it is unclear how computational inefficiency could be leveraged to build more robust classifiers using existing techniques.

Candidate approaches could be to train much larger models (e.g., with an exponential number of parameters), or to perform an exhaustive architecture search to find more robust models. Yet, note that the robust classifier constructed in our reduction only uses its unbounded computational power at *inference time*. That is, the classifier that is built in Theorem 4 uses a pre-trained detector model as a black-box subroutine (which is presumed to be efficient), and then solves a non-convex optimization problem at inference time. The classifier built in our reduction is thus presumably weaker than a robust classifier that can be *trained* with unbounded computational power.

To summarize, **when a detector defense claims a high robust accuracy, this implies the existence of a *concretely instantiatable, sample efficient robust classifier with an inefficient inference procedure* with the same robust accuracy.**

If this inefficient classifier is much more robust than the current state-of-the-art, it thus means that either: (1) the detector defense inadvertently made a huge breakthrough in robust *classification*; or (2) the detector defense's claims are incorrect...

Our results do not imply that the latter option is necessarily the right one. But given how challenging robust classification is proving to be, we have reason to be skeptical of such a major breakthrough (even for inefficient classifiers). To compound this, many proposed detection defenses are quite *simple*,[2] and reject adversarial inputs based on some standard statistical test over a neural network's features. It would thus be particularly surprising if such simple techniques could yield robust *classifiers*, given that most "simple" approaches to adversarial robustness (denoising, compression, randomness, etc.) are known to be ineffective (He et al., 2017). It is thus not too surprising that a number of the detector defenses that we survey hereafter have later been broken by stronger attacks (Carlini & Wagner, 2017; Tramèr et al., 2020). Our reduction would have suggested that such a break was likely to happen.

Of course, one could always take the optimistic interpretation that the detector defense's claims *are correct*, and that the defense does in fact imply a huge breakthrough in robust classification. We then argue that detector defenses which claim such high robustness should also explicitly claim a breakthrough in robust classification (since our theorem proves that these claims are equivalent). In turn, these defenses should then be met with the scientific skepticism and scrutiny that such a breakthrough claim naturally entails.

3.2. Experimental Setup

We choose 14 detector defenses from the literature (see Table 1). Our selection of these defenses was partially motivated by a pragmatic consideration on the easiness of translating the defenses' claims into a bound on the robust risk with detection $R^{\epsilon}_{\text{adv-det}}$. Indeed, some defenses simply report a single AUC score for the detector's performance, from which we cannot derive a useful bound on the robust risk. We thus focus on defenses that either directly report a robust error akin to Definition 3, or that provide concrete pairs of false-positive and false-negative rates (e.g., a full ROC curve). In the latter case, we compute a "best-effort" bound on the robust risk with detection[3] as:

$$R^{\epsilon}_{\text{adv-det}}(f_{\texttt{det}}) \leq \text{FPR} + \text{FNR} + R(f_{\texttt{det}}) \ , \tag{4}$$

where FPR and FNR are the detector's false-positive and false-negative rates for a fixed detection threshold, and $R(f_{\texttt{det}})$ is the defense's standard risk (i.e., the test error on natural examples).

The above union bound in Equation (4) is quite pessimistic, as we may over-count examples that lead to multiple sources of errors (e.g., a natural input that is misclassified and erroneously detected). The true robustness claim made by these detector defenses might thus be stronger than what we obtain from our bound (note that this only strengthens our claims). We encourage future detection papers to report their adversarial risk with detection, $R^{\epsilon}_{\text{adv-det}}$, to facilitate direct comparisons with robust classifiers using our reduction.

[1] Some works show that for certain "unnatural" distributions, computational inefficiency is necessary to build robust classifiers (Garg et al., 2020; Bubeck et al., 2018). Yet, since we presume that the human perceptual system *is* robust to small perturbations on natural data (e.g., such as CIFAR-10), there must exist some efficient natural process to achieve robustness on such data.

[2] This simplicity is sometimes justified with the assumption that detecting adversarial examples is a much easier problem than classifying them. As we show, this assumption is flawed.

[3] Many detector defenses report performance against a set of *fixed* (non-adaptive) attacks. We interpret these results as being an approximation of the worst-case risk.

Table 1. For each detector defense, we compute a (best-effort) bound on the claimed robust risk with detection $R^{\epsilon}_{\text{adv-det}}$ using Equation (4), and report the complement (the robust accuracy with detection), $1 - R^{\epsilon}_{\text{adv-det}}$. For each detector's robustness claim (at distance ϵ), we report the state-of-the-art robust classification accuracy for attacks at distance $\epsilon/2$, denoted $1 - R^{\epsilon/2}_{\text{adv}}$. Detection defense claims that imply a higher robust classification accuracy than the current state-of-the-art are highlighted in red.

Dataset	**Defense**	**Norm**	ϵ	$1 - R^{\epsilon}_{\text{adv-det}}$	$1 - R^{\epsilon/2}_{\text{adv}}$
MNIST	Grosse et al. (2017)	ℓ_∞	0.5	$\geq 98\%$	94%
	Ma et al. (2018)	ℓ_2	4.2	$\geq 99\%$	72%
	Raghuram et al. (2021)	ℓ_2	8.9	$\geq 74\%$	0%
CIFAR-10	Yin et al. (2020)	ℓ_2	1.7	$\geq 90\%$	66%
	Feinman et al. (2017)	ℓ_2	2.7	$\geq 43\%$	36%
	Miller et al. (2019)	ℓ_2	2.9	$\geq 75\%$	30%
	Raghuram et al. (2021)	ℓ_2	4.0	$\geq 56\%$	10%
	Ma & Liu (2019)	ℓ_∞	$4/255$	$\geq 96\%$	85%
	Roth et al. (2019)	ℓ_∞	$8/255$	$\geq 66\%$	79%
	Lee et al. (2018)	ℓ_∞	$20/255$	$\geq 81\%$	59%
	Li et al. (2019)	ℓ_∞	$26/255$	$\geq 80\%$	44%
ImageNet	Xu et al. (2018)	ℓ_2	1.0	$\geq 67\%$	54%
	Ma & Liu (2019)	ℓ_∞	$2/255$	$\geq 68\%$	55%
	Jha et al. (2019)	ℓ_∞	$2/255$	$\geq 30\%$	55%
	Hendrycks & Gimpel (2017)	ℓ_∞	$10/255$	$\geq 76\%$	30%
	Yu et al. (2019)	ℓ_∞	$26/255$	$\geq 7\%$	5%

The 14 detector defenses use three datasets: MNIST, CIFAR-10 and ImageNet, and consider adversarial examples under the ℓ_∞ or ℓ_2 norms. Given a claim of robust detection at distance ϵ, we contrast it to a state-of-the-art robust classification result for distance $\epsilon/2$:

- On MNIST with ℓ_∞ attacks, we use the adversarially-trained TRADES classifier (Zhang et al., 2019) and measure robust error with the Square attack (Andriushchenko et al., 2020).
- On MNIST with ℓ_2 attacks, we use the adversarially-trained classifier from Tramèr & Boneh (2019) and measure robust error with PGD (Madry et al., 2018).
- On CIFAR-10, for both ℓ_∞ and ℓ_2 attacks we use the adversarially-trained classifier of Rebuffi et al. (2021) (trained without external data), and attack it using the APGD-CE attack from AutoAttack (Croce & Hein, 2020).
- For ImageNet, for both ℓ_∞ and ℓ_2 attacks we use adversarially-trained classifiers and PGD attacks from Engstrom et al. (2019).

We also consider two *certified* defenses for ℓ_∞ attacks on CIFAR-10: the robust classifier of Zhang et al. (2020a), and a recent certified detector of Sheikholeslami et al. (2021).

3.3. Results

As we can see from Table 1, most defenses claim a detection performance that implies a far greater robust accuracy than our current best robust classifiers. To illustrate with a concrete example, the CIFAR-10 detector of Miller et al. (2019) claims to achieve robust accuracy with detection of 75% for ℓ_2 attacks with $\epsilon = 2.9$. Using Theorem 4, this implies an inefficient classifier with robust accuracy of 75% for ℓ_2 attacks with $\epsilon = 2.9/2 = 1.45$. Yet, the current state-of-the-art robust accuracy for such a perturbation budget is only 30% (Rebuffi et al., 2021). If this detector defense's robustness claim were correct, it would imply a remarkable breakthrough in robust classification!

Why do many of these defenses claim robust accuracies that appear far beyond the current state-of-the-art? A primary reason is that the vast majority of the above detector defenses do not consider evaluations against *adaptive attacks* (Carlini et al., 2019; Athalye et al., 2018; Tramèr et al., 2020). That is, these defenses show that they can detect *some fixed attacks*, and thereafter conclude that the detector is robust against *all attacks*. As in the case of robust classifiers, such an evaluation is clearly insufficient. Some defenses do evaluate against adaptive adversaries, but likely fail to build a sufficiently strong attack to reliably approximate the worst-case robust risk. Because of the lack of a strong comparative baseline, it is not always immediately clear that these results are overly strong.

For example, the recent work of Raghuram et al. (2021, ICML Long Talk) builds a detector on MNIST with a FNR of $\leq 5\%$ at a FPR of $\leq 20\%$, for *adaptive* ℓ_2 attacks bounded by $\epsilon = 8.9$. Yet, this perturbation bound is much larger than the average distance between an MNIST image and the nearest image from a different class! Thus, an attack within this perturbation bound can trivially reduce the detector's accuracy to chance. On CIFAR-10, the same detector achieves 95% clean accuracy, and a FNR of $\leq 19\%$ at a FPR of $\leq 20\%$ for *adaptive* ℓ_2 attacks bounded by $\epsilon = 4$. Using Equation (4), this yields a bound on the robust accuracy with detection of $1 - R^{\epsilon}_{\text{adv-det}}(f_{\texttt{det}}) \geq 1-(5\%+19\%+20\%) = 56\%$. In contrast, the best robust classifier we are aware of for ℓ_2 attacks bounded by $\epsilon = 2$ achieves robust accuracy of only 10% (Rebuffi et al., 2021). In summary, the adaptive attack considered in this detector defense's evaluation is highly unlikely to be a good approximation of a worst-case attack, and this defense can probably be broken by stronger attacks (or if not, the defense should deservedly claim to have made a huge breakthrough in robust classification!).

Certified classifiers and detectors. In Table 2, we look at the robust accuracy with detection and standard robust accuracy achieved by *certified* defenses (for which the claimed robustness numbers are necessarily mathematically correct).

We note that our reduction is not as meaningful in the case of certified defenses, since it is highly plausible that computational inefficiency *can* be leveraged to build better certified classifiers. Indeed, given any robust classifier (e.g., an adversarially trained model), the classifier's robustness can always be certified inefficiently (by enumerating over all points within an ϵ-ball). Thus, the existence of an inefficient classifier with higher *certified* robustness than the state-of-the-art would not be particularly surprising.

Nevertheless, we find that existing results for certified classifiers and detectors perfectly match what is implied by our reduction (up to $\pm 2\%$ error). For example, Zhang et al. (2020a) achieve 39% certified robust accuracy on CIFAR-10 for perturbations of ℓ_∞-norm below $^{4}/_{255}$. Together with Theorem 5, this implies an inefficient detector with 39% robust detection accuracy for perturbations of ℓ_∞-norm below $^{8}/_{255}$. The recent work of Sheikholeslami et al. (2021) nearly matches that bound (37% robust accuracy with detection), with a defense that has the advantage of being concretely efficient.

These results give additional credence to our thesis: with current techniques, robust classification is indeed approximately twice as hard (in terms of the perturbation bounds covered) than robust detection.

Table 2. Certified robust accuracy $1 - R^{\epsilon/2}_{\text{adv}}$ for the defense of Zhang et al. (2020a), and certified robust accuracy with detection $1 - R^{\epsilon}_{\text{adv-det}}$ for the defense of Sheikholeslami et al. (2021).

$\boldsymbol{\epsilon}$	$\mathbf{1 - R^{\epsilon}_{\text{adv-det}}}$	$\mathbf{1 - R^{\epsilon/2}_{\text{adv}}}$
$^{8}/_{255}$	37%	39%
$^{16}/_{255}$	32%	33%

4. Extensions and Open Problems

The main open problem raised by our work is whether it is possible to show an *efficient* reduction between classification and detection of adversarial examples, but this seems implausible—at least with our minimum distance decoding approach.

Another interesting question is whether a similar reduction can be shown for robustness to less structured perturbations than ℓ_p balls and other metric spaces. For example, there has been a line of research on defending against *adversarial patches* (Brown et al., 2017), using empirical (Hayes, 2018; Naseer et al., 2019; Chou et al., 2020) and certifiable techniques (Chiang et al., 2020; Zhang et al., 2020b; Xiang et al., 2021). To use our result, we would have to define some metric to measure the size of an adversarial patch's perturbation. Yet, the size of a patch is typically defined by the number of contiguously perturbed pixels, which does not define a metric—in particular, it does not satisfy the triangular inequality which is necessary in our proofs.

Finally, similar hardness reductions as ours might exist between other approaches for building robust classifiers. For example, the question of whether (test-time) randomness can be leveraged to build more robust models is also intriguing. Empirical defenses that use randomness can be notoriously hard to evaluate (Athalye et al., 2018; Tramèr et al., 2020), so a reduction similar to ours might be useful in showing that we should not expect such approaches to bare fruit.

5. Conclusion

We have shown formal reductions between robust classification with, and without, a detection option. Our results show that significant progress on one of these two tasks implies similar progress on the other—unless computational inefficiency can somehow be leveraged to build more robust models. This raises the question on whether we should spend our efforts on studying both of these tasks, or focus our efforts on a single one.

On one hand, the two tasks represent different ways of tackling a common goal, and working on either task might result in new techniques or ideas that apply to the other task as well. On the other hand, our reductions show that unless we make progress on both tasks, work on one of the tasks can merely aim to match the robustness of our inefficient constructions, whilst improving their computational complexity.

We believe our reduction will serve as a useful sanity-check when assessing the claims of future detector defenses. Detector defenses' robustness evaluations have received less stringent scrutiny than robust classifiers over the past years, perhaps in part due to a lack of strong comparative baselines. Instead of having to wait until some detector defense's claims pass the test-of-time, we show that detection results can be directly contrasted against long-standing results for robust classification.

When applying this approach to past detector defenses, we find that many make robustness claims that imply significant breakthroughs in robust classification. We believe our reduction could have been useful in highlighting the suspiciously strong claims made by many of these defenses—before they were explicitly broken by stronger attacks.

Acknowledgments

Thanks to thank Nicholas Carlini and Wieland Brendel for helpful discussions, as well as to Alex Ozdemir for suggesting the connection between our results and minimum distance decoding.

References

Aaronson, S. Ten signs a claimed mathematical breakthrough is wrong. https://www.scottaaronson.com/blog/?p=304, Jan 2008.

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

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, 2018.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 387–402. Springer, 2013.

Brown, T. B., Mané, D., Roy, A., Abadi, M., and Gilmer, J. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

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

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *AISec*, pp. 3–14. ACM, 2017.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., and Madry, A. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. S. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pp. 11192–11203, 2019.

Chiang, P.-y., Ni, R., Abdelkader, A., Zhu, C., Studer, C., and Goldstein, T. Certified defenses for adversarial patches. *arXiv preprint arXiv:2003.06693*, 2020.

Chou, E., Tramèr, F., and Pellegrino, G. Sentinet: Detecting physical attacks against deep learning systems. In *Workshop on Deep Learning Security*, 2020.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

Croce, F., Andriushchenko, M., Sehwag, V., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Engstrom, L., Ilyas, A., Salman, H., Santurkar, S., and Tsipras, D. Robustness (python library), 2019. URL https://github.com/MadryLab/robustness.

Feinman, R., Curtin, R. R., Shintre, S., and Gardner, A. B. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

Garg, S., Jha, S., Mahloujifar, S., and Mohammad, M. Adversarially robust learning could leverage computational hardness. In *Algorithmic Learning Theory*, pp. 364–385. PMLR, 2020.

Grosse, K., Manoharan, P., Papernot, N., Backes, M., and McDaniel, P. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.

Hayes, J. On visible adversarial perturbations & digital watermarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1597–1604, 2018.

He, W., Wei, J., Chen, X., Carlini, N., and Song, D. Adversarial example defenses: Ensembles of weak defenses are not strong. In *USENIX Workshop on Offensive Technologies*, 2017.

Hendrycks, D. and Gimpel, K. Early methods for detecting adversarial images. In *International Conference on Learning Representations*, 2017.

Jha, S., Raj, S., Fernandes, S. L., Jha, S. K., Jha, S., Verma, G., Jalaian, B., and Swami, A. Attribution-driven causal analysis for detection of adversarial examples. *arXiv preprint arXiv:1903.05821*, 2019.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018.

Li, Y., Bradshaw, J., and Sharma, Y. Are generative classifiers more robust to adversarial attacks? In *International Conference on Machine Learning*, pp. 3804–3814. PMLR, 2019.

Ma, S. and Liu, Y. Nic: Detecting adversarial samples with neural network invariant checking. In *Proceedings of the 26th Network and Distributed System Security Symposium (NDSS 2019)*, 2019.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Schoenebeck, G., Song, D., Houle, M. E., and Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Miller, D., Wang, Y., and Kesidis, G. When not to classify: Anomaly detection of attacks (ada) on dnn classifiers at test time. *Neural computation*, 31(8):1624–1670, 2019.

Naseer, M., Khan, S., and Porikli, F. Local gradients smoothing: Defense against localized adversarial attacks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1300–1307. IEEE, 2019.

Pang, T., Zhang, H., He, D., Dong, Y., Su, H., Chen, W., Zhu, J., and Liu, T.-Y. Adversarial training with rectified rejection. *arXiv preprint arXiv:2105.14785*, 2021.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018.

Raghuram, J., Chandrasekaran, V., Jha, S., and Banerjee, S. A general framework for detecting anomalous inputs to dnn classifiers. In *International Conference on Machine Learning*, 2021.

Rebuffi, S.-A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., and Mann, T. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021.

Roth, K., Kilcher, Y., and Hofmann, T. The odds are odd: A statistical test for detecting adversarial examples. In *International Conference on Machine Learning*, 2019.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *Advances In Neural Information Processing Systems*, pp. 5019–5031, 2018.

Sheikholeslami, F., Lotfi, A., and Kolter, J. Z. Provably robust classification of adversarial examples with detection. In *International Conference on Learning Representations*, 2021.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Tramèr, F. and Boneh, D. Adversarial training and robustness for multiple perturbations. In *Advances In Neural Information Processing Systems*, 2019.

Tramèr, F., Carlini, N., Brendel, W., and Madry, A. On adaptive attacks to adversarial example defenses. In *Conference on Neural Information Processing Systems*, 2020.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

Uesato, J., Alayrac, J.-B., Huang, P.-S., Stanforth, R., Fawzi, A., and Kohli, P. Are labels required for improving adversarial robustness? *arXiv preprint arXiv:1905.13725*, 2019.

Wong, E. and Kolter, Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pp. 5283–5292, 2018.

Xiang, C., Bhagoji, A. N., Sehwag, V., and Mittal, P. Patchguard: A provably robust defense against adversarial patches via small receptive fields and masking. In *30th $\{$USENIX$\}$ Security Symposium ($\{$USENIX$\}$ Security 21)*, 2021.

Xu, W., Evans, D., and Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Network and Distributed System Security Symposium*, 2018.

Yin, X., Kolouri, S., and Rohde, G. K. Gat: Generative adversarial training for adversarial example detection and robust classification. In *International Conference on Learning Representations*, 2020.

Yu, T., Hu, S., Guo, C., Chao, W.-L., and Weinberger, K. Q. A new defense against adversarial images: Turning a weakness into a strength. In *Advances in Neural Information Processing Systems*, 2019.

Zhai, R., Cai, T., He, D., Dan, C., He, K., Hopcroft, J., and Wang, L. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.

Zhang, H., Chen, H., Xiao, C., Li, B., Boning, D., and Hsieh, C.-J. Towards stable and efficient training of verifiably robust neural networks. In *International Conference on Learning Representations*, 2020a.

Zhang, Z., Yuan, B., McCoyd, M., and Wagner, D. Clipped bagnet: Defending against sticker attacks with clipped bag-of-features. In *2020 IEEE Security and Privacy Workshops (SPW)*, pp. 55–61. IEEE, 2020b.