Detection of Iterative Adversarial Attacks via Counter Attack

Matthias Rottmann\textsuperscript{1} · Kira Maag\textsuperscript{2} · Mathis Peyron\textsuperscript{3} · Hanno Gottschalk\textsuperscript{4} · Nataša Krejić\textsuperscript{5}

Received: 14 May 2021 / Accepted: 7 July 2023 / Published online: 4 August 2023
© The Author(s) 2023

Abstract

Deep neural networks (DNNs) have proven to be powerful tools for processing unstructured data. However, for high-dimensional data, like images, they are inherently vulnerable to adversarial attacks. Small almost invisible perturbations added to the input can be used to fool DNNs. Various attacks, hardening methods and detection methods have been introduced in recent years. Notoriously, Carlini–Wagner (CW)-type attacks computed by iterative minimization belong to those that are most difficult to detect. In this work we outline a mathematical proof that the CW attack can be used as a detector itself. That is, under certain assumptions and in the limit of attack iterations this detector provides asymptotically optimal separation of original and attacked images. In numerical experiments, we experimentally validate this statement and furthermore obtain AUROC values up to 99.73\% on CIFAR10 and ImageNet. This is...
in the upper part of the spectrum of current state-of-the-art detection rates for CW attacks.

**Keywords** Deep neural networks · Adversarial attacks · Counter attacks · Asymptotically perfect detection

**Mathematics Subject Classification** 49J52 · 62H35 · 68T10

### 1 Introduction

For many applications, deep learning has shown to outperform concurring machine learning approaches by far [14, 18, 35]. Especially, when working with high-dimensional input data like images, deep neural networks (DNNs) show impressive results. As discovered by [37], this remarkable performance comes with a downside. Very small (noise-like) perturbation added to an input image can result in incorrect predictions with high confidence [10, 37]. Such adversarial attacks are usually crafted by performing a projected gradient descent method for solving a constrained optimization problem. The optimization problem is formulated as the least change of the input image that yields a change in the class predicted by the DNN. The new class can be either a class of choice (targeted attack) or an arbitrary but different one (untargeted attack). Many other types of attacks such as the fast signed gradient method [10] and DeepFool [26] have been introduced, but these methods do not fool DNNs reliably. Carlini & Wagner (CW) extended [37] with a method that reliably attacks deep neural networks while controlling important features of the attack like sparsity and maximum size over all pixels, see [3]. This method aims at finding a targeted or an untargeted attack where the distance between the attacked image and the original one is minimal with respect to a chosen $\ell_p$ distance. Distances of choice within CW-type frameworks are mostly $\ell_p$ with $p = 0, 2, \infty$. Mixtures of $p = 1, 2$ have also been proposed by [5]. Except for the $\ell_0$ distance, these attacks are perceptually extremely hard to detect, cf. Fig. 2. Minimization of the $\ell_0$ distance minimizes the number of pixels changed, but also changes these pixels maximally and the resulting spikes are easy to detect. This changes for $p > 0$. Typically, one distinguishes between three different attack scenarios. **White box** attack: the attacker has access to the DNN’s parameters / the whole framework including defense strategies. **Black box** attack: the attacker does not know the DNN’s parameters / the whole framework including defense strategies. **Gray box** attack: in-between white and black box, e.g., the attacker might know the framework but not the parameters used. The CW attack currently is one of the most efficient white box attacks.

#### 1.1 Defense Methods

Several defense mechanisms have been proposed to either harden neural networks or to detect adversarial attacks. One such hardening method is the so-called *defensive distillation* [28]. With a single re-training step, this method provides strong security,
however not against CW attacks. Training for robustness via adversarial training is another popular defense approach, see, e.g., [10, 19, 24, 25, 39]. See also [30] for an overview regarding defense methods. Most of these methods are not able to deal with attacks based on iterative methods for constrained minimization like CW attacks. In a white box setting, the hardened networks can still be susceptible to fooling.

1.2 Detection Methods

There are numerous detection methods for adversarial attacks. In many works, it has been observed and exploited that adversarial attacks are less robust to random noise than non-attacked clean samples, see, e.g., [2, 31, 38, 42]. This robustness issue can be utilized by detection and defense methods. In [31], a statistical test is proposed for a white box setup. Statistics are obtained under random corruptions of the inputs and observing the corresponding softmax probability distributions. An adaptive noise reduction approach for detection of adversarial examples is presented in [21]. JPEG compression [22, 29], similar input transformations [13] and other filtering techniques [20] have demonstrated to filter out many types of adversarial attacks as well. Some approaches also work with several images or sequences of images to detect adversarial images, see, e.g., [6, 12]. Approaches that not only take input or output layers into account, but also hidden layer statistics, are introduced in [4, 45]. Auxiliary models trained for robustness and equipped for detecting adversarial examples are presented in [44]. Recently, GANs have demonstrated to defend DNNs against adversarial attacks very well, see [32]. This approach called defense-GAN iteratively filters out the adversarial perturbation. The filtered result is then presented to the original classifier. Semantic concepts in image classification tasks pre-select image data that only cover a tiny fraction of the expressive power of rgb images. Therefore, training datasets can be embedded in lower-dimensional manifolds and DNNs are only trained with data within the manifolds and behave arbitrarily in perpendicular directions. These are also the directions used by adversarial perturbations. Thus, the manifold distance of adversarial examples can be used as criterion for detection in [16]. As opposed to many of the adversarial training and hardening approaches, most of the works mentioned in this paragraph are able to detect CW attacks with AUROC values up to 99% [23]. For an overview we refer to [43].

The idea of detecting adversarial attacks by applying another attack (counter attack) to both attacked and non-attacked data (images as well as text), instead of applying random perturbations as previously discussed, was first proposed in [40, 41] (termed adv$^{-1}$) and re-invented by [46] (termed undercover attack). The idea is to measure the perturbation strength of the counter attack in a given $\ell_p$ norm and then to discriminate between $\ell_p$ norms of perturbations for images that were attacked previously and images that are still original (non-attacked). The motivation of this approach is similar to the one for utilizing random noise, i.e., attacked images tend to be less robust to adversarial attacks than original images. Indeed, there is a difference between the methods proposed by [40, 41] and [46]. While the former measure counter attack perturbation norms on the input space, the latter do so on the space of softmax probabilities. All three works are of empirical nature and present results for a range of
Fig. 1 Sketch of the action of the counter attack, $x^*$ is the stationary point of the first attack, $x_k$ the final iterate of the first attack and $x_{k,j}$ the final iterate of the second attack.

attack methods (including FGSM and CW) where they apply the same attack another time (to both attacked and non-attacked data) and also perform cross method attacks, achieving detection accuracies that in many test are state-of-the-art and beyond. As opposed to many other types of detection methods, counter attack methods tend to detect stronger attacks such as the CW attack much more reliably than less strong attacks. Here, strong refers to attacks that fool DNNs reliably while producing very small perturbations.

1.3 Our Contribution

In this paper we establish theory for the CW attack and the counter attack framework [40, 41] applied to the CW attack. As opposed to other types of attack (not based on constraint optimization), the inherent properties of the CW attack constitute an alternative motivation for the counter attack method that turns out to lead to provable detection:

The CW attack minimizes the perturbation in a chosen norm such that the class prediction changes. Thus, an iterative sequence is supposed to terminate closely behind the decision boundary. This also holds for the counter attack. Therefore, in a statistical sense, the counter attack is supposed to produce much smaller perturbations on already attacked images than on original images, see also Fig. 1. This point of view does not necessarily generalize to other types of attacks that are not based on constrained optimization. With this work, we are the first to mathematically prove the following results for the $\ell_2$ CW attack:

- For specific choice of penalty term, the CW attack converges to a stationary point. If the CW attack is successful (the predicted class of the attacked image is different from the prediction for the original image), the CW attack converges to a stationary point on the class boundary.
– If the counter attack is started at an iteration of the first attack that is sufficiently close to the limit point of the attack sequence, the counter attack asymptotically perfectly separates attacked and non-attacked images by means perturbation strength, i.e., by $\ell_p$ norm of the perturbation produced by the counter attack.

We are therefore able to explain many of the results found in [40, 41, 46]. It is exactly the optimality of the CW attack that leaves a treacherous footprint in the attacked data. Note that we do not claim that the counter attack cannot be bypassed as well; however, our work contributes to understanding the nature of CW attacks and the corresponding counter attack.

We complement the mathematical theory with numerical experiments on 2D examples where the assumptions of our proofs are satisfied and study the relevant parameters such as the number of iterations of the primary attack, initial learning rate of the secondary attack and others. We consecutively step by step relax the assumptions, showing that the detection accuracy in terms of area under the receiver operator curve (AUROC) remains still close to 100%. This is complemented with a study on the dependence of the problem dimension. Lastly, we present numerical results for the CW (counter) attack with default parameters applied to CIFAR10 and ImageNet that extend the studies found in [40, 41, 46]. We again study the dependence on the number of primary attack iterations and furthermore perform cross attacks for different $\ell_p$ norms that maintain high AUROC values of up to 99.73%.

2 Detection by Counter Attack

In this section, we first introduce deep neural networks with ReLU activation functions which are widely used in practice and outline why such networks yield piecewise affine functions if the final network layer is equipped with a linear activation function. Afterwards, we introduce the CW attack as well as its counter attack and provide theoretical insights on their convergence to stationary points. These theoretical insight rely on the geometry induced by piecewise affine functions.

2.1 ReLU Networks as Piecewise Affine Functions

We outline why ReLU networks yield piecewise affine functions if the activation function in the ultimate layer is linear. We proceed similarly to [15] and refer to [1] for further details. A function $h : \mathbb{R}^n \rightarrow \mathbb{R}$ is called piecewise affine if there exists a finite set of polytopes $\{Q_j : j = 1, \ldots, s\}$ (also called linear regions of $h$) such that $h$ is an affine function when restricted to every $Q_j$ and $\bigcup_{j=1}^{s} Q_j = \mathbb{R}^n$. Using piecewise affine activation functions like ReLU or LeakyReLU, feedforward neural networks that are linear in the output layer yield continuous piecewise affine functions [1]. This includes, for example, fully connected and convolutional layers as well as max pooling.

Here we restrict ourselves to a network with a one-dimensional output, the extension to $c$ outputs works component-wise and is straightforward. Since convolutional layers are a special case of fully connected layers, we use fully connected layers in the
following. We denote by \( r : \mathbb{R} \to \mathbb{R} \) with \( r(u) = \max\{0, u\} \) the ReLU activation function and by \( L + 1 \) the number of network layers. For layer \( l, l = 1, \ldots, L + 1 \), the weights matrix and the bias vector are given by \( W^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}} \) and \( b^{(l)} \in \mathbb{R}^{m_l} \), respectively. We initialize \( m_0 = n \) and \( r(h^{(0)}(x)) = x \) for \( x \in \mathbb{R}^n \). The output of every layer is defined by

\[
h^{(l)}(x) = W^{(l)} r(h^{(l-1)}(x)) + b^{(l)}, \quad l = 1, \ldots, L,
\]

and the entire network’s output is obtained by

\[
h^{(L+1)}(x) = W^{(L+1)} r(h^{(L)}(x)) + b^{(L+1)}.
\]

Let be \( U^{(l)} \in \mathbb{R}^{m_l \times m_l} \),

\[
U^{(l)}(x)_{ij} = \begin{cases} 
\text{sign}(h^{(l)}_i(x)) & \text{if } i = j, \\
0 & \text{else}
\end{cases}
\]

and \( V^{(l)} \in \mathbb{R}^{m_l \times m_l} \),

\[
V^{(l)}(x)_{ij} = \begin{cases} 
1 & \text{if } i = j \text{ and } h^{(l)}_i(x) > 0, \\
0 & \text{else}
\end{cases}
\]

two diagonal matrices for \( l = 1, \ldots, L \). Thus, \( h^{(l)}(x) \) can be written as composition of affine functions

\[
h^{(l)}(x) = W^{(l)} V^{(l-1)}(x) (W^{(l-1)} V^{(l-2)}(x) \times \left( \ldots (W^{(1)}(x) + b^{(1)}) \ldots \right) + b^{(l-1)}) + b^{(l)}.
\]

This equation can be simplified to \( h^{(l)}(x) = T^{(l)} x + t^{(l)} \) with \( T^{(l)} \in \mathbb{R}^{m_l \times n} \),

\[
T^{(l)} = W^{(l)} \left( \prod_{j=1}^{l-1} V^{(l-j)}(x) W^{(l-j)} \right),
\]

and \( t^{(l)} \in \mathbb{R}^{m_l} \),

\[
t^{(l)} = b^{(l)} + \sum_{j=1}^{l-1} \left( \prod_{k=1}^{l-j} W^{(l+1-k)} V^{(l-k)}(x) \right) b^{(j)}.
\]
The polytope $Q(x)$ is characterized as an intersection of $M$ half spaces with

$$Q(x) = \bigcap_{l=1}^{L} \bigcap_{k=1}^{m_l} \Omega_{l,k},$$

where the half spaces are given by

$$\Omega_{l,k} = \{ y \in \mathbb{R}^n | U^{(l)}(x)(T_k^{(l)}y + t_k^{(l)}) \geq 0 \}. $$

The affine restriction of $h$ to $Q(x)$ can then be written as

$$h^{(L+1)}(y) \big|_{Q(x)} = T^{(L+1)}y + t^{(L+1)},$$

which is affine on $Q(x)$. Note that the number of such polytopes $Q(x)$ is upper bounded by $2^M$.

### 2.2 The CW Attack

Let $C = \{1, \ldots, c\}$ denote the set of $2 \leq c \in \mathbb{N}$ distinct classes and let $I = [0, 1]$ denote the closed unit interval. An image is an element $x \in I^n$. Let $\varphi : I^n \rightarrow I^c$ be a continuous function given by a ReLU-DNN, almost everywhere differentiable, that maps $x$ to a probability distribution $y = \varphi(x) \in I^c$ with $\sum_{i=1}^{c} y_i = 1$ and $y_i \geq 0$. Note that this includes convolutional layers (among others) as they can be viewed as fully connected layers with weight sharing and sparsity. Furthermore, $\kappa : I^n \rightarrow C$ denotes the map that yields the class index, i.e.,

$$\kappa(x) = \begin{cases} 
  i & \text{if } y_i > y_j \quad \forall j \in C \setminus \{i\}, \\
  0 & \text{else.}
\end{cases}$$

This is a slight modification of the arg max function as $K_i = \{ x \in I^n : \kappa(x) = i \}$ for $i > 0$ gives the set of all $x \in I^n$ predicted by $\varphi$ to be a member of class $i$ and for $i = 0$ we obtain the set of all class boundaries with respect to $\varphi$. Given $x \in \mathbb{R}^n$, the $\ell_p$ “norm” is defined as follows: for $p \in \mathbb{N}$, $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$, for $p = 0$, $\|x\|_0 = |\{x_i > 0\}|$, and for $p = \infty$, $\|x\|_\infty = \max_i |x_i|$.

The corresponding $\ell_p$ distance measure is given by $\text{dist}_p(x, x') = \|x - x'\|_p$, and the $n$-dimensional open $\ell_p$-neighborhood with radius $\varepsilon$ and center point $x_0 \in \mathbb{R}^n$ is $B_p(x_0, \varepsilon) = \{ x \in \mathbb{R}^n : \text{dist}_p(x, x_0) < \varepsilon \}$.

For any image $x_0 \in I^n$, the CW attack introduced in [3] can be formulated as the following optimization problem:

$$\min_{x \in I^n} \text{dist}_p(x_0, x) \quad \text{s.th. } \kappa(x) \neq \kappa(x_0) \quad x \in I^n. $$
Several reformulations of problem (12) are proposed in [3]. The condition $\kappa(x) \neq \kappa(x_0)$ is replaced by a continuous function $f(x)$ specified further in the next section. The reformulation of (12) we adopt here is based on the penalty approach. The objective in (12) is replaced by $\text{dist}_p(x_0, x)^p$ for $p \in \mathbb{N}$ and left unchanged for $p = 0, \infty$. Therefore, the problem we consider from now on is

$$F(x) := \text{dist}_p(x_0, x)^p + A f(x) \rightarrow \min, \quad x \in I^n$$

(13)

for the penalty parameter $A$ large enough. In their experiments, Carlini & Wagner perform a binary search for the constant $A$ such that after the final iteration $\kappa(x) \neq \kappa(x_0)$ is almost always satisfied. For further algorithmic details, we refer to [3]. Two illustrations of attacked images are given in Fig. 2. In particular, for the ImageNet dataset even the perturbations themselves are imperceptible.

2.3 The CW Counter Attack

The goal of the CW attack is to find a minimizer of the distance of $x_0$ to the closest class boundary via $\text{dist}_p(x_0, x) \rightarrow \min$ under the constraint that $x$ lies beyond the boundary. The more eager the attacker minimizes the perturbation $x - x_0$ in a chosen $p$-norm, the more likely it is that the iterative point obtained during the minimization procedure $x_k$ is close to a class boundary. Hence, when estimating the distance of $x_k$ to the closest class boundary by performing another CW attack, taking $x_k$ as the starting point and generating an iterative sequence denoted by $x_{k,j}$, it is to be expected that

$$\text{dist}_p(x_k, x_{k,j}) \ll \text{dist}_p(x_0, x_k),$$

(14)

for $j$ sufficiently large, cf. Fig. 1. This motivates our claim that the CW attack itself is a good detector for CW attacks. A non-attacked image $x_0 \in I^n$ is likely to have a greater distance to the closest class boundary than an $x_k \in I^n$ which has already

Fig. 2 An illustration of attacked images for CIFAR10 (left) and ImageNet2012 (right). Each package of three images contains from left to right: input image, attack and the resulting adversarial image.
been exposed to a CW attack. In practice, it cannot be guaranteed that the CW attack finds a point \( x^* \) which is a minimizer of (13) and hence we cannot guarantee that \( x_k \) (an approximation of \( x^* \)) is close to the class boundary. However, we can show that we can find a stationary point of problem (13) and in the case of a successful attack (\( \kappa(x) \neq \kappa(x_0) \)) \( x^* \) is guaranteed to lie on the decision boundary.

### 2.4 Theoretical Considerations

In this section, we prove what has been announced in Sect. 1.3 and proceed in three steps:

1. Firstly, we prove that the CW attack converges to a stationary point \( x^* \) and that in case of a successful attack (changing the predicted class) \( x^* \) is located on the decision boundary.
2. Secondly, in case the Euclidean distance between the final iterate \( x_k \) and \( x^* \) is at most \( \epsilon \) where the latter is sufficiently small, then the counter attack is not leaving a \( 3\epsilon \) ball around \( x^* \).
3. Lastly, we use the previous result and translate it into a statistical argument proving that we, in the limit of \( k \to \infty \), achieve asymptotically perfect separability of attacked and non attacked images.

#### 2.4.1 Convergence of the CW Attack

Let \( t := \kappa(x_0) \) denote the original class predicted by the neural network \( \varphi \) for the input \( x_0 \). We assume that \( p = 2 \) and from now on denote \( \text{dist} := \text{dist}_2 \) and \( \| \cdot \| := \| \cdot \|_2 \).

Furthermore, we fix a choice for \( f \) which is

\[
 f(x) = \max\{ Z_t(x) - \max_{i \neq t} Z_i(x), 0 \},
\]

where \( Z \) denotes the neural network \( \varphi \), but without the final softmax activation function. Note that \( f \) in (15) is a construction for an untargeted attack. Choosing a penalty term for a targeted attack does not affect the arguments provided in this section, the convergence on the CW attack to a stationary point only requires assumptions on the geometry of \( \text{Im}(f) \) that are fulfilled for both the targeted and the untargeted attacked. Furthermore, let \( F := \{ x \in I^n : \kappa(x) \neq t \} \) be the feasible region of (13), \( \mathcal{N}F := \{ x \in I^n : \kappa(x) = t \} \) be the infeasible region, and \( \partial F = \partial \mathcal{N}F \) denote the class boundary between \( F \) and \( \mathcal{N}F \). Note that, by the definition of (11), both \( F \) and \( \mathcal{N}F \) are open sets relative to \( I^n \). The disjoint union of \( F, \mathcal{N}F \) and \( \partial F \) yields \( I^n \). Furthermore, \( x \in \mathcal{N}F \) implies \( f(x) > 0 \) while \( x \in F \) or \( x \in \partial F \) implies \( f(x) = 0 \), cf. also Fig. 3.

As outlined in Sect. 2.1, for a ReLU-DNN the input space \( I^n \) can be decomposed into a finite number of polytopes \( \{Q_i\}_{i=1}^s \) such that \( Z \) is affine on each \( Q_i \), see [8, 34]. In [8], each of the polytopes \( Q_i \) is expressed by a set of linear constraints, therefore being convex. For a locally Lipschitz function \( g \), the Clarke generalized gradient is
Fig. 3 An example of $f$ for $n = 2$. This example contains three polytopes $Q_i$ intersecting with $\partial F$. When also two polytopes $Q_{i_1}, Q_{i_2}$ intersect with each other in $x \in \partial F$, then $\nabla_C f(x)$ contains two gradients $g_1 \perp \partial F \cap Q_{i_1}, g_2 \perp \partial F \cap Q_{i_2}$. Both of them are also orthogonal to $\partial F \cap Q_{i_1} \cap Q_{i_2}$ defined by

$$\nabla_C g(x) := \text{co}\{ \lim_{k \to \infty} \nabla g(x_k) : x_k \to x \land \exists \nabla g(x_k) \}, \quad (16)$$

where $\text{co}$ stands for the convex hull of a set of vectors. Let $S$ be the stationary set of problem (13),

$$S = \{ x \in \mathbb{R}^n : 0 \in \nabla_C F(x) + N_{I^n}(x) \} \quad (17)$$

where $\nabla_C F(x)$ is the set of all generalized gradients and $N_{I^n}(x) = \{ z \in \mathbb{R}^n : z^T (z' - x) \leq 0 \ \forall z' \in I^n \}$ is the normal cone to the set $I^n$ at point $x$. Note that $N_{I^n}(x) = \{0\}$ for $x \in (0, 1)^n$. In general, computing the set of generalized gradients is not an easy task, but given the special structure of $f$—piece-wise linear—it can be done relatively easily in this case. Namely, the set $\nabla_C f(x)$ is a convex hull of the gradients of linear functions that are active at the point $x$, [33]. Therefore, by [7] (Corollary 2, p. 39), the generalized gradient $G(x)$ of $F(x)$ has the form

$$G(x) = \{ Ag(x) + 2(x - x_0), \ g(x) \in \nabla_C f(x) \}. \quad (18)$$

The projected generalized gradient method [36] is defined as follows. Let $P(x)$ denote the orthogonal projection of $x$ onto $I^n$. Given a non-increasing learning schedule $\{ \alpha_k \}$ such that

$$\lim_{k \to \infty} \alpha_k = 0 \text{ and } \sum_{k=1}^{\infty} \alpha_j = \infty, \quad (19)$$
the iterative sequence is generated as
\[ x_{k+1} = P(x_k - \alpha_k G_k), \quad G_k \in \nabla_C F(x_k) \] (20)
for \( k = 0, 1, \ldots \). In this paper we add the condition \( \sum_{k=1}^{\infty} \alpha_k^2 < \infty \) which strengthens the conditions (19) and does not alter the statements from [36].

**Remark 2.1** We now elaborate on an observation that will be used several times later on. The manifold given by \( \{(f(x), x) : x \in I^n\} \) is a \( n \)-dimensional manifold in \( I^n \times \mathbb{R}_{\geq 0} \). Therein, \( \partial F \) is given by a composition of sections of affine hyperplanes of \( n - 1 \) dimensions as they fulfill the condition \( Z_t(x) - \max_{j \neq t} \{Z_j(x)\} = 0 \), cf. (15). If a polytope \( Q_i \) intersects with \( \partial F \) and \( Q_j \cap \partial F \) is \( n - 1 \)-dimensional, then the gradient \( g \in \nabla_C f(x) \) for \( x \in Q_j \cap \partial F \) (and \( x \notin Q_{i'} \), \( i' \neq i \)) is orthogonal to \( Q_i \cap \partial F \) since the latter is a contour line of \( f \). Similarly, if \( \nabla_C f(x) \) contains multiple gradients \( g_i \) corresponding to different polytopes \( Q_i \), and for each of them \( Q_i \cap \partial F \) is \( n - 1 \)-dimensional, then all gradients are orthogonal to \( \partial F \). Since \( Z \) and therefore \( f \) depend on the weights of the neural network that are learned in a stochastic optimization, it is unlikely that \( Q_i \cap \partial F \) will be less than \( n - 1 \)-dimensional, if it is non-empty. Hence, in what follows we will not consider this case. For brevity, we also refer to this by saying that there are no polytope boundaries at \( \partial F \), which we assume from now on. For an illustration on the geometry of \( f \) in the case of \( n = 2 \), see figure 3.

We now prepare two ingredients for a statement on the convergence of CW attacks. The first one is that all stationary points in \( F \cup \partial F \) are isolated w.r.t. \( F \cup \partial F \). The second is that all sequences that visit \( F \) infinitely many times remain in
\[ D(\delta) := F \cap \partial F \cap \{x \in I^n : \text{dist}(x, \partial F) < \delta\} \] (21)
for a \( \delta > 0 \) and \( \alpha_k \downarrow 0 \) small enough as well as \( A \) large enough.

We now make a few preparatory remarks for the statement on the isolatedness of stationary points. Due to the weights of the neural network and therefore the \( n - 1 \)-dimensional hyperplanes that define polytope boundaries being a result of a gradient descent procedure, it is for instance in the 2D case unlikely that more than two 1D hyperplanes will intersect in the same point. Analogously, the intersection of \( n \) hyperplanes of dimension \( n - 1 \) is a point which is again unlikely to be contained in another \( n - 1 \)-dimensional hyperplane. Hence, we can assume that up to \( s' \leq n \) gradients \( \{g_i\}_{i=1}^{s'} \) are independent without making a strong assumption. We argue similarly for a normal cone \( N_{f^n}(x^*) \) of dimension \( m' \) being spanned by \( m' \leq n \) signed canonical basis vectors \( \{e_i\}_{i=1}^{m'} \) (with a corresponding sign \( \pm 1 \) such that they point outside of \( I^n \) which we, however, do not explicitly state here). They will take a role equivalent to the one of \( \{g_i\}_{i=1}^{s'} \) in the proof of the subsequent theorem. We now fix some notation. For the sake of brevity, let \( g_i \) be the unique gradient \( g(x') \) of \( f \) corresponding to some point \( x' \) contained in the inner of the polytope \( Q_i \), i.e.,
\[ g_i := g(x') \in \nabla_C f(x') \quad \text{for} \quad x' \in Q_i \setminus \partial Q_i, \] (22)
Let us first assume that Case 1: $\lambda_i(x^*) \geq 0, i \in A_Q(x^*)$ and $\mu_j(x^*) \geq 0, j \in A_E(x^*)$ such that

$$A \left( \sum_{i \in A_Q(x^*)} \lambda_i(x^*)g_i + \sum_{j \in A_E(x^*)} \mu_j(x^*)e_j \right) = 2(x_0 - x^*).$$

We introduce active sets with respect to the coefficients $\lambda_i(x^*)$ and $\mu_j(x^*)$, i.e.,

$$A_\lambda(x^*) = \{ i \in \mathbb{N} : \lambda_i(x^*) > 0 \} \text{ and } A_\mu(x^*) = \{ j \in \mathbb{N} : \mu_j(x^*) > 0 \}. \quad (25)$$

Obviously, $A_\lambda(x^*) \subseteq A_Q(x^*), A_\mu(x^*) \subseteq A_E(x^*)$ and we can replace in $(24) A_Q(x^*)$ by $A_\lambda(x^*)$ and $A_E(x^*)$ by $A_\mu(x^*)$. Also, $|A_\mu(x^*)| \leq n$ and $\bigcup_{j \in A_E(x^*)} \{ e_j \}$ are linearly independent.

**Theorem 2.1** Let $x^* \in F \cup \partial F$ be a stationary point of (13). That is, $x^* \in \bigcap_{i \in A_Q(x^*)} Q_i \bigcap_{j \in A_E(x^*)} E_j$. Assume that $\{ g_i \}_{i \in A_Q(x^*)} \cup \{ e_j \}_{j \in A_E(x^*)}$ are linearly independent or span the whole $\mathbb{R}^n$. Then $x^*$ is isolated in $F \cup \partial F$ for $A$ large enough.

**Proof** A stationary point $x^* \in S \subseteq I^n$ is isolated, if there exists $\varepsilon > 0$ such that $x^*$ is the only stationary point in $B(x^*, \varepsilon)$. For the sake of brevity, let

$$A_\xi^* := A_\xi(x^*), \quad \xi \in \{ \lambda, \mu \}, \quad \text{and} \quad A_R^* := A_R(x^*), \quad R \in \{ Q, E \}, \quad (26)$$

as well as

$$\lambda_i^* := \lambda_i(x^*) \quad \text{and} \quad \mu_j^* := \mu_j(x^*) \quad (27)$$

for $i \in A_\lambda^*$ and $j \in A_E^*$.

We proceed by considering two cases: $x^* \in \partial F$ and $x^* \in F$.

**Case 1:** Let us first assume that $x^* \in \partial F$. Then

$$\nabla_C f(x^*) = \text{co}(\{ Ag_i : i \in A_Q^* \} \cup \{ 0 \}) \quad (28)$$

and $|A_Q^*| \geq 1$. Notice that $A_E^*$ is empty if $x^* \in (0, 1)^n$.

Let us consider two sub-cases: 1.1) $\{ g_i \}_{i \in A_Q^*} \cup \{ e_j \}_{j \in A_E^*}$ are linearly independent and 1.2) the vectors $\{ g_i \}_{i \in A_Q^*} \cup \{ e_j \}_{j \in A_E^*}$ span the whole $\mathbb{R}^n$. 

 Springer
Case 1.1: Let us further assume that \( \{ g_i \}_{i \in A_Q^*} \cup \{ e_j \}_{j \in A_E^*} \) are linearly independent. Due to \( x^* \in S \), we can write

\[
A \left( \sum_{i \in A_Q^*} \lambda_i^* g_i + \sum_{j \in A_E^*} \mu_j^* e_j \right) = 2(x_0 - x^*). \tag{29}
\]

Notice that the second sum of (29) vanishes when \( x^* \notin \partial \mathcal{I}^n \). Assume now that \( x^* \) is not isolated, i.e., for all \( \varepsilon > 0 \) there exists a stationary point \( \tilde{x}(\varepsilon) \in B(x^*, \varepsilon) \). Note that the function \( \tilde{x}(\varepsilon) \) is continuous in \( \varepsilon = 0 \).

Three cases are possible: \( \tilde{x}(\varepsilon) \in \partial \mathcal{F} \), \( \tilde{x}(\varepsilon) \in \mathcal{F} \) or \( \tilde{x}(\varepsilon) \in \mathcal{N} \mathcal{F} \). In at least one of these sets, there will be infinitely many \( \tilde{x}(\varepsilon) \) as \( \varepsilon \to 0 \). Hence, w.l.o.g. we can assume that \( \tilde{x}(\varepsilon) \) is always chosen in this set. We consider the two cases \( \tilde{x}(\varepsilon) \in \partial \mathcal{F} \) and \( \tilde{x}(\varepsilon) \in \mathcal{F} \) below, the third case \( \tilde{x}(\varepsilon) \in \mathcal{N} \mathcal{F} \) will be treated by the subsequent Lemma 2.2.

Again for the sake of brevity, let

\[
\tilde{A}_\xi(\varepsilon) := A_\xi(\tilde{x}(\varepsilon)), \quad \xi \in \{ \lambda, \mu \} \quad \text{and} \quad \tilde{A}_R(\varepsilon) := A_R(\tilde{x}(\varepsilon)), \quad R \in \{ Q, E \}, \tag{30}
\]

as well as

\[
\tilde{\lambda}_i(\varepsilon) := \lambda_i(\tilde{x}(\varepsilon)) \quad \text{and} \quad \tilde{\mu}_j(\varepsilon) := \mu_j(\tilde{x}(\varepsilon)) \tag{31}
\]

for \( i \in \tilde{A}_Q(\varepsilon) \) and \( j \in \tilde{A}_E(\varepsilon) \).

Case 1.1.1: Let us assume that \( \tilde{x}(\varepsilon) \in \partial \mathcal{F} \) holds. Let \( \varepsilon \) be small enough such that \( \tilde{A}_R(\varepsilon) \subseteq A_R^* \), \( R = Q, E \) and

\[
\tilde{x}(\varepsilon) \in \bigcap_{i \in \tilde{A}_Q(\varepsilon)} Q_i \bigcap_{j \in \tilde{A}_E(\varepsilon)} E_j \tag{32}
\]

with \( \tilde{A}_Q(\varepsilon) \neq \emptyset \). As \( \tilde{x}(\varepsilon) \) is stationary, we obtain

\[
A \left( \sum_{i \in \tilde{A}_Q(\varepsilon)} \tilde{\lambda}_i(\varepsilon) g_i + \sum_{j \in \tilde{A}_E(\varepsilon)} \tilde{\mu}_j(\varepsilon) e_j \right) = 2(x_0 - \tilde{x}(\varepsilon)). \tag{33}
\]

We consider

\[
d(\varepsilon) = \frac{\tilde{x}(\varepsilon) - x^*}{\| \tilde{x}(\varepsilon) - x^* \|} \tag{34}
\]
and conclude

\[ d(\varepsilon) = \frac{A}{2\|\tilde{x}(\varepsilon) - x^*\|} \left( \sum_{i \in A_Q^*} (\lambda_i^* - \tilde{\lambda}_i(\varepsilon)) g_i + \sum_{j \in A_E^*} (\mu_j^* - \tilde{\mu}_j(\varepsilon)) e_j \right) \]  

(35)

with \( \tilde{\lambda}_i(\varepsilon) = 0 \) for \( i \in A_Q^* \backslash \tilde{A}_\lambda(\varepsilon) \) and \( \tilde{\mu}_j(\varepsilon) = 0 \) for \( j \in A_E^* \backslash \tilde{A}_\mu(\varepsilon) \). Thus,

\[ 1 = \|d(\varepsilon)\| = \frac{A}{2\|\tilde{x}(\varepsilon) - x^*\|} \left( \sum_{i \in A_Q^*} (\lambda_i^* - \tilde{\lambda}_i(\varepsilon)) g_i + \sum_{j \in A_E^*} (\mu_j^* - \tilde{\mu}_j(\varepsilon)) e_j \right). \]  

(36)

As \( \|\tilde{x}(\varepsilon) - x^*\| < \varepsilon \), for \( \varepsilon \to 0 \) we conclude from the linear independence of \( \{g_i\}_{i \in A_Q^*} \cup \{e_j\}_{j \in A_E^*} \) that

\[ \tilde{\lambda}_i(\varepsilon) \to \lambda_i^* \quad \text{and} \quad \tilde{\mu}_j(\varepsilon) \to \mu_j^* \]  

(37)

for \( i \in A_Q^* \) and \( j \in A_E^* \). Since \( x^*, \tilde{x}(\varepsilon) \in \bigcap_{i \in \tilde{A}_Q(\varepsilon)} Q_i \cap \bigcap_{j \in \tilde{A}_E(\varepsilon)} E_j \cap \partial F \), it follows that

\[ d(\varepsilon)^T g_i = 0 = d(\varepsilon)^T e_j \quad \text{for} \quad i \in \tilde{A}_Q(\varepsilon), \quad j \in \tilde{A}_E(\varepsilon), \]  

(38)

cf. also remark 2.1. We obtain by (37) that

\[ A_{\xi}^* \subseteq \tilde{A}_{\xi}(\varepsilon), \quad \xi \in \{\lambda, \mu\}, \]  

(39)

for \( \varepsilon \) small enough. Thus,

\[ 2\|\tilde{x}(\varepsilon) - x^*\|d(\varepsilon) = \sum_{i \in \tilde{A}_\lambda(\varepsilon)} \tilde{\lambda}_i(\varepsilon) g_i - \sum_{i \in A_{\lambda}^*} \lambda_i^* g_i + \sum_{j \in \tilde{A}_\mu(\varepsilon)} \tilde{\mu}_j(\varepsilon) e_j - \sum_{j \in A_{\mu}^*} \mu_j^* e_j \]  

\[ = \sum_{i \in \tilde{A}_\lambda(\varepsilon)} (\tilde{\lambda}_i(\varepsilon) - \lambda_i^*) g_i + \sum_{j \in \tilde{A}_\mu(\varepsilon)} (\tilde{\mu}_j(\varepsilon) - \mu_j^*) e_j , \]  

(40)

which implies

\[ 2\|\tilde{x}(\varepsilon) - x^*\| = 2d(\varepsilon)^T \|\tilde{x}(\varepsilon) - x^*\|d(\varepsilon) \]  

\[ = \sum_{i \in \tilde{A}_\lambda(\varepsilon)} (\tilde{\lambda}_i(\varepsilon) - \lambda_i^*) d(\varepsilon)^T g_i + \sum_{j \in \tilde{A}_\mu(\varepsilon)} (\tilde{\mu}_j(\varepsilon) - \mu_j^*) d(\varepsilon)^T e_j = 0 \]  

(41)

yielding a contradiction.

The case \( \tilde{x}(\varepsilon) \in F \) is remaining. The case \( \tilde{x}(\varepsilon) \in NF \) will be treated by the subsequent Lemma 2.2.
Case 1.2: Assume now that $\tilde{x}(\varepsilon) \in \mathcal{F}$, $\|x^* - \tilde{x}(\varepsilon)\| < \varepsilon$ is a stationary point. If $\tilde{x}(\varepsilon) \notin \partial I^n$ it is clear that $\tilde{x}(\varepsilon)$ cannot be a stationary point. This is due to $\nabla C \tilde{x}(\varepsilon) = 2(x_0 - \tilde{x}(\varepsilon))$, which vanishes only if $\tilde{x}(\varepsilon) = x_0$, but $x_0 \in \mathcal{N}\mathcal{F}$. Hence, this implies $\tilde{x}(\varepsilon) \in \partial I^n$ and that (32) and (33) hold with $\tilde{A}_\lambda(\varepsilon) = \emptyset$, i.e.,

$$\tilde{x}(\varepsilon) \in \bigcap_{j \in \tilde{A}_E(\varepsilon)} E_j \quad \text{and} \quad 2(x_0 - \tilde{x}(\varepsilon)) = A \left( \sum_{j \in \tilde{A}_\mu(\varepsilon)} \tilde{\mu}_j(\varepsilon) e_j \right).$$

(42)

Consequently, we can repeat the reasoning from case 1.1.1 with $\tilde{A}_\lambda(\varepsilon) = \emptyset$ for any $\varepsilon > 0$. However, for $\varepsilon$ small enough, we obtain $A^*_\lambda \subseteq \tilde{A}_\lambda(\varepsilon) = \emptyset$, which is a contradiction to $A^*_\lambda \neq \emptyset$.

Case 1.2: Now assume that span$\{g_i\}_{i \in A^*_\lambda} \cup \{e_j\}_{j \in A^*_E}$ = $\mathbb{R}^n$. Let $\varepsilon > 0$ and let us consider the active sets $\tilde{A}_\lambda(\varepsilon) \subseteq \{1, \ldots, s\}$, $\tilde{A}_\mu(\varepsilon) \subseteq \{1, \ldots, m\}$ of which there can exist at most $2^s$ and $2^m$ different ones, respectively. Hence, at least one of those choices for $\tilde{A}_\lambda(\varepsilon)$ and $\tilde{A}_\mu(\varepsilon)$ must occur infinitely often for a sequence $\varepsilon \to 0$. Hence, for $\varepsilon > 0$ small enough, we can assume w.l.o.g. that $\tilde{A}_\lambda(\varepsilon)$ and $\tilde{A}_\mu(\varepsilon)$ are constant. For $A$ large enough we can thus write

$$d(\varepsilon) = \frac{A}{2\|\tilde{x}(\varepsilon) - x^*\|} \left( \sum_{i \in \tilde{A}^*_\lambda(\varepsilon)} (\lambda_i^* - \tilde{\lambda}_i(\varepsilon))g_i + \sum_{j \in \tilde{A}^*_\mu(\varepsilon)} (\mu_j^* - \tilde{\mu}_j(\varepsilon))e_j \right).$$

(43)

If the vectors $\{g_i\}_{i \in \tilde{A}_\lambda(\varepsilon)} \cup \{e_j\}_{j \in \tilde{A}_\mu(\varepsilon)}$ are linearly independent, we reside with case 1.1. If $\{g_i\}_{i \in \tilde{A}_\lambda(\varepsilon)} \cup \{e_j\}_{j \in \tilde{A}_\mu(\varepsilon)}$ are linearly dependent, we choose subsets of $\tilde{A}_\lambda(\varepsilon)$ and $\tilde{A}_\mu(\varepsilon)$ such that the preceding equation holds and the corresponding subsets are again linearly independent and span an $n-1$-dimensional subspace. Note that $n-1$ is the maximum dimension since all $g_i$ and $e_j$ are orthogonal to $d(\varepsilon)$. Hence, we can again repeat the reasoning from case 1.1. For the previous arguments, it is important to realize that the previous cases 1.1.1 and 1.1.2 do not require nonnegative coefficients $\lambda_i$, $\mu_j$. Only linear independence of the set of all corresponding vectors $g_i$ and $e_j$ is required.

Case 2: With similar arguments as used in the case $x^* \in \partial \mathcal{F}$, it can be shown that there are no stationary points inside of $\mathcal{F} \setminus \partial I^n$. W.r.t. $\mathcal{F} \cap \partial I^n$, stationary points are isolated by the considerations from case 1.1.2.

By the preceding theorem, we analyzed the nature of stationary points in $\mathcal{F} \cup \partial \mathcal{F}$ which will help us to use a result of [36] to prove the convergence of the projected sub-gradient method for sequences $\{x_k\}$ that are of interest to us. In particular, we are not interested in sequences $\{x_k\}$ that for $k$ greater than some $k_0$ get stuck in $\mathcal{N}\mathcal{F}$ as they cannot be considered as successful adversarial attacks.

The CW attack is a particular type of the projected sub-gradient method. It simplifies the general method in two points:
- It assumes that no \(x_k\) is contained in an intersection of (at least two) polytopes. This is likely since these intersections are probably sections of lower-dimensional subspaces, recall remark 2.1.
- It does not consider the normal cone \(N_{I^n}(x_k)\) although we included it.

The former implies that it is sufficient to only consider a single gradient \(g_i\) for computing \(x_{k+1}\) from \(x_k\). Still the theoretical analysis needs to deal with generalized gradients as stationary points are likely to be at polytope boundaries. Hence, our analysis of the stationary points in Theorem 2.1 is done for the general case, but in subsequent considerations we will use single gradients \(g_i\). The fact that the CW attack does not consider the normal cone \(N_{I^n}(x_k)\) allows us to exclude it from further considerations and the CW attack will not converge to any stationary point in \(\partial I^n \cap F\). In practice, we can efficiently filter out stationary points that are very close to \(\partial I^n\) since the probability distribution corresponding to original data points \(x_0 \in I^n\) has a measure of zero on \(\partial I^n\). Furthermore, in what follows, we will omit the projection \(P\) since for \(x_k \in I^n\) it is clear that

\[
\|P(x_{k+1}) - x\| \leq \|x_{k+1} - x\| \quad \forall x \in I^n. \tag{44}
\]

In order to characterize the sequences of interest and fix their properties, we formulate two further statements in the following. For the subsequent lemma, let \(Q_{s+1} = \partial F\) and let

\[
\mathcal{A}(x) := \{i \in \mathbb{N} : x \in Q_i\} \tag{45}
\]

be the set of all polytopes \(Q_i\), including the boundary \(\partial F\), that contain \(x\). Analogously, \(\mathcal{M}(x, \varepsilon)\) is the set of all indices in \(\{1, \ldots, s+1\}\) corresponding to the \(Q_i\) that intersect \(B(x, \varepsilon)\). That is,

\[
\mathcal{M}(x, \varepsilon) := \{i \in \mathbb{N} : Q_i \cap B(x, \varepsilon) \neq \emptyset\}. \tag{46}
\]

**Lemma 2.1** Let

\[
\delta_1 := \min_{x \in I^n} \sup_{\varepsilon} \left\{ \varepsilon : \bigcap_{i \in \mathcal{M}(x, \varepsilon)} Q_i \neq \emptyset \right\} \tag{47}
\]

be the maximal ball radius where for any \(x \in I^n\), all polytopes intersecting with that ball have at least one intersection point in common, then it holds that \(\delta_1 > 0\).

The proofs of Lemma 2.1 and 2.2 are given in appendix A.

**Lemma 2.2** Let \(\delta_1\) as in Lemma 2.1 and \(\delta_2 := \min_{Q_i \cap \partial F = \emptyset} \text{dist}(\partial F, Q_i)\), i.e., \(\delta_2\) smaller than the minimal distance between \(\partial F\) and any polytope not intersecting with the boundary of the feasible set \(\partial F\). Let \(\delta \leq \min\{\delta_1, \delta_2\}\), \(x_{k-1} \in F\), \(x_k \in \mathcal{N} \mathcal{F}\). Let \(\alpha_k \downarrow 0\), then

\[
\text{dist}(\partial F, x_\ell) < \delta \quad \tag{48}
\]
for \( \ell \geq k \), if \( A \) large enough and \( \alpha_k \) small enough.

**Definition 2.1** A CW attack sequence \( \{x_k\} \) is called *successful* if \( x_k \in \mathcal{F} \) for infinitely many \( k \).

Recall (21), i.e., \( D(\delta) = \mathcal{F} \cup \partial \mathcal{F} \cup \{ x \in \mathbb{R}^n : \text{dist}(x, \partial \mathcal{F}) < \delta \} \). The preceding lemma implies that for any successful attack for \( A \) large enough and \( \alpha_k \) small enough, all \( x_k \in D(\delta) \) for \( k \) large enough. We have also shown that all stationary points in \( \mathcal{F} \cup \partial \mathcal{F} \) are isolated in \( \mathcal{F} \cup \partial \mathcal{F} \). By the considerations of the induction step of Lemma 2.2, for \( \delta \) small enough there are no stationary points in \( \mathcal{N} \mathcal{F} \cap D(\delta) \). We have shown that all points in \( \mathcal{N} \mathcal{F} \cap D(\delta) \) are attracted by points on \( \partial \mathcal{F} \). By this, we conclude that all stationary points in \( D(\delta) \) are isolated. Hence, we can utilize a convergence result from [36] for the projected gradient descent method and adapt it to the family of successful CW attacks:

**Theorem 2.2** Let \( \{x_k\} \) be a successful attack. Then, for \( \delta > 0 \) small enough the set of all stationary points in \( D(\delta) \), i.e., \( S \cap D(\delta) \), is nowhere dense and the CW attack converges to an \( x^* \in S \cap D(\delta) \).

By Lemma 2.2, the sequence \( x_k \in D(\delta) \) for \( k \) large enough and \( \delta \) small enough. Since \( S \cap D(\delta) \) is nowhere dense, the conditions of [36, Theorem 4.1] are satisfied and the proof of the preceding theorem follows. It turns out that all successful CW attacks end up at the boundary \( \partial \mathcal{F} \). We say that a CW attack is successful if the sequence \( \{x_j\} \) converges to \( x^* \) and \( x^* \notin \mathcal{N} \mathcal{F} \).

**Lemma 2.3** Any successful CW attack converges to a stationary point on the boundary \( \partial \mathcal{F} \) of the feasible set.

The proof of this lemma is given in Appendix A.

### 2.4.2 Boundedness of the Counter Attack

Let us now consider the counter attack and therefore let \( \mathcal{F}^{(2)} \subseteq \mathcal{N} \mathcal{F}^{(1)} := \mathcal{N} \mathcal{F} \) denote the counter attack’s feasible region and \( \mathcal{N} \mathcal{F}^{(2)} \subseteq \mathcal{F}^{(1)} := \mathcal{F} \) the infeasible region.

We seek to minimize the functional

\[
F^{(2)}(x) := \text{dist}(x_k, x)^2 + Bf^{(2)}(x), \text{ s.t. } x \in \mathbb{R}^n
\]

where \( x_k \) is the final iterate of the primary attack and \( B \) is suitably chosen. The case \( p = 2 \) is considered again. The penalty \( f^{(2)} \) is chosen as in (15) but with \( t = \kappa(x_k) \). Let w.l.o.g. \( \{Q_1, \ldots, Q_{s'}\} \) be the set of all polytopes that fulfill \( x^* \in Q_i, \ i = 1, \ldots, s' \).

We assume that the first attack is successful and has iterated long enough such that the final iterate \( x_k \) of the first attack is so close to its stationary point that \( \text{dist}(x^*, x_k) < \varepsilon \) and \( B(x^*, 3\varepsilon) \subseteq \bigcup_{i=1}^{s'} Q_i \subseteq (0, 1)^n \). For solving (49) we will consider the same iterative method as before, i.e., we will assume that the iterative sequence is defined as

\[
x_{k,j+1} = P(x_{k,j} - \alpha_{k,j} G^{(2)}(x_{k,j})), \ j = 0, 1, \ldots
\]
with $P(\cdot)$ being the orthogonal projection onto $I^n$.

**Theorem 2.3** Consider the preceding assumptions and further assume that $f^{(2)}$ has no zero gradients inside $\bigcup_{i=1}^{N} Q_i$ and there are no polytope boundaries at $\partial F$. The counter attack is defined by (49) with $x_k \in N\mathcal{F}^{(2)}$, it stops when reaching $\mathcal{F}^{(2)}$, uses sufficiently small initial step sizes $\alpha_k$ and a sufficiently large $B$. Then the counter attack iterations $\{x_k\}$ stay within $\mathcal{B}(x^*, 3\varepsilon)$ for any $x_k \in B(x^*, 3\varepsilon)$.

**Proof** From now on, let $\mathcal{B} := \bigcup \{Q_i \cap \partial Q_i : Q_i \cap \partial \mathcal{F}^{(2)} \neq \emptyset\}$ and

$$c := \min_{z \in \mathcal{B}} \|g^{(2)}(z)\| > 0, \quad C := \max_{z \in \mathcal{B}} \|g^{(2)}(z)\|.$$  \hfill (51)

Note that these symbols were previously used for the first attack, but will from now on be used for the counter attack. Denote by $x \in B(x^*, 3\varepsilon)$ an arbitrary iteration $x_{k,j}$ of (50) and with $\alpha$ the corresponding learning coefficient $\alpha_{k,j}$. We will show that $P(x - x^* - \alpha G^{(2)}(x)) \in B(x^*, 3\varepsilon)$ for all $x \in B(x^*, 3\varepsilon)$, so the whole iterative sequence belongs to $B(x^*, 3\varepsilon)$. Given that $x - x^* \in (0, 1)^n$ for all $x \in B(x^*, 3\varepsilon)$ we have $\|P(x - x^* - \alpha G^{(2)}(x))\| \leq \|x - x^* - \alpha G^{(2)}(x)\|$ and we can omit the projection operator from now on. Take $B \geq \frac{8\varepsilon}{c}$. Let us consider two cases separately, the first one being $x \in B(x^*, 3\varepsilon) \setminus B(x^*, 2\varepsilon)$ and the second $x \in B(x^*, 2\varepsilon)$.

Let $x \in B(x^*, 3\varepsilon) \setminus B(x^*, 2\varepsilon)$ and $G^{(2)}(x) \in \nabla C f^{(2)}(x)$ be any gradient with corresponding $g^{(2)}(x) \in \nabla C f^{(2)}(x)$ which is constant on a given polytope $Q_i$ (containing $x$). We can assume that $x \in N\mathcal{F}^{(2)}$ since we are done otherwise. Furthermore, let $\theta$ be the angle between $x_k - x$ and $x^* - x$ (cf. Fig. 4). Then $\cos(\theta) = \frac{(x_k - x)^T(x^* - x)}{\|x_k - x\| \|x^* - x\|}$ where $\|\cdot\| = \|\cdot\|_2$. Given that $x^* - x_k = x^* - x - (x_k - x^*)$ we have

$$\cos(\theta) = \frac{\|x_k - x\|^2 + \|x^* - x\|^2 - \|x^* - x_k\|^2}{2 \|x_k - x\| \|x^* - x\|} \geq \frac{4\varepsilon^2 + \varepsilon^2 - \varepsilon^2}{2 \cdot 3\varepsilon \cdot 4\varepsilon} = \frac{1}{6}. \hfill (52)$$
Since \( G^{(2)}(x) = 2(x - x_k) + B g^{(2)}(x) \), we obtain

\[
G^{(2)}(x)^T(x - x^*) = 2(x - x_k)^T(x - x^*) + B g^{(2)}(x)^T(x - x^*) > 0
\]

\[
> 2(x - x_k)^T(x - x^*) \geq 4\varepsilon^2 \cos(\theta) \geq \frac{2\varepsilon^2}{3}.
\] (53)

Note that, \( g^{(2)}(x)^T(x - x^*) > 0 \) follows from \( g^{(2)}(x) \) being orthogonal to \( \partial \mathcal{F}^{(2)} \) and \( x - x^* \) being an ascent direction. The latter holds due to \( x \in \mathcal{Q}_i \cap N \mathcal{F}^{(2)} \) which implies \( f^{(2)}(x) > 0 \) and the fact that \( x^* \in \mathcal{Q}_i \) as well as \( f^{(2)}(x^*) = 0 \). Therefore, \( \|G^{(2)}(x)\|^2 \) can be bounded from above as follows:

\[
4\|x - x_k\|^2 + 4B(x - x_k)^Tg^{(2)}(x) + B^2\|g^{(2)}(x)\|^2
\leq 4(4\varepsilon)^2 + 4\frac{8\varepsilon}{c}4\varepsilon C + \frac{64\varepsilon^2}{c^2}C^2 = 64\varepsilon^2\left(1 + \frac{C}{c}\right)^2.
\] (54)

Hence, the difference in distance to \( x^* \) when performing a gradient descent step is

\[
\|x - x^*\|^2 - \|x - x^* - \alpha G^{(2)}(x)\|^2 = 2\alpha G^{(2)}(x)^T(x - x^*) - \alpha^2\|G^{(2)}(x)\|^2
\geq 2\alpha \cdot \frac{2\varepsilon^2}{3} - \alpha^264\varepsilon^2\left(1 + \frac{C}{c}\right)^2.
\] (55)

The latter expression is greater than zero if the step-size schedule \( \alpha \) is small enough, i.e., \( \alpha < 1/(16(1 + \frac{C}{c})^2) \). Hence, the distance to \( x^* \) decreases.

In the second case, \( x \in B(x^*, 2\varepsilon) \), we require \( \alpha \|G^{(2)}(x)\| < \varepsilon \) to ensure \( \|x - x^* - \alpha BG^{(2)}(x)\| \leq 3\varepsilon \). A simple derivation shows that this condition leads to a weaker bound on \( \alpha \). This concludes the proof. \( \square \)

Note that the preceding theorem and proof do not depend on whether we choose the first attack to be targeted or untargeted.

### 2.4.3 Asymptotically Perfect Separation

We now take into account the stochastic effects that stem from choosing an arbitrary initial image \( x \in I^n \) represented by a random variable \( X \). Let \( k \) be the number of iterates of the original CW attack and let \( X_k \in \mathcal{F}^{(1)} \) be the final iterate after \( k \geq 1 \) iterations, starting at \( X_0 = X \). For any random variables \( Y, Z \) with values in \( I^n \) let

\[
D(Y) = \text{dist}(Y, \mathcal{F}) = \inf_{y \in \mathcal{F}} \text{dist}(Y, y) \text{ and } D^{(0,k)} = \text{dist}(X_0, X_k) \geq D(X_0), \quad (56)
\]

provided \( X_k \in \mathcal{F}^{(1)} \), which means that the \( k \)-th iterate is a successful attack. For \( \tau \in \mathbb{R} \), we define the cumulative density function corresponding to a random variable

\( \square \) Springer
Fig. 5 A sketch of the information theoretic definition of AUROC. $\Delta F(Y \leq \tau)$ denotes a segment on the horizontal axis. From the limit $\Delta F(Y \leq \tau) \to 0$ as in the Riemann integral, we obtain (57)

$$D$$ (representing a random distance) by $F_D(\tau) = P(D \leq \tau) = D_*(P((\leq \infty, \tau]))$ where the push forward measure is $D_*(P(E) = P(D^{-1}(E))$ for all $E$ in the Borel $\sigma$-algebra.

The area under receiver operator characteristic curve (AUROC) of $D(Y)$ and $D(Z)$ is a measure that quantifies comprehensively how well the values of the random variables $D(Y)$ and $D(Z)$ can be separated by different thresholds $\tau$. More precisely, assuming that statistically $D(Y)$ is smaller than $D(Z)$, then, for a given threshold $\tau$, the true positive rate is given by $F_D(Y)(\tau)$ whereas the false positive rate is given by $F_D(Z)(\tau)$. Thus, a point of the ROC curve induced by $\tau$ given by $(F_D(Y)(\tau), F_D(Z)(\tau))$ measures how well the threshold $\tau$ separates values of $D(Y)$ from values of $D(Z)$, cf. Fig. 5. Since $D(Y)$ and $D(Z)$ are non-negative, the AUROC is given by

$$\text{AUROC}(D(Y), D(Z)) = \int_{\mathbb{R}^+} F_D(Y)(\tau) dF_D(Z)(\tau) = \int_{\mathbb{R}^+} F_D(Y)(\tau) dD(Z)_*P(\tau).$$

(57)

The following lemma formalizes under realistic assumptions that we obtain perfect separability of $D(X)$ and $D(X_k)$ as we keep iterating the initial CW attack, i.e., $k \to \infty$.

**Lemma 2.4** Let $D(X) \geq 0$ with $P(D(X) = 0) = 0$ and $D(X_k) \to 0$ for $k \to \infty$ weakly by law. Then,

$$\text{AUROC}(D(X_k), D(X)) \xrightarrow{k \to \infty} 1.$$

(58)
Proof Let $\delta_0$ be the Dirac measure in 0 with distribution function $F_{\delta_0}(z) = \{1 \text{ for } z \geq 0, \ 0 \text{ else}\}$. By the characterization of weak convergence in law by the Helly–Bray Lemma [9], $F_{D(X_k)}(\tau) \to F_{\delta_0}(\tau)$ for all $\tau$ where $F_{\delta_0}(\tau)$ is continuous. This is the case for all $\tau \neq 0$. As $P(D(X) = 0) = 0$, this implies $F_{D(X_k)} \to F_{\delta_0}$ $D(X)_*P$-almost surely. Furthermore, it holds that $|F_{D(X_k)}(\tau)| \leq 1$. Hence, by Lebesgue’s theorem of dominated convergence

$$AUROC(D(X_k), D(X)) = \int_{\mathbb{R}} F_{D(X_k)}(\tau) dD(X)_*P(\tau) \xrightarrow{k \to \infty} \int_{\mathbb{R}} F_{\delta_0}(\tau) dD(X)_*P(\tau) = \int_{\mathbb{R}_+} dD(X)_*P(\tau) = 1 \quad (59)$$

since $D(X) \geq 0$ by assumption. □

Now, let $X$ be a random input image with a continuous density. Assuming that decision boundaries of the neural net have Lebesgue measure zero, $X \in \mathcal{NF}^{(1)}$ holds almost surely, and therefore, $D(X) > 0$ almost surely, such that indeed $P(D(X) = 0) = 0$. Furthermore, let $X_k$ be the $k$-th iterate inside $\mathcal{F}^{(1)}$ of the CW attack starting with $X$. Conditioned on the event that the attack is successful, we obtain that $X_k \xrightarrow{k \to \infty} X^* \in \partial \mathcal{F}$, thus

$$\text{dist}(X_k, X^*) \xrightarrow{k \to \infty} 0 \quad (60)$$

almost surely for the conditional probability measure, which we henceforth use as the underlying measure. Now there exists a learning rate schedule $\alpha_{k,j}$ and a penalty parameter $A$ such that for all steps $X_{k,j}$ of the counter attack originating at $X_k$, the distance

$$\tilde{D}^{(k,j)} = \text{dist}(X_k, X_{k,j}) \leq 8 \text{dist}(X_k, X^*) \xrightarrow{k \to \infty} 0 \quad (61)$$

almost surely. Note that the latter inequality holds since we can choose $\varepsilon$ in Theorem 2.3, such that $\text{dist}(x_k, x^*) > \varepsilon/2$ and then obtain the constant by the triangle inequality. Let $X_{k,j^*}$ be the first iterate of the CW counter attack in $\mathcal{F}^{(2)}$. We consider

$$\tilde{D}^{(k,j^*)} = \text{dist}(X_k, X_{k,j^*}) \leq 8 \text{dist}(X_k, X^*) \xrightarrow{k \to \infty} 0 \quad (62)$$

almost surely. Hence, we obtain by application of Lemma 2.4 to (62):

**Theorem 2.4** Under the assumptions outlined above, we obtain the perfect separation of the distribution of the distance metric $\tilde{D}^{(k,j^*)}$ of the CW counter attack from the distribution of the distance metric $D^{(0,k)}$ of the original CW attack, i.e.,

$$AUROC \left( D^{(0,k)}, \tilde{D}^{(k,j^*)} \right) \xrightarrow{k \to \infty} 1. \quad (63)$$
In practice, the preceding theorem has the following consequences: The more eager the attacker is to minimize the attack perturbation applied to the original image $X_0$ (i.e., the larger $k$), the easier it is for the counter attack to detect the attack. This is implied by the asymptotically perfect separability of attack distances $D^{(0,k)}$ and counter attack distances $\bar{D}^{(k,j^*)}$.

3 Numerical Experiments

We now demonstrate how our theoretical considerations apply to numerical experiments. First, we introduce a two-dimensional classification problem which we construct in such a way that the theoretical assumptions are strictly respected and the mechanisms underlying our proof of perfect separability in the large iteration limit of the primary attack are illustrated.

3.1 Experiments with a 2D Example

For the 2D case, we create a dataset based on the two moons example. We use 2,000 data points to train a classifier and 300 data points as test set achieving a test accuracy of 93.33%. The test dataset and corresponding class predictions are shown in Fig. 6.
As classifier, a shallow ReLU network with one hidden layer consisting of 8 neurons is considered. The decomposition of this network into a finite set of polytopes \( \{ Q_j \}_{j=1}^t \) is given in Fig. 6 (top right). A corresponding surface plot of \( f(x) \) is shown in Fig. 6 (bottom left).

A stationary point \( x^* \) needs to fulfill

\[
2(x_0 - x^*) = A \sum_{i=1}^{t} \lambda_i g_i(x)
\]

for \( \lambda_i \geq 0 \) and \( \sum_{i=1}^{t} \lambda_i \leq 1 \) where \( g_i \) is the gradient of the \( i \)th polytope \( Q_i \) and w.l.o.g. \( x^* \in Q_i, i = 1, \ldots, t, t \geq 1 \). Hence, a penalization strength \( A \) is sufficient if it fulfills

\[
2 \|x_0 - x^*\| \leq 2\sqrt{n} \leq A \cdot c
\]

where \( n \) is the input dimension and \( c \) as in (51) the smallest gradient norm of all polytopes that intersect with the decision boundary. It follows that

\[
A \geq \frac{2\sqrt{n}}{c}
\]

is sufficiently large. Consequently, the penalization strength of the primary attack and the learning rate schedule are defined by

\[
A \in \left[ \frac{2\sqrt{n}}{c}, 100 \right] \quad \text{and} \quad \alpha_j = \alpha_0 \frac{j_0}{j_0 + j},
\]

respectively. The latter choice allows us to control the \( \sum_j \alpha_j \) in our experiments. For the input dimension \( n = 2 \), we considered up to 2,048 iterations for the primary attack and adjusted \( j_0 = 100 \) such that \( \sum_{j=1}^{2048} \alpha_j > \sqrt{2} \). We chose \( \alpha_0 = 0.01 \) for the primary attack. For the counter attack we use the parameters

\[
B = \frac{8\varepsilon}{c} \quad \text{and} \quad \alpha_{k,0} = \frac{1}{16(1 + \frac{\varepsilon}{c})^2}
\]

fulfilling the assumptions of Theorem 2.3. For numerical experiments we need to derive a lower bound.

In our experiments we use the CleverHans framework [27]. Unfortunately, this framework does not allow to easily access the weights of the neural network (which would help to compute polytopes). Therefore, we had to discretize the domain up to an order of \( 10^{-4} \) and compute decision and polytope boundaries via network inference. Hence, all numbers presented in this section are only accurate up to the discretization error. Consequently, we chose \( \varepsilon = \|x_k - x^*\| + 2 \cdot 10^{-4} \) individually for each data point \( x_0 \). The quantiles of calculated \( \varepsilon \) values for the different number of iterations are presented in Fig. 6 (bottom right). The counter attack performs up to 2,048 iterations.
(independently of the number of iterations of the primary attack) and stops in iteration $j^*$ when $x_{k, j^*} \in \mathcal{F}(2)$ for the first time ($x_{k, j^*}$ a realization of $X_{k, j^*}$). For all tests, we randomly split the test set into two equally sized and distinct portions $\mathcal{X}$ and $\mathcal{X}$. For $\mathcal{X}$, we run a primary attack as well as the counter attack generating values $D(0, k)$ and $\bar{D}(k, j^*)$. For $\mathcal{X}$, we only run a counter attack and obtain $D(0, j^*)$. We used the CleverHans framework [27] which includes the CW attack with $\ell_2$. The implementation contains a confidence parameter $\eta$, reformulating (15) into $f(x) = \max\{Z_t(x) - \max_{i \neq t} [Z_i(x)] - \eta, 0\}$. In accordance to our theory, we consider $\eta = 0$, recovering (15).

To empirically validate the statement in Theorem 2.3, we construct the corresponding $B(x^*, 3\varepsilon)$ and use all data points for which $B(x^*, 3\varepsilon) \subset \bigcup_{i=1}^{s'} Q_i$ with $x^* \in Q_i, \; i = 1, \ldots, s'$ is fulfilled. Figure 6 (bottom right) suggests that more iterations could be considered since stationarity is not yet achieved. However, the CW attack is already computationally costly for 2,048 iterations. Nevertheless, we observe in Fig. 7 (left) that the considered number of iterations is sufficient for perfect detection in terms of AUROC. We also registered return rates, i.e., the percentage of cases with successful primary attacks where we obtain $\kappa(x_0) = \kappa(x_{k, j^*})$. In a two class setting, this is guaranteed for all successful primary attacks being close enough to the decision boundary. Figure 7 (right) shows that the return rate indeed tends to 100% for an increasing number of iterations. We observe in our tests, independently of the number of iterations of the primary attack, that $x_k \in B(x^*, \varepsilon)$ implies $x_{k, j^*} \in B(x^*, 3\varepsilon)$ for all the data points. Also when the condition $B(x^*, 3\varepsilon) \subset \bigcup_{i=1}^{s'} Q_i$ is violated we stopped when $x_{k, j^*} \in \mathcal{F}(2)$ and still found $x_{k, j^*} \in B(x^*, 3\varepsilon)$.

### 3.2 Experiments with CIFAR10 and ImageNet2012

In the following, we evaluate the dependence on the number of primary attack iterations and perform cross attacks for different $\ell_p$ norms on the CIFAR10 and the ImageNet2012 dataset. In Sect. 3.2.1, we apply CleverHans studying the primary attack iterations for the CIFAR10 dataset [17] which consists of tiny $32 \times 32$ rgb images from 10 classes, containing 50k training and 10k test images. We trained a CNN consisting of 7 convolutional layers and achieved a test accuracy of 78.55%. Since CleverHans
only provides $\ell_2$ attacks, we also deploy the framework provided with [3] to study separability in different $\ell_p$ norms. For further experiments on separability and targeted attacks, the framework presented in [3] including pre-trained models is considered and applied to two datasets, CIFAR10 and ImageNet2012. For the CIFAR10 dataset, we used a network with 4 convolutional layers and 2 dense ones as included and used per default in the code. For the ImageNet2012 high resolution rgb images, we used a pre-trained Inception network ([11], trained on the standard subset of 1,000 classes). All networks were trained with default train/val/test splittings.

In contrast to the 2D experiments in Sect. 3.1, we cannot compute the stationary points in higher dimensions anymore as this is computationally prohibitive. Hence we adjust the following parameters independently of the chosen framework: For parameters $A$ and $B$, we run a binary search in the range 0 to $10^{10}$ (default in both frameworks). We perform 1000 attack and counter attack iterations, if not stated otherwise. For the counter attack we set the initial learning rate $\alpha_0 = 0.01$ (which is the same value as for the primary attack). In the CleverHans framework, we modify the learning rate schedule of the primary and counter as defined in (67) with $n_0 = 100$, while in the other framework the Adam optimizer is used as default. For each of the two datasets, CIFAR10 and ImageNet2012, we randomly split the test set into two equally sized portions $\mathcal{X}$ and $\bar{\mathcal{X}}$. For one portion we compute the values $D(0,j)$ and for the other one $D(0,k)$ and $\bar{D}(k,j)$, such that they refer to two distinct sets of original images. Since CW attacks can be computationally demanding, we chose the sample sizes for $\mathcal{X}$ and $\bar{\mathcal{X}}$ as stated in Table 1. This experimental setup yields a CW attack which is successful on all attacked images when performing 1000 attack iterations.

### 3.2.1 Primary Attack Iterations and Separability

Performing experiments on the CIFAR10 dataset with the CleverHans framework analogously to the experiments with the 2D example presented in Sect. 3.1, we observe a similar behavior as in Fig. 7, see Fig. 8, although we are not able to check whether our theoretical assumptions hold. We performed at most 1,024 counter attack iterations due to the computational cost of the CW attack. For 1,024 primary attack iterations, only 5 examples did not return to their original class, 3 of them remained in the primary attack’s class, 2 moved to a different one.

---

1. [http://image-net.org](http://image-net.org)
3.2.2 Separability of Different Distance Measures on CIFAR10 and ImageNet2012

In order to discuss the distance measures $D^{(0,j)}$ and $\tilde{D}^{(k,j)}$ for different $\ell_p$ norms, we now alter the notation. Let $\gamma_p : I^n \rightarrow I^n$ be the function that maps $x \in \mathcal{X}$ to its $\ell_p$ attacked counterpart for $p = 0, 2, \infty$. In order to demonstrate the separability of $\mathcal{X}$ and $\gamma_p(\tilde{X})$ under a second $\ell_q$ attack, we compute the two scalar sets

$$D_q = \{\text{dist}_q(x, \gamma_q(x)) : x \in \mathcal{X}\}, \quad \tilde{D}_{p,q} = \{\text{dist}_q(\gamma_p(\tilde{x}), \gamma_q(\gamma_p(\tilde{x}))) : \tilde{x} \in \tilde{X}\}(69)$$

for $q = 0, 2, \infty$. As mentioned previously, we altered the code to the one provided with [3]. Besides CIFAR10 we consider ImageNet2012\(^2\) with high resolution rgb images. From now on we used default parameters and performed 1,000 iterations for both attacks, respectively. For the number of images used, see Table 1. Since the $\ell_p$ norms used in our tests are equivalent except for $\ell_0$, we expect that cross attacks, i.e., the case $p \neq q$ is supposed to yield a good separation of $D_q$ and $\tilde{D}_{p,q}$ (cf. (69)). Table 2 shows results for cross attacks with AUROC values up to 99.73%. Each column shows the detection performance of a norm $\ell_q$. In both cases, for CIFAR10 and ImageNet2012, when comparing the different columns of both plots we observe a superiority of the $\ell_2$ norm. In our tests we observe that the $\ell_2$ norm also requires lowest computational effort, thus the $\ell_2$ norm might be favorable from both perspectives. Noteworthily, there is also only a minor performance degradation when going from CIFAR10 to ImageNet2012 although the perturbations introduced by the $\ell_p$ attacks, in particular for $p = 2$, are almost imperceptible, cf. also Fig. 2. The inferiority of the $\ell_0$ attack might be due to the less granular iteration steps, changing pixel values maximally. In Table 3 we report the number of returned examples and return rates. For both datasets we observe strong return rates although there is no theoretical guarantee in a multi class setting. For further insights on the distribution of distances, we include violin plots visualization the distributions of $D_p$ and $\tilde{D}_{p,p}$, i.e., for the case $p = q$, see Figs. 9 and 10. For both datasets, the distributions of $D_p$ and $\tilde{D}_{p,p}$ are also visually well separable.

\(^2\) http://image-net.org
Table 2  AUROC values for the different cross attacks performed to separate $D_q$ and $\bar{D}_{p,q}$, the $\ell_q$ norm is the one used for detection

| Attack  | CIFAR10 |  | ImageNet2012 |  |  |
|---------|---------|---|--------------|---|---|
| 2nd ↓ \ 1st → | $p = 0$ (%) | $p = 2$ (%) | $p = \infty$ (%) | $p = 0$ (%) | $p = 2$ (%) | $p = \infty$ (%) |
| $q = 0$ | 91.86  | 99.62 | 99.64 | 98.73  | 99.30 | 99.73 |
| $q = 2$ | 91.25  | 99.13 | 99.17 | 95.82  | 96.04 | 96.75 |
| $q = \infty$ | 89.23  | 96.81 | 96.29 | 96.31  | 94.55 | 95.61 |

Table 3  Number of returnees as well as return rates

| Dataset   | $\ell_0$ | $\ell_2$ | $\ell_\infty$ |
|-----------|----------|----------|---------------|
| Returnees | CIFAR10  | 4994     | 4997          | 995          |
| Percentage| CIFAR10  | 99.88%   | 99.94%        | 99.5%        |
| Returnees | ImageNet2012 | 500    | 486           | 488          |
| Percentage| ImageNet2012 | 100%   | 97.20%        | 97.6%        |

All primary attacks were successful

Fig. 9  Violin plots displaying the distributions of distances in $D_p$ (top) and $\bar{D}_{p,p}$ (bottom) from (69) for the ImageNet dataset, $p = 0, 2, \infty$ in ascending order from left to right. The vertical red line indicates the optimal separation threshold

Fig. 10  Violin plots displaying the distributions of distances in $D_p$ (top) and $\bar{D}_{p,p}$ (bottom) from (69) for the ImageNet dataset, $p = 0, 2, \infty$ in ascending order from left to right. The vertical red line indicates the optimal separation threshold

3.2.3 Targeted Attacks on CIFAR10

So far, all results presented have been computed for untargeted attacks. In principle a targeted attack $\gamma_p$ only increases the distances $\text{dist}_q(\gamma_p(\overline{x}), \overline{x})$ while the distance measures $\text{dist}_q(\gamma_q(\gamma_p(\overline{x})), \gamma_q(\overline{x}))$ corresponding to another untargeted attack $\gamma_q$ are supposed to remain unaffected. Thus, targeted attacks should be even easier to detect as
confirmed by Fig. 11. However, intuitively the counter attack might be less successful in returning to the original class than in the untargeted case. For the untargeted primary attack on CIFAR10 with $p = q = 2$, we obtained a return rate of 99.94% (cf. Table 3); this number is indeed reduced to 73.32% when performing a targeted primary attack.

### 3.3 Comparison with State-of-the-Art Defense Methods

Recent works report their numbers in a high heterogeneity w.r.t. considered attacks, evaluation metrics and datasets. This makes a clear comparison difficult. Table 4 summarizes a comparison with state-of-the-art detection methods. Our approach is situated in the upper part of the spectrum. PDG $\ell_2$ denotes the projected gradient
descent method where gradients are normalized in an $\ell_2$ sense, see [24]. This attack is often considered in tests; however, for most detection methods PDG $\ell_2$ might be easier to detect than the CW attack as PDG $\ell_2$ does not aim at minimizing the perturbation strength. For the detection by attack (DBA) approach [46] which is similar to ours, it was observed empirically that eager attacks like CW are easier to detect than others. In light of our theoretical statements, this observation is to be expected. The results presented by [40, 41] are limited to MNIST where the detection of attacks is simpler. Their findings are in line with those in [46], for the setting where both attacks are CW attacks, the authors of [40, 41] state a return rate of 99.6% which is in line with what we observe. For additional interesting numerical comparisons of counter attacks based on attacks different to CW, we refer to [40, 41, 46].

4 Conclusion & Outlook

We presented a mathematical proof for asymptotically optimal detection of CW attacks via CW counter attacks. In numerical experiments we confirmed that the number of iterations indeed increases the separability of attacked and non-attacked images. Even when further relaxing the theoretical assumptions and considering default parameters, we still obtain detection rates of up to 99.73% AUROC that are in the upper part of the spectrum of detection rates reported in the literature. For future work we plan to study the effect of introducing a confidence parameter in the primary CW attack. We expect that the treacherous efficiency of the CW attack can still be noticed in that case. Furthermore we plan to investigate under which conditions stationary points of the CW attack are local minimizers and to consider theoretically the cases $p \neq 2$. Our CleverHans branch is publicly available under https://github.com/kmaag/cleverhans.

Acknowledgements We would like to thank the anonymous reviewers whose constructive comments helped us to improve the paper.

Funding Open Access funding enabled and organized by Projekt DEAL. The work of N. Krejić is supported by Provincial Secretariat for Higher Education and Scientific Research of Vojvodina, grant no. 142-451-2593/2021-01/2.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.
A Supplementary Material on the Theory

Lemma 2.1 Let

\[ \delta_1 := \min_{x \in I^n} \sup_{\varepsilon} \{ \varepsilon : \bigcap_{i \in M(x, \varepsilon)} Q_i \neq \emptyset \} \]  

(47)

be the maximal ball radius where for any \( x \in I^n \), all polytopes intersecting with that ball have at least one intersection point in common, then it holds that \( \delta_1 > 0 \).

**Proof** Let us assume the contrary, i.e., \( \delta_1 = 0 \). Then there exists an \( \hat{x} \in I^n \) such that for all \( \varepsilon > 0 \) it holds that \( \bigcap_{i \in M(\hat{x}, \varepsilon)} Q_i = \emptyset \). (70)

This implies that for all \( \varepsilon > 0 \) there exists an \( i \in M(\hat{x}, \varepsilon) \) such that \( \hat{x} \notin Q_i \), and equivalently \( i \notin A(\hat{x}) \). Since the set of all possible \( i \notin A(\hat{x}) \) is finite and the \( Q_i \) are compact, it follows that

\[ \delta := \min_{i \notin A(\hat{x})} \text{dist}(Q_i, \hat{x}) > 0. \]  

(71)

Hence, for \( \varepsilon < \delta \) we obtain \( M(\hat{x}, \varepsilon) = A(\hat{x}) \) and there exists an \( i \in M(\hat{x}, \varepsilon) = A(\hat{x}) \) such that \( \hat{x} \notin Q_i \). This is a contradiction to the definition of \( A(\hat{x}) \). \( \square \)

Lemma 2.2 Let \( \delta_1 \) as in Lemma 2.1 and \( \delta_2 := \min_{Q_i \cap \partial F = \emptyset} \text{dist}(\partial F, Q_i) \), i.e., \( \delta_2 \) smaller than the minimal distance between \( \partial F \) and any polytope not intersecting with the boundary of the feasible set \( \partial F \). Let \( \delta \leq \min\{\delta_1, \delta_2\} \), \( x_{k-1} \in F \), \( x_k \in NF \). Let \( \alpha_k \downarrow 0 \), then

\[ \text{dist}(\partial F, x_\ell) < \delta \]  

(48)

for \( \ell \geq k \), if \( A \) large enough and \( \alpha_k \) small enough.

**Proof** Let \( g_i \) denote the gradient corresponding to \( Q_i \), \( i = 1, \ldots, s' \), where w.l.o.g. \( \{Q_i\}_{i=1}^{s'} \) is exactly the set of polytopes that have non-empty intersection with \( \partial F \). Let

\[ c := \min_{i=1, \ldots, s'} \|g_i\| > 0, \quad C := \max_{i=1, \ldots, s'} \|g_i\|. \]  

(72)

W.l.o.g. let \( x_k \) be contained in \( Q_1 \) with corresponding gradient \( g := g_1 \). We can assume that \( \alpha_k \) is small enough such that

\[ \|x_{k+1} - x_k\|^2 = \alpha_k^2 \|2(x_k - x_0) + Ag\|^2 \leq \alpha_k^2 (4\|x_k - x_0\|^2 + 2A\|x_k - x_0\| \cdot \|g\| + A^2\|g\|^2) \leq \alpha_k^2 (4n + 2\sqrt{n}AC + A^2C^2) = \alpha_k^2 (2\sqrt{n} + AC)^2 < \frac{\delta^2}{16} \]  

(73)
by requiring
\[ \alpha_k < \frac{\delta}{(8\sqrt{n} + 4AC)^2}. \]  

The preceding conservative bound with squared denominator shall help us again later on. We now proceed via induction and prove the following: In the base step, \( x_k \) is contained in \( Q_1 \) with \( \text{dist}(x_k, \partial F) < 2\alpha_{k-1}\sqrt{n} \) and \( \|x_k - x_{k-1}\| < 3\alpha_{k-1}\sqrt{n} \). By Lemma 2.1 we obtain that \( Q_1 \cap \partial F \neq \emptyset \) and there exists an \( x^* \) that is contained in the intersection of all polytopes intersecting with \( B(x_k, 3\alpha_{k-1}\sqrt{n}) \). We show in the base step that \( \|x^* - x_k\| < \delta \) for \( \alpha_k \) sufficiently small. In the induction step, we prove for \( \ell \geq k \) that if \( x_\ell \) is contained in some \( Q_1 \) and \( x^* \) is contained in the intersection of all polytopes intersecting with \( B(x^*, \delta) \) and \( \|x_\ell - x^*\| < \delta \), then also \( \|x_{\ell+1} - x^*\| < \delta \).

**Base step—case 1:** W.l.o.g. let \( x_k \in Q_1 \). We assume that \( B(x_k, 3\alpha_{k-1}\sqrt{n}) \cap Q_i = \emptyset \) for any \( i \geq 2 \). By (73), we obtain \( x_k \in B(x_{k-1}, 3\alpha_{k-1}\sqrt{n}) \) and our assumption implies that the line segment between \( x_{k-1} \) and \( x_k \) intersects only with a single polytope, namely \( Q_1 \). We denote by
\[
\text{co}\{x_k, x_{k-1}\} = \{\lambda x_k + (1-\lambda)x_{k-1} : \lambda \in [0, 1]\} \subseteq Q_1 \cup \mathcal{F} \tag{75}
\]
the line segment between \( x_k \) and \( x_{k-1} \). Then, there exists \( x^* \in \text{co}\{x_k, x_{k-1}\} \cap \partial F \subseteq Q_1 \). This implies by (73)
\[
\|x_k - x^*\| < \|x_k - x_{k-1}\| \leq 2\alpha_{k-1}\sqrt{n} < \delta. \tag{76}
\]

**Base step—case 2:** Assume \( B(x_k, 3\alpha_{k-1}\sqrt{n}) \cap Q_i \neq \emptyset \), w.l.o.g. \( i = 1, \ldots, j, j \geq 2 \), i.e., \( B(x_k, 3\alpha_{k-1}\sqrt{n}) \) intersects with at least two polytopes. Since we ruled out the case to end up in a polytope not intersecting with \( \partial F \) by the choice of \( \delta_2 \), it follows that \( Q_i \cap \partial F \neq \emptyset \) for all \( i = 1, \ldots, j \) and there exists \( x^* \in \partial F \cap \bigcup_{i=0}^j Q_i \) by Lemma 2.1. We again consider the line segment \( \text{co}\{x_k, x_{k-1}\} \). In what follows we construct in each of the subsequent cases a line segment inside the ball \( B(x_k, 3\alpha_{k-1}\sqrt{n}) \) such that this particular line segment intersects with a polytope, which does not contain any of the two end points of the line segment. We then construct a triangle with that line segment and \( x^* \). By upper bounding the cosine of the angle at \( x^* \), we then derive the desired bound on the distance between \( x_k \) and \( x^* \).

**Case 2.1:** Assume that the line segment \( \text{co}\{x_k, x_{k-1}\} \) intersects with at least two polytopes, i.e., w.l.o.g. \( \text{co}\{x_k, x_{k-1}\} \cap Q_i \neq \emptyset \) for \( i = 1, \ldots, j' \) and \( 2 \leq j' \leq j \), for at least one \( i \in \{1, \ldots, j'\} \), \( \text{co}\{x_k, x_{k-1}\} \cap Q_i \) is more than singleton and both \( x_k \) and \( x_{k-1} \) are contained in at least one additional polytope (or \( \partial F \cup \mathcal{F} \)) different from \( Q_i \) (cf. Fig. 12). Clearly, \( \text{co}\{x_k, x_{k-1}\} \) also intersects with \( \mathcal{F} \). For such a situation, we upper bound the cosine of the angle \( \theta \) in Fig. 12. To make this procedure applicable to similar cases, we proceed generically. Let \( R_i := Q_i, i = 1, \ldots, s, R_{s+1} := \partial F \cup \mathcal{F} \) and
\[
R := \{ (z_a, z_b) \in I^n \times I^n : z_a \in R_{i_1}, z_b \in R_{i_2}, i_1 \neq i_2 \text{ and } \exists i_3 \neq \{i_1, i_2\} \text{ s.t. } \text{co}\{z_a, z_b\} \cap R_{i_3} = \text{co}\{z_{a'}, z_{b'}\} \text{ with } z_{a'} < z_{b'} \} \tag{77}
\]
Fig. 12 An exemplary 2D illustration of case 2.1 of the base step. The angle \( \theta \) satisfies \( \cos(\theta) \leq \gamma_{\text{max}} < 1 \). Also note the scale invariance of such scenarios.

For any such \((z_a, z_b) \in \mathcal{R}\) with \( \|z_a - z_b\| < \delta \), there exists a \( z^* \) by Lemma 2.1 and we can consider the triangle given by \( z_a, z_b \) and \( z^* \). Consequently,

\[
\gamma_{\text{max}} := \sup_{(z_a, z_b) \in \mathcal{R}, \|z_a - z_b\| < \delta} \frac{(z_a - z^*)^T (z_b - z^*)}{\|z_a - z^*\| \cdot \|z_b - z^*\|} < 1, \tag{78}
\]

also cf. Fig. 12 with \( z_a = x_{k-1} \) and \( z_b = x_k \). A cosine of 1 would yield a degenerate triangle and therefore at least one degenerate polytope \( Q_i \), \( i = 1, \ldots, s \) with empty interior.

Now, we consider the triangle given by \( x_k, x_{k-1} \) and \( x^* \). Extending the triangle by an altitude \( h \perp (x_{k-1} - x^*) \) to a triangle given by \( h, x_k - x^* \) and \( \beta(x_{k-1} - x^*) \) for some \( \beta > 0 \), it follows that

\[
\|h\| \leq \|x_{k-1} - x_k\| < 3\alpha_{k-1}\sqrt{n}. \tag{79}
\]

Via the law of sines, we obtain

\[
\frac{3\alpha_{k-1}\sqrt{n}}{\|x_k - x^*\|} > \sqrt{1 - \gamma_{\text{max}}^2} \tag{80}
\]

which implies

\[
\|x_k - x^*\| < \frac{3\alpha_{k-1}\sqrt{n}}{\sqrt{1 - \gamma_{\text{max}}^2}}. \tag{81}
\]

Hence, for \( \alpha_{k-1} \leq \sqrt{1 - \gamma_{\text{max}}^2}\delta/(3\sqrt{n}) \), which for \( \alpha_k \downarrow 0 \) we may assume as well, we obtain

\[
\|x_k - x^*\| < \delta. \tag{82}
\]

**Case 2.2:** Assume that the line segment between \( x_k \) and \( x_{k-1} \) intersects only with a single polytope, i.e., \( \text{co}\{x_k, x_{k-1}\} \subseteq Q_1 \cap \mathcal{F} \). In what follows, we consider two
different sub-cases where we either conclude that $x^*$ is already contained in the ball $B(x_k, 3\alpha_{k-1}\sqrt{n})$ or we construct a triangle as in case 2.1, again yielding the desired conclusion as in (82).

**Case 2.2.1:** Let $x' \in B(x_k, 3\alpha_{k-1}\sqrt{n})$, then $x' \in Q_i$ for some $i \in \{1, \ldots, j\}$. Assume that

$$\text{co}\{x', x_{k-1}\} \subseteq Q_i \cup \mathcal{F}$$

for all $x' \in B(x_k, 3\alpha_{k-1}\sqrt{n})$. Hence, for all $x'$ in polytope intersections we obtain that the corresponding line segment $\text{co}\{x', x_{k-1}\}$ is contained in the intersection as well. That is, for all $J \subseteq \{1, \ldots, j\}$ with $x' \in \bigcap_{i \in J} Q_i \cap B(x_k, 3\alpha_{k-1}\sqrt{n}) \neq \emptyset$ we obtain

$$\text{co}\{x', x_{k-1}\} \subseteq \bigcap_{i \in J} Q_i \cup \mathcal{F}. \quad (84)$$

Now let

$$J_{\text{max}} \in \arg \max_{J' \subseteq \{1, \ldots, j\}} \left\{|J'| : \bigcap_{i \in J'} Q_i \cap B(x_k, 3\alpha_{k-1}\sqrt{n}) \neq \emptyset\right\} \quad (85)$$

a maximal subset of indices $i \in \{1, \ldots, j\}$ such that the intersection of all the polytopes $Q_i$, $i \in J_{\text{max}}$, and the ball is non-empty. For $J_{\text{max}} = \{1, \ldots, j\}$, then by (84) there exists an $x^* \in \bigcap_{i \in J_{\text{max}}} B(x_k, 3\alpha_{k-1}\sqrt{n}) \cap \partial \mathcal{F}$ and thus $\|x_k - x^*\| < 3\alpha_{k-1}\sqrt{n} < \delta$.

Hence, let us assume $|J_{\text{max}}| < j$. That is, there exists at least one polytope that is not intersecting with a maximal intersection of polytopes inside the ball. This implies the existence of at least one $j^* \in \{1, \ldots, j\} \setminus J_{\text{max}}$ with $Q_{j^*} \cap B(x_k, 3\alpha_{k-1}\sqrt{n}) \neq \emptyset$. Let

$$\hat{x} = \arg \min_{x \in \bigcap_{i \in J_{\text{max}}} Q_i} \{\text{dist}(x, Q_{j^*}) : j^* \notin J_{\text{max}}\} \quad (86)$$

and

$$(x', j^*) = \arg \min_{x \in Q_{j^*}, j^* \notin J_{\text{max}}} \text{dist}(x, \hat{x}), \quad (87)$$

i.e., we also fix the corresponding $j^*$. This implies that

$$\text{co}\{\hat{x}, x'\} \cap Q_{j^*} \neq \emptyset \quad \text{and} \quad \text{co}\{\hat{x}, x'\} \bigcap_{i \in J_{\text{max}}} Q_i \neq \emptyset, \quad (88)$$

i.e., $\text{co}\{\hat{x}, x'\}$ intersects with at least three polytopes (due to $|J_{\text{max}}| \geq 2$) and

$$0 < \|\hat{x} - x'\| < 6\alpha_{k-1}\sqrt{n}. \quad (89)$$
The triangle given by $x', \hat{x}$ and $x^*$ can be treated as in case 2.1, thus concluding for

$$\alpha_{k-1} < \frac{\sqrt{1 - \gamma_{\text{max}}^2}}{6\sqrt{n}}$$  \hfill (90)

that

$$\|x_k - x^*\| < \delta. \hfill (91)$$

**Case 2.2.2:** Assume there exists an $x' \in B(x_k, 3\alpha_{k-1}\sqrt{n})$ such that $\text{co}\{x', x_{k-1}\}$ intersects with at least two $Q_i$. Then,

$$\|x' - x_{k-1}\| < 6\alpha_{k-1}\sqrt{n}. \hfill (92)$$

The triangle given by $x', x_{k-1}$ and $x^*$ can be treated as in case 2.1. Thus, we conclude for $\alpha_{k-1}$ as in (90) that

$$\|x_k - x^*\| < \delta. \hfill (93)$$

Hence, for $\alpha_{k-1}$ small enough such that (74) and (90) hold, this concludes the base step.

**Induction step:** We now perform the induction step and assume that $x_\ell$ is contained in some $Q_1$ and $x_\ell \in B(x^*, \delta)$ with $x^*$ as in the base step. Hence, if $x_{\ell+1}$ is contained in a different polytope than $x_\ell$, they still share an $x^*$ which is contained in both of the polytopes.

We need to prove that $x_{\ell+1} \in B(x^*, \delta)$. If $x_\ell \in B(x^*, \frac{3}{4}\delta)$, this follows directly by (73). Hence, let us assume $x_\ell \notin B(x^*, \frac{3}{4}\delta)$. Clearly, the following holds:

$$\|x_{\ell+1} - x^*\|^2 = \|x_{\ell+1} - x_\ell\|^2 + 2(x_{\ell+1} - x_\ell)^T(x_\ell - x^*) + \|x_\ell - x^*\|^2. \hfill (94)$$

Hence, it suffices to prove

$$\|x_{\ell+1} - x_\ell\|^2 + 2(x_{\ell+1} - x_\ell)^T(x_\ell - x^*) < 0. \hfill (95)$$

Analogously to the choice of $x^*$ in the base step but more generally, let

$$x_{\partial F}(x) := \begin{cases} \arg\min_{x' \in \partial F} \text{dist}(x, x') & \text{if } B(x, \delta) \cap \partial F \subseteq Q(x), \\ \arg\min_{x' \in \partial F \cap j \in M(x, \delta) Q_j} \text{dist}(x, x') & \text{else}. \end{cases} \hfill (96)$$

and

$$\gamma_{\min} := \min_j \min_{x \in Q_j} \frac{(x - x_{\partial F}(x))^Tg_j}{\|x - x_{\partial F}(x)\| \cdot \|g_j\|}, \hfill (97)$$
then $\gamma_{\text{min}} > 0$ due to sufficient overlap of the $\frac{3}{4}\delta$-balls of each of the cases in (96). Inserting the iteration prescript for $x_{\ell+1}$ into (95), we obtain that statement (95) holds, iff

$$4\alpha_{\ell}^2\|x_{\ell} - x_0\|^2 + 2\alpha_{\ell}^2A(x_{\ell} - x_0)^T g + \alpha_{\ell}^2A^2\|g\|^2 - 4\alpha_{\ell}(x_{\ell} - x_0)^T(x_{\ell} - x^*) - 4\alpha_{\ell}Ag^T(x_{\ell} - x^*) < 0.$$ (98)

Dividing on both sides of the inequality by $\alpha_{\ell}$, the former holds, if

$$4\alpha_{\ell}\sqrt{n}C + \alpha_{\ell}A^2C^2 < 4(x_{\ell} - x_0)^T(x_{\ell} - x^*) + 4Ag^T(x_{\ell} - x^*).$$ (99)

which is implied by

$$\alpha_{\ell}(2\sqrt{n} + CA)^2 < -4\sqrt{n}\delta + 4A\gamma_{\text{min}}c \frac{3}{4}\delta.$$ (100)

Using (74), the former is implied by

$$\delta + 4\sqrt{n}\delta < 3A\gamma_{\text{min}}c\delta.$$ (101)

This holds for $n \geq 1$ if

$$A > \frac{5\sqrt{n}}{3c\gamma_{\text{min}}}$$ (102)

which we may demand by the assumptions of the theorem. The lemma is proved. □

**Lemma 2.3** Any successful CW attack converges to a stationary point on the boundary $\partial F$ of the feasible set.

**Proof** Let $\{x_k\}$ be an iterative sequence such that $\lim_{k \to \infty} x_k = x^*$. We are going to show that for any $x_k \in F$ there exists $\ell \in \mathbb{N}$ such that $x_{k+\ell} \in N F$ and hence we cannot have $x^* \in F$, i.e., $x^* \in \partial F$ holds.

For all $x_k \in F$ we have $F(x_k) = \text{dist}(x_0, x_k)^2$ and thus $G_k = \nabla \text{dist}(x_0, x_k)^2 = 2(x_k - x_0)$. As $x_k \in I^n$ we have that $\|G_k\| \leq C_1$ for an appropriate constant $C_1 > 0$. Furthermore, for all $x_k \in F$ we have $\text{dist}(x_0, x_k)^2 > C_2$ for an appropriate constant $C_2 > 0$. The latter holds due to $x_0 \in \mathcal{N} F$. Given the properties of the learning rate $\alpha_k$, there exists $k_0$ such that for all $k \geq k_0$ we have

$$\alpha_k^2C_1^2 - 2\alpha_kC_2 < 0.$$ (103)

Let us consider $x_k \in F$, $k \geq k_0$. Then,

$$\|x_{k+1} - x_0\|^2 = \|P(x_k - \alpha_k G_k) - P(x_0)\|^2 \leq \|x_k - \alpha_k G_k - x_0\|^2 = \|x_k - x_0\|^2 + \alpha_k^2\|G_k\|^2 - 2\alpha_k(x_k - x_0)^T G_k$$

$$= \|x_k - x_0\|^2 + \alpha_k^2\|G_k\|^2 - 2\alpha_k\text{dist}(x_0, x_k)^2 < \|x_k - x_0\|^2,$$ (104)
due to (103). Assume now that for any $\ell \in \mathbb{N}$ we have $x_k, x_{k+1}, \ldots, x_{k+\ell} \in \mathcal{F}$. Repeating the above reasoning we get

$$
\|x_{k+\ell} - x_0\|^2 \leq \|x_k - x_0\|^2 + C_2^2 \sum_{i=0}^{\ell-1} \alpha_{k+i}^2 - 2C_2 \sum_{i=0}^{\ell-1} \alpha_{k+i},
$$

(105)

and using the facts $\lim_{\ell \to \infty} \sum_{i=0}^{\ell-1} \alpha_{k+i}^2 < \infty$ and $\lim_{\ell \to \infty} \sum_{i=0}^{\ell-1} \alpha_{k+i} = \infty$, we would have, by (105), that $\|x_{k+\ell} - x_0\|^2 < 0$. Therefore, there exists an $\ell$ such that $x_{k+\ell} \in \mathcal{N}\mathcal{F}$.

$\square$

**References**

1. Arora, R., Basu, A., Mianjy, P., Mukherjee, A.: Understanding deep neural networks with rectified linear units. In: Bengio, Y., LeCun, Y. (eds.) International Conference on Learning Representations (ICLR) (2018)
2. Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing Robust Adversarial Examples. In: Bach, F. (ed.) International Conference on Machine Learning (ICML), pp. 284–293. PMLR, NY (2018)
3. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: Butler, K.R.B. (ed.) Symposium on Security and Privacy (SP), pp. 39–57. IEEE, New York (2017)
4. Carrara, F., Becarelli, R., Caldelli, R., Falchi, F., Amato, G.: Adversarial examples detection in features distance spaces. In: Ferrari, V., Sminchisescu, C., Hebert, M., Weiss, Y. (eds.) European Conference on Computer Vision (ECCV) Workshops (2018)
5. Chen, P.Y., Sharma, Y., Zhang, H., Yi, J., Hsieh, C.J.: EAD: elastic-net attacks to deep neural networks via adversarial examples. In: Zilberstein, S. (ed.) Proceeding of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
6. Chen, S., Carlini, N., Wagner, D.: Stateful detection of black-box adversarial attacks. In: X. Xing, Y.H. Lin (eds.) Proceedings of the ACM Workshop on Security and Privacy on Artificial Intelligence, pp. 30–39 (2020). https://doi.org/10.1145/3385003.3410925
7. Clarke, F.: Optimization and Nonsmooth Analysis. SIAM, Philadelphia (1990). https://doi.org/10.1137/1.9781611971309
8. Croce, F., Andriushchenko, M., Hein, M.: Provable robustness of ReLU networks via maximization of linear regions. In: K. Chaudhuri, M. Sugiyama (eds.) International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 2057–2066. PMLR (2019)
9. Ferguson, T.: A Course in Large Sample Theory. Springer, Berlin (1996)
10. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. stat 1050, 20 (2015)
11. Google: Tensorflow inception network. http://download.tensorflow.org/models/image/imagenet/inception-2015-12-05.tgz (2015)
12. Grosse, K., Manoharan, P., Papernot, N., Backes, M., McDaniel, P.D.: On the (statistical) detection of adversarial examples. CoRR abs/1702.06280 (2017). http://arxiv.org/abs/1702.06280
13. Guo, C., Rana, M., Cisse, M., van der Maaten, L.: Countering adversarial images using input transformations. In: Y. Bengio, Y. LeCun (eds.) International Conference on Learning Representations (ICLR) (2018)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: R. Bajcsy, F.F. Li, T. Tuytelaars (eds.) Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
15. Hein, M., Andriushchenko, M., Bitterwolf, J.: Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: L. Davis, P. Torr, S.C. Zhu (eds.) Conference on Computer Vision and Pattern Recognition (CVPR), pp. 41–50. IEEE/CVF (2019). https://doi.org/10.1109/CVPR.2019.00013
16. Jha, S., Jang, U., Jha, S., Jalaian, B.: Detecting adversarial examples using data manifolds. In: J. Shea (ed.) IEEE Military Communications Conference (MILCOM), pp. 547–552 (2018). https://doi.org/10.1109/MILCOM.2018.8599691
17. Krizhevsky, A.: Learning multiple layers of features from tiny images (2009). https://www.cs.toronto.edu/~kriz/cifar.html
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates Inc, Red Hook (2012)
19. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial machine learning at scale. In: Y. Bengio, Y. LeCun (eds.) International Conference on Learning Representations (ICLR) (2017)
20. Lee, S., Park, S., Lee, J.: Defensive denoising methods against adversarial attacks. In: C.J. Lin, H. Xiong (eds.) ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2018)
21. Liang, B., Li, H., Su, M., Li, X., Shi, W., Wang, X.: Detecting adversarial image examples in deep neural networks with adaptive noise reduction. IEEE Trans. Dependable Secur. Comput. 18(1), 72–85 (2021). https://doi.org/10.1109/TDSC.2018.2874243
22. Liu, Z., Liu, Q., Liu, T., Xu, N., Lin, X., Wang, Y., Wen, W.: Feature distillation: Dnn-oriented jpeg compression against adversarial examples. In: Davis, L., Torr, P., Zhu, S.C. (eds.) 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 860–868. IEEE (2019)
23. Lu, J., Issaranon, T., Forsyth, D.: Safetynet: Detecting and rejecting adversarial examples robustly. In: K. Ikeuchi, G. Medioni, M. Pelillo (eds.) Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 446–454 (2017)
24. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: Y. Bengio, Y. LeCun (eds.) International Conference on Learning Representations (ICLR) (2018)
25. Metzen, J.H., Genevieve, T., Fischer, V., Bischoff, B.: On detecting adversarial perturbations. In: Y. Bengio, Y. LeCun (eds.) International Conference on Learning Representations (ICLR) (2017)
26. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: R. Bajcsy, F.F. Li, T. Tuytelaars (eds.) Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2574–2582 (2016)
27. Papernot, N., Faghri, F., Carlini, N., Goodfellow, I., Feinman, R., Kurakin, A., Xie, C., Sharma, Y., Brown, T., Roy, A., Matyas, A., Behzadan, V., Hambardzumyan, K., Zhang, Z., Juang, Y.L., Li, Z., Sheatsley, R., Garg, A., Uesato, J., Gierke, W., Dong, Y., Berthelot, D., Hendricks, P., Rauber, J., Long, R.: Technical report on the cleverhans v2.1.0 adversarial examples library. arXiv preprint arXiv:1610.00768 (2018)
28. Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks. In: Locasto, M. (ed.) Symposium on Security and Privacy (SP), pp. 582–597. IEEE, New York (2016)
29. Prakash, A., Moran, N., Garber, S., DiLillo, A., Storer, J.: Protecting JPEG images against adversarial attacks. In: Data Compression Conference, pp. 137–146. IEEE (2018). https://doi.org/10.1109/DCC.2018.00022
30. Ren, K., Zheng, T., Qin, Z., Liu, X.: Adversarial attacks and defenses in deep learning. Engineering 6(3), 346–360 (2020). https://doi.org/10.1016/j.eng.2019.12.012
31. Roth, K., Kilcher, Y., Hofmann, T.: The odds are odd: A statistical test for detecting adversarial examples. In: E. Xing (ed.) International Conference on Machine Learning (ICML), pp. 5498–5507. PMLR (2019)
32. Samangouei, P., Kabkab, M., Chellappa, R.: Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In: Y. Bengio, Y. LeCun (eds.) International Conference on Learning Representations (ICLR) (2018)
33. Scholtes, S.: Introduction to Piecewise Differentiable Equations. SpringerBriefs in Optimization. Springer, New York (2012). https://doi.org/10.1007/978-1-4614-4340-7
34. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning—From Theory to Algorithms. Cambridge University Press, Cambridge (2014)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Y. Bengio, Y. LeCun (eds.) International Conference on Learning Representations (ICLR) (2015)
36. Sodolov, M.V., Zavriev, S.K.: Error stability properties of generalized gradient-type algorithms. JOTA 98, 663–680 (1998). https://doi.org/10.1023/A:1022680114518
37. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: Y. Bengio, Y. LeCun (eds.) International Conference on Learning Representations (ICLR) (2014)

38. Taran, O., Rezaeifar, S., Holotyak, T., Voloshynovskiy, S.: Defending against adversarial attacks by randomized diversification. In: L. Davis, P. Torr, S.C. Zhu (eds.) Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11226–11233 (2019)

39. Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., Gu, Q.: On the convergence and robustness of adversarial training. In: Chaudhuri, K., Salakhutdinov, R. (eds.) International Conference on Machine Learning (ICML), vol. 97, pp. 6586–6595. PMLR, Long Beach, California (2019)

40. Worzyk, N., Kramer, O.: Adversarials −1: Defending by attacking. In: M. Vellasco, P. Estevez (eds.) International Joint Conference on Neural Networks, IJCNN 2018, Rio de Janeiro, Brazil, July 8-13, 2018, pp. 1–8. IEEE (2018). https://doi.org/10.1109/IJCNN.2018.8489630

41. Worzyk, N., Kramer, O.: Properties of adv-1 - adversarials of adversarials. In: European Symposium on Artificial Neural Networks, ESANN 2018, Bruges, Belgium, April 25-27, 2018 (2018)

42. Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: Mitigating adversarial effects through randomization. In: Y. Bengio, Y. LeCun (eds.) International Conference on Learning Representations (ICLR) (2018)

43. Xu, H., Ma, Y., Liu, H., Deb, D., Liu, H., Tang, J., Jain, A.K.: Adversarial attacks and defences in images, graphs and text: a review. Int. J. Autom. Comput. 17(2), 151–178 (2020). https://doi.org/10.1007/s11633-019-1211-x

44. Yin, X., Kolouri, S., Rohde, G.K.: Divide-and-conquer adversarial detection. CoRR abs/1905.11475 (2019). http://arxiv.org/abs/1905.11475

45. Zheng, Z., Hong, P.: Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks. In: S. Bengio, H. Wallach (eds.) Proceedings of the International Conference on Neural Information Processing Systems, NIPS, pp. 7924–7933. Curran Associates Inc., USA (2018)

46. Zhou, Q., Zhang, R., Wu, B., Li, W., Mo, T.: Detection by attack: detecting adversarial samples by undercover attack. In: Chen, L., Li, N., Liang, K., Schneider, S. (eds.) Computer Security - ESORICS 2020, pp. 146–164. Springer, Cham (2020)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.