On Adversarial Examples and Stealth Attacks in Artificial Intelligence Systems

1st Ivan Y. Tyukin
School of Mathematics and Actuarial Science
University of Leicester
and *Norwegian University of Science and Technology*
and *Saint-Petersburg State Electrotechnical University*

Leicester, LE1 7RH, UK
and Trondheim, Norway
and Saint-Petersburg, Russia
I.Tyukin@le.ac.uk

2nd Desmond J. Higham
School of Mathematics
University of Edinburgh

Edinburgh, EH9 3FD, UK
d.j.higham@ed.ac.uk

3rd Alexander N. Gorban
School of Mathematics and Actuarial Science
University of Leicester
and *Lobachevsky University*

Leicester, LE1 7RH, UK
and Nizhni Novgorod, Russia
a.n.gorban@le.ac.uk

***Abstract*—In this work we present a formal theoretical framework for assessing and analyzing two classes of malevolent action towards generic Artificial Intelligence (AI) systems. Our results apply to general multi-class classifiers that map from an input space into a decision space, including artificial neural networks used in deep learning applications. Two classes of attacks are considered. The first class involves adversarial examples and concerns the introduction of small perturbations of the input data that cause misclassification. The second class, introduced here for the first time and named *stealth attacks*, involves small perturbations to the AI system itself. Here the perturbed system produces whatever output is desired by the attacker on a specific small data set, perhaps even a single input, but performs as normal on a validation set (which is unknown to the attacker).**

We show that in both cases, i.e., in the case of an attack based on adversarial examples and in the case of a stealth attack, the dimensionality of the AI's decision-making space is a major contributor to the AI's susceptibility. For attacks based on adversarial examples, a second crucial parameter is the absence of local concentrations in the data probability distribution, a property known as Smeared Absolute Continuity. According to our findings, robustness to adversarial examples requires either (a) the data distributions in the AI's feature space to have concentrated probability density functions or (b) the dimensionality of the AI's decision variables to be sufficiently small. We also show how to construct stealth attacks on high-dimensional AI systems that are hard to spot unless the validation set is made exponentially large.

***Index Terms*—Adversarial examples, adversarial attacks, stochastic separation theorems, artificial intelligence, machine learning**

NOTATION

- $\mathbb{R}$ denotes the field of real numbers, $\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} |\ x \geq 0\}$, and $\mathbb{R}^n$ stands for the n-dimensional linear real vector space;
- $\mathbb{N}$ denotes the set of natural numbers;
- symbols $\boldsymbol{x} = (x_1, \ldots, x_n)$ will denote elements of $\mathbb{R}^n$;
- $(\boldsymbol{x}, \boldsymbol{y}) = \sum_k x_k y_k$ is the inner product of $\boldsymbol{x}$ and $\boldsymbol{y}$, and $\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x}, \boldsymbol{x})}$ is the standard Euclidean norm in $\mathbb{R}^n$;
- $\mathbb{B}_n$ denotes the unit ball in $\mathbb{R}^n$ centered at the origin:
$$\mathbb{B}_n = \{\boldsymbol{x} \in \mathbb{R}^n |\ \|\boldsymbol{x}\| \leq 1\};$$
- $\mathbb{B}_n(r, \boldsymbol{y})$ stands for the ball in $\mathbb{R}^n$ of radius $r > 0$ centered at $\boldsymbol{y}$:
$$\mathbb{B}_n(r, \boldsymbol{y}) = \{\boldsymbol{x} \in \mathbb{R}^n |\ \|\boldsymbol{x} - \boldsymbol{y}\| \leq r\};$$
- $\mathbb{S}_{n-1}(r, \boldsymbol{y})$ stands for the $n-1$ sphere in $\mathbb{R}^n$ that is centered at $\boldsymbol{y}$ and has a radius r:
$$\mathbb{S}_{n-1}(r, \boldsymbol{y}) = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x} - \boldsymbol{y}\| = r\};$$
- V_n is the n-dimensional Lebesgue measure, and $V_n(\mathbb{B}_n)$ is the volume of unit n-ball;

I. BACKGROUND AND MOTIVATION

The application of Artificial Intelligence (AI) and Machine Learning methods has produced numerous success stories in recent years [1], [2], [3]. Examples where it has been reported that human levels of performance can be matched or exceeded include identification of breast cancer [4], detection of objects hidden from view [5], mastery of board games [6], optimization of new imaging techniques [7], and development of systems for autonomous self-driving cars [8].

Existing breakthroughs are clearly stimulating further research and encouraging the broad deployment of such systems in practice. However, in a field of research where, for example, traffic "Stop" signs on the roadside can be misinterpreted as speed limit signs when minimal graffiti is added [9], many commentators are asking whether current solutions are sufficiently robust, resilient, and trustworthy; and how such issues should be quantified and addressed. Marcus [10] outlines ten concerns about the current state of deep learning, one of which is that "Deep learning thus far works well as an approximation, but its answers often cannot be fully trusted."

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Examples of undesirable, unintended, and unexpected behavior of otherwise sophisticated deep learning systems raising further questions around the issues of resilience and trustworthiness of data-driven AI systems have been extensively reported and discussed in the literature on *adversarial images* [11], [12].

Adversarial images arise when specially chosen perturbations, effectively imperceptible to the human eye, cause misclassification in an AI system, or, indeed, simultaneously across a range of AI systems. The existence of adversarial images illustrates the risks associated with the deployment of data-driven neural network-based decision-making and raises important questions around responsible research and innovation (RRI) [13], [14]. There are now many constructive approaches for the generation of adversarial attacks; for example, [15], [16], [17], [18], [19], [20], [21]. On the other hand, techniques that aim to identify or guard against such attacks have also been developed; for example, [22], [23], [24], [25], [26], [27], leading to a version of conflict escalation where attack and defence strategies become increasingly ingenious.

Against this backdrop, the work in [28] looks at a higher-level question: are there fundamental reasons that make adversarial examples difficult to thwart? The authors developed arguments based on various versions of the isoperimetric inequality to determine a set of conditions under which adversarial examples occur with probability close to one in a very general setting (see [28] for further details).

In this work, we use a different set of tools to derive alternative conditions under which the existence of adversarial examples is essentially unavoidable for general classifiers. In addition, we introduce a second type of risk, relating to malicious, targeted behavior that we refer to as a *stealth attack*. In this scenario, an attacker (who may, for example, be a mischievous, disgruntled, malevolent or corrupt member of a large software development team) has access to the actual code implementing the AI system. Such an attacker is capable of changing, adding or replacing a single or a small number of nodes with the aim of altering the behavior of the system. To evade detection, the perturbed system must show little if any deviation from the nominal system's expected performance on some finite verification set $\mathcal{V}$, making the attack transparent to the AI's owners and users. At the same time, on a data set or even a single input $\boldsymbol{x}'$ that is known only to the attacker, the system must generate a response which the attacker desires but which is different from the nominal system's output. (So, for example, there may be a particular image whose classification the attacker wishes to override.)

If the verification set $\mathcal{V}$ is available to the attacker and the attacker is allowed to change a significant portion of the nominal AI system (e.g., parameters and connections of neurons in the network) then it is technically plausible and operationally simple to execute such an attack by re-training. Large systems in which the total number of parameters of the altered part exceeds the cardinality of $\mathcal{V} \cup \boldsymbol{x}'$ are particularly vulnerable to alterations of this type. Indeed, it is well-known that $n+1$ generic points in $\mathbb{R}^n$ are linearly separable. Experiments in [29] showed that simple shallow yet sufficiently large neural networks may achieve perfect finite sample expressivity as soon as the number of parameters exceeds the number of data points (cf. [30]).

We have in mind the more challenging case when i) the set $\mathcal{V}$ and its cardinality is unknown to the attacker and ii) the attacker may change only a single element (albeit with its weights and parameters) in the system.

II. General Framework

We will study both adversarial examples and stealth attacks in a single, generic, setting. We suppose that the system is modeled by a map

$$\mathcal{F} : \mathbb{B}_n \to \mathbb{R}. \tag{1}$$

The map may represent a multi-class classifier, implemented e.g., by a neural network, defined on a set $\Phi \subset \mathbb{B}_n$ of the feature vectors $\boldsymbol{x} \in \Phi$. The nature and the origin of the feature vectors and the map itself are not important for our analysis. The map can be viewed as a transformation modelled by one or a few fully connected layers inside a deep neural network; it may also describe the entire input-output behavior of the system. What is important, however, is that the feature vectors $\boldsymbol{x}$ are elements of a high-dimensional vector space $\mathbb{R}^n$.

Using this model, we formally analyze the inevitability of both adversarial examples and stealth attacks. With respect to the problem of adversarial examples (Theorem 1), we formulate a relationship between a given classifier and statistical properties of the data (Assumption 1) that leads inevitably to the existence of adversarial examples. A key element of these conditions is the Smeared Absolute Continuity (SmAC) property of the probability distribution introduced in [31]. A similar condition is imposed in [28] in the form of the assumption of an upper bound for the probability density function. Here, however, we do not require that the latter property holds for the entire distribution. If n is sufficiently large then for the existence of an $(\varepsilon + \Delta)$-adversarial example (ε may be chosen arbitrarily small) it is sufficient that

i) the SmAC condition holds in some ball of non-zero measure, and
ii) for any point on the boundary of that ball there is an element of a different class within distance Δ.

We also provide an explicit estimate of the dimension n at which such examples become probable.

The new concept of a stealth attack, where an opponent modifies a small part of the backbone system in a way that impacts only specific inputs, is formalized (9). Our results show that stealth attacks are surprisingly easy to construct for large enough n. In particular, we find that if the cardinality M of the verification set $\mathcal{V}$ is smaller than 2^n then these attacks can be produced by a modification of a single node in the system and without any knowledge of the verification data (Theorem 2).

The rest of the manuscript is organized as follows: in Section III we quantify probabilities of adversarial examples for a broad class of data distributions satisfying the SmAC

condition, Section IV presents conditions and possible scenarios for stealth attacks, and Section V concludes the paper.

III. Adversarial examples

Consider a standard multi-class classification problem in which each element $\boldsymbol{x} \in \Phi$ is associated with a label $\ell \in \mathcal{L}$ from a finite set $\mathcal{L}$ of labels. We assume that the pairs $(\boldsymbol{x}, \ell)$ are drawn from some probability distribution with the corresponding probability density function:

$$p: \ \mathbb{B}_n \times \mathcal{L} \to \mathbb{R}_{\geq 0}.$$

The distribution as well as the probability density functions are supposed to be *unknown* but their existence is assumed. The backbone/legacy AI system is hence a classifier which for a given $\boldsymbol{x} \in \Phi$ aims at predicting its label ℓ.

Definition 1: For the given classification map $\mathcal{F}$, an element $\boldsymbol{x} \in \mathbb{B}_n$ admits a δ-adversarial example $\boldsymbol{y}(\boldsymbol{x})$ if

$$\mathcal{F}(\boldsymbol{x}) \neq \mathcal{F}(\boldsymbol{y}(\boldsymbol{x})) \text{ and } \|\boldsymbol{x} - \boldsymbol{y}(\boldsymbol{x})\| \leq \delta, \ \boldsymbol{y}(\boldsymbol{x}) \in \mathbb{B}_n.$$

In what follows we will determine a set of conditions on the classifier map $\mathcal{F}$ and the data distribution for which adversarial examples exist and the probability of their occurrence is non-zero and sometimes could be even exponentially "close" to 1 with respect to n.

Let A be an element of the label set $\mathcal{L}$. We denote

$$p_A(\boldsymbol{x}) = p(\boldsymbol{x}|\ell = A), \ P(A) = \int_{\mathbb{B}_n} p(\boldsymbol{x}, A) d\boldsymbol{x},$$
$$p(\boldsymbol{x}|\ell = A) = \frac{p(\boldsymbol{x}, A)}{P(A)}. \tag{2}$$

Assumption 1: There exists a label $A \in \mathcal{L}$ and an associated set $\mathcal{C}_A \subset \mathbb{B}_n$, a number $r_A \in (0, 1)$, a vector $\boldsymbol{x}_A \in \mathbb{B}_n$, a positive constant $C > 0$, and a number $\nu \in (0, 1]$ such that

A1) The set $\mathcal{C}_A$ is contained in $\mathbb{B}_n(r_A, \boldsymbol{x}_A)$.

A2) $\mathcal{F}(\boldsymbol{x}) = A$ for all $\boldsymbol{x} \in \mathcal{C}_A$, and there is a $\Delta > 0$ such that for any $\boldsymbol{x} \in \mathbb{S}_{n-1}(r_A, \boldsymbol{x}_A)$ there exists a $\boldsymbol{y}(\boldsymbol{x})$:

$$\mathcal{F}(\boldsymbol{y}(\boldsymbol{x})) \neq A, \quad \|\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{x}\| \leq \Delta.$$

A3) The probability density function p_A satisfies

$$p_A(\boldsymbol{x}) \leq \frac{C}{V_n(\mathbb{B}_n)} \frac{1}{r_A^n} \text{ for all } \boldsymbol{x} \in \mathbb{B}_n(r_A, \boldsymbol{x}_A),$$
$$\text{and} \int_{\mathcal{C}_A} p_A(\boldsymbol{x}) d\boldsymbol{x} \geq \nu > 0. \tag{3}$$

Conditions $A1$ – $A3$ in Assumption 1 formalize a relationship between the given classification map $\mathcal{F}$ and statistical properties of the pair $(\boldsymbol{x}, \ell)$ which, as we shall see later, lead to the risk of emergence of adversarial examples. In particular, Assumption 1 ensures that

- The probability that the event $\boldsymbol{x} \in \mathcal{C}_A, \ell = A$ occurs is at least $P(A)\nu$, and the corresponding conditional probability density p_A satisfies a form of the Smeared Absolute Continuity condition in the domain $\mathcal{C}_A$ [31] (condition $A3$).
- Any $\boldsymbol{x}$ from the set $\mathcal{C}_A$ is interpreted as an element of class A by the map $\mathcal{F}$, and a Δ-neighborhood of any element $\boldsymbol{x}$ on the boundary of the set $\mathbb{B}_n(r_A, \boldsymbol{x}_A) \supset \mathcal{C}_A$ contains at least one element $\boldsymbol{y}(\boldsymbol{x})$ to which the map $\mathcal{F}$ assigns a label that is different from A (condition $A2$). The latter part of the condition will obviously hold if

  $$\mathcal{F}(\boldsymbol{x}) \neq A \text{ for all}$$
  $$\boldsymbol{x} \in \mathbb{B}_n(r_A + \Delta, \boldsymbol{x}_A) \cap \mathbb{B}_n \setminus \mathbb{B}_n(r_A, \boldsymbol{x}_A).$$
- A non-empty set for which the above properties hold exists (the set $\mathcal{C}_A$) and is in the interior of some n-ball in $\mathbb{B}_n$ (condition $A1$).

Under these conditions the following statement holds.

Theorem 1: Consider a classification map $\mathcal{F}$ and a probability distribution with probability density function p satisfying Assumption 1. Let a sample $(\boldsymbol{x}, \ell)$ be drawn from this distribution and let ε be chosen arbitrarily in $(0, r_A)$. Then the probability that $\boldsymbol{x}$ admits an $(\varepsilon + \Delta)$-adversarial example is at least

$$P(A) \max \left\{ \nu - C \left(1 - \frac{\varepsilon}{r_A}\right)^n, 0 \right\}. \tag{4}$$

Proof of Theorem 1 Let us fix an $0 < \varepsilon < r_A$ and let P^* be the probability of the event

$$\boldsymbol{x} \in \mathbb{B}_n(\boldsymbol{x}_A, r_A) \setminus B_n(\boldsymbol{x}_A, r_A - \varepsilon), \ \ell = A.$$

Then according to Assumption 1 (condition $A2$) and Definition 1, the probability that a $\Delta + \varepsilon$ adversarial example exists for the given classifier is P^*. The probability P^* can be estimated as

$$P^* = P(A) P(\boldsymbol{x} \in \mathbb{B}_n(\boldsymbol{x}_A, r_A) \setminus B_n(\boldsymbol{x}_A, r_A - \varepsilon) | \ell = A).$$

Consider

$$P(\boldsymbol{x} \in B_n(\boldsymbol{x}_A, r_A - \varepsilon) | \ell = A)$$
$$= \int_{B_n(\boldsymbol{x}_A, r_A - \varepsilon)} p(\boldsymbol{x}|\ell = A) d\boldsymbol{x}$$
$$= \int_{B_n(\boldsymbol{x}_A, r_A - \varepsilon)} p_A(\boldsymbol{x}) d\boldsymbol{x}.$$

According to (2) and (3),

$$\int_{B_n(\boldsymbol{x}_A, r_A - \varepsilon)} p_A(\boldsymbol{x}) d\boldsymbol{x} \leq \int_{B_n(\boldsymbol{x}_A, r_A - \varepsilon)} \frac{C}{V_n(\mathbb{B}_n) r_A^n} d\boldsymbol{x}$$
$$C \frac{(r_A - \varepsilon)^n}{r_A^n} = C \left(1 - \frac{\varepsilon}{r_A}\right)^n.$$

Using conditions $A1$ and $A3$ from Assumption 1, we can obtain the following estimate

$$P(\boldsymbol{x} \in \mathbb{B}_n(\boldsymbol{x}_A, r_A) \setminus B_n(\boldsymbol{x}_A, r_A - \varepsilon) | \ell = A) =$$
$$P(\boldsymbol{x} \in \mathbb{B}_n(\boldsymbol{x}_A, r_A) | \ell = A) -$$
$$P(B_n(\boldsymbol{x}_A, r_A - \varepsilon) | \ell = A) \geq \nu - C \left(1 - \frac{\varepsilon}{r_A}\right)^n.$$

The value of P^* can now be estimated from below as

$$P(A) \left(\nu - C \left(1 - \frac{\varepsilon}{r_A}\right)^n \right),$$

and hence the statement follows $\square$.

Using the well-known inequality

$$(1-x)^{1/x} < e^{-1}, \quad x \in (0,1),$$

the following exponential lower bound estimate for (4) holds:

$$P(A)\max\left\{\nu - C\exp\left(-\frac{n\varepsilon}{r_A}\right), 0\right\}.$$

Remark 1: According to Theorem 1, if the classifier and the data probability distribution satisfy Assumption 1, then $(\varepsilon + \Delta)$-adversarial examples are expected to occur if the dimensionality n of the feature space is sufficiently large:

$$n > (\log\nu - \log C)\left[\log\left(1 - \frac{\varepsilon}{r_A}\right)\right]^{-1}. \tag{5}$$

Moreover, if C is independent of n, then the probability that the data sample admits a $(\varepsilon + \Delta)$-adversarial example approaches $P(A)\nu$ exponentially fast with dimension n.

Remark 2: For classifiers operating in dimensions satisfying (5) one can now easily derive a bound on the probability of occurrence of an $(\varepsilon+\Delta)$- adversarial example in a sample of N i.i.d. random data points. In particular, under the assumptions of Theorem 1, the probability that at least one $(\varepsilon + \Delta)$-adversarial example occurs this sample is not smaller than

$$1 - \left[1 - P(A)\left(\nu - C\left(1 - \frac{\varepsilon}{r_A}\right)^n\right)\right]^N.$$

IV. Stealth Attacks to the backbone AI

The susceptibility of decision-making in AI systems operating in high-dimensional space to small adversarial perturbations of the data is just one facet of the larger topic of robust, resilient, and ultimately verifiable AI performance. In this subsection we formally define and study the related but distinct issue of stealth attacks.

To set-up our framework, consider the classification map (1)

$$\mathcal{F} : \mathbb{B}_n \to \mathbb{R}$$

modelling the backbone AI system. In addition to this map, consider

$$\begin{aligned} &\mathcal{F}_a : \mathbb{B}_n \times \Theta \to \mathbb{R} \\ &\mathcal{F}_a(\cdot, \boldsymbol{\theta}) = \mathcal{F}(\cdot) + \mathfrak{A}(\cdot, \boldsymbol{\theta}), \end{aligned} \tag{6}$$

where the term

$$\mathfrak{A} : \mathbb{B}_n \times \Theta \to \mathbb{R}$$

models a stealth attack on the original backbone system $\mathcal{F}$, and $\Theta \subset \mathbb{R}^m$ is an associated set of parameters.

A case of significant practical interest arises when the term $\mathfrak{A}$ can be expressed using just a single Rectified Linear Unit (ReLU function), [32] (see, for example, [33] or [34] for information regarding basic nonlinear elements)

$$\begin{aligned} &\mathfrak{A}(\cdot, (\boldsymbol{w}, b)) = D\mathrm{ReLU}((\cdot, \boldsymbol{w}) - b), \\ &\mathrm{ReLU}(s) = \max\{s, 0\} \end{aligned} \tag{7}$$

or a sigmoid

$$\begin{aligned} &\mathfrak{A}(\cdot, (\boldsymbol{w}, b)) = D\sigma((\cdot, \boldsymbol{w}) - b), \\ &\sigma(s) = \frac{1}{1 + \exp(-s)}, \end{aligned} \tag{8}$$

with $D > 0$ being a positive constant. It is convenient to denote

$$\mathfrak{A}(\cdot, (\boldsymbol{w}, b)) = Dg((\cdot, \boldsymbol{w}) - b),$$

where the function g is either ReLU or sigmoid, depending on the case, and $(\boldsymbol{w}, b) = \boldsymbol{\theta}$ are its relevant parameters. We are now ready to formally introduce the following stealth attack problem

Problem 1 (Stealth Attack on $\mathcal{F}$): Consider a classification map $\mathcal{F}$ defined by (1) and modelling a backbone AI. Suppose that an owner of the AI system or a network has a finite validation or verification set

$$\mathcal{V} \subset \mathbb{B}_n.$$

The validation set $\mathcal{V}$ is kept secret and is assumed to be *unknown* to an attacker. The cardinality of $\mathcal{V}$ is bounded from above by some constant M, and this bound is known to the attacker.

The attacker seeks to modify the map $\mathcal{F}$ and replace it by $\mathcal{F}_a$ constructed in accordance with (6), (7) or (6), (8) and such that for some given $\varepsilon > 0$, $\Delta > 0$ and an element $\boldsymbol{x}' \in \mathbb{B}_n$, known to the attacker but unknown to the owner of the map $\mathcal{F}$, the following properties hold:

$$\begin{aligned} &\|\mathcal{F}(\boldsymbol{x}) - \mathcal{F}_a(\boldsymbol{x}, (\boldsymbol{w}, b))\| \le \varepsilon \ \forall \ \boldsymbol{x} \in \mathcal{V} \\ &\mathcal{F}_a(\boldsymbol{x}', (\boldsymbol{w}, b)) = \mathcal{F}(\boldsymbol{x}') + \Delta. \end{aligned} \tag{9}$$

In words, the stealth attack has an imperceptible effect on the validation set, since $\varepsilon > 0$ can be made arbitrarily small, but makes the desired modification of the backbone AI (with arbitrarily large $\Delta > 0$) for the target input $\boldsymbol{x}'$.

We say that $\mathfrak{A}$ is a solution of this problem if it satisfies (9). The next statement provides an efficient mechanism for constructing such solutions.

Theorem 2: Consider Problem 1, and let $\boldsymbol{x}'$ be a vector that is randomly drawn from the equidistribution in $\mathbb{B}_n$. Then the probability that

$$\begin{aligned} &\mathfrak{A}(\cdot, (\kappa\boldsymbol{x}', b)) = Dg((\cdot, \kappa\boldsymbol{x}') - b), \\ &b = \kappa\left(\frac{1+\gamma}{2}\right)\|\boldsymbol{x}'\|^2, \end{aligned} \tag{10}$$

where κ and D are chosen so that

$$\begin{aligned} &Dg\left(-\kappa\frac{1-\gamma}{2}\|\boldsymbol{x}'\|^2\right) \le \varepsilon \text{ and} \\ &Dg\left(\kappa\frac{1-\gamma}{2}\|\boldsymbol{x}'\|^2\right) \ge \Delta, \ \gamma \in (0,1), \end{aligned} \tag{11}$$

is a solution of Problem 1 is at least

$$1 - M\left(\frac{1}{2\gamma}\right)^n.$$

Proof of Theorem 2. Let us pick $\gamma \in (0,1)$ and let $\boldsymbol{x}'$ be such that

$$\gamma(\boldsymbol{x}', \boldsymbol{x}') = \gamma \|\boldsymbol{x}'\|^2 > (\boldsymbol{x}', \boldsymbol{x}_i), \text{ for all } \boldsymbol{x}_i \in \mathcal{V}.$$

Set

$$\begin{aligned} \boldsymbol{w} &= \kappa \boldsymbol{x}', \ \kappa > 0, \\ b &= \kappa \left(\frac{1+\gamma}{2} \right) \|\boldsymbol{x}'\|^2, \end{aligned}$$

and observe that

$$\mathfrak{A}(\cdot, (\boldsymbol{w}, b)) = Dg\left(\kappa \left((\cdot, \boldsymbol{x}') - \left(\frac{1+\gamma}{2} \right) \|\boldsymbol{x}'\|^2 \right) \right),$$

where we recall that g is either ReLU or sigmoid. Consider

$$\|\mathcal{F}(\boldsymbol{x}_i) - \mathcal{F}_a(\boldsymbol{x}_i, (\boldsymbol{w}, b))\| = |\mathfrak{A}(\cdot, (\boldsymbol{w}, b))|.$$

Since the function g is monotone,

$$|\mathfrak{A}(\boldsymbol{x}_i, (\boldsymbol{w}, b))| \leq Dg\left(-\kappa \left(\frac{1-\gamma}{2} \|\boldsymbol{x}'\|^2 \right) \right) \ \forall \ \boldsymbol{x}_i \in \mathcal{V}.$$

Denote

$$z = \frac{1-\gamma}{2} \|\boldsymbol{x}'\|^2$$

and pick the values of D and κ so that

$$Dg(-\kappa z) \leq \varepsilon \text{ and } Dg(\kappa z) \geq \Delta.$$

Given that $\mathrm{ReLU}(s) = 0$ for all $s \leq 0$ and that the sigmoidal function is strictly increasing with $g(0) \neq 0$, such choice is always possible.

Finally, let $\boldsymbol{x}'$ be drawn from the equidistribution in $\mathbb{B}_n$. Then the probability that

$$\gamma(\boldsymbol{x}', \boldsymbol{x}') > (\boldsymbol{x}', \boldsymbol{x}_i), \text{ for all } \boldsymbol{x}_i \in \mathcal{V}$$

is at least

$$1 - M\left(\frac{1}{2\gamma} \right)^n.$$

(see Proposition 1 of [31]). This completes the proof. □

Remark 3: If $g = \mathrm{ReLU}$ then the value of ε in Theorem 2 can be set to 0 which in turn implies that the stealth map $\mathcal{F}_a$ is indistinguishable from $\mathcal{F}$ on the verification set $\mathcal{V}$:

$$\mathcal{F}_a(\boldsymbol{x}) = \mathcal{F} \ \forall \ \boldsymbol{x} \in \mathcal{V}.$$

Remark 4: The statement of Theorem 2 can be adjusted to include the class of functions g:

$$\lim_{s \to -\infty} g(s) = 0, \ \lim_{s \to \infty} g(s) = 0, \ g(0) = 1.$$

In this case the value of b in (10) should change to

$$b = \kappa \|\boldsymbol{x}'\|^2$$

and condition (11) will need to become

$$Dg\left(-\kappa(1-\gamma)\|\boldsymbol{x}'\|^2 \right) \leq \varepsilon \text{ and } D \geq \Delta, \ \gamma \in (0,1).$$

This extends the results to bell-shaped functions g such as the Gaussian and also opens possibilities to use general sigmoidal functions σ to construct such g:

$$g(s) = \frac{\sigma(s) - \sigma(s+a)}{\sigma(0) - \sigma(0+a)}, \quad \sigma(0) - \sigma(0+a) \neq 0.$$

V. Conclusion

In this work we set up a formal framework for analyzing two classes of malevolent action towards generic AI systems. These systems include neural networks but generally could be of a rather arbitrary type. The first class, adversarial examples, concerns small perturbations of the input data that cause misclassification. Such perturbations have been widely studied in recent years, mostly from an empirical perspective. The second class, introduced here for the first time and named stealth attacks, involve small perturbations to the AI system itself. Here the perturbed system produces whatever output is desired by the attacker on a specific small data set, perhaps even a single input, but performs as normal on a validation set (which is unknown to the attacker).

In both cases, we identified the dimensionality of the AI's decision-making space as a major factor in its susceptibility.

With regard to adversarial examples, a second crucial aspect influencing the risk of adversarial attacks is the absence or presence of local concentrations in the data probability distribution (Smeared Absolute Continuity condition). According to our findings, a robust system should either have concentrated probability density functions or its dimensionality must be reduced to avoid the effects of the measure concentration.

Concerning stealth attacks on the backbone AI, we note that systems with ReLU activation functions are particularly prone to adversarial modifications which are hard to spot without resorting to exponentially large in dimension, 2^n, verification sets. Single-node adversarial alterations involving differentiable activation functions may need to have large Lipschitz constants (i.e., the values of κ, D in Theorem 2). Lipschitz constants calculated over a data sample have been used extensively as an indicator of network quality [33] (the smaller the better). Here we have shown that these are not only mere quality indicators; large Lipschitz constants in networks and systems with differentiable activation functions are also consistent with susceptibility to stealth attack.

Many relevant questions, however, remain. In particular, we did not consider here probabilities of noise-induced misclassifications. We also did not try to produce the tightest possible probability estimates. Addressing these, and related issues, will be the focus of future work.

Acknowledgement

Desmond J. Higham was supported by EP/M00158X/1 from the EPSRC/RCUK Digital Economy Programme and EPSRC Programme Grant EP/P020720/1; Alexander N. Gorban and Ivan Y. Tyukin were supported the Ministry of Science and Higher Education of Russian Federation (Project No. 14.Y26.31.0022).

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[2] S. Rogers and M. Girolami, *A First Course in Machine Learning*, 2nd ed. London: CRC Press, 2016.

[3] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[4] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. C. Corrado, A. Darzi *et al.*, "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[5] P. Caramazza, A. Boccolini, D. Buschek, M. Hullin, C. F. Higham, R. Henderson, R. Murray-Smith, and D. Faccio, "Neural network identification of people hidden from view with a single-pixel, single-photon detector," *Scientific reports*, vol. 8, no. 1, pp. 1–6, 2018.

[6] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 2529, pp. 484–489, 2016.

[7] C. F. Higham, R. Murray-Smith, M. J. Padgett, and M. P. Edgar, "Deep learning for real-time single-pixel video," *Scientific Reports*, vol. 8, p. 2369, 2018.

[8] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[9] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song, "Robust physical-world attacks on machine learning models," *CoRR*, vol. abs/1707.08945, 2017. [Online]. Available: http://arxiv.org/abs/1707.08945

[10] G. Marcus, "Deep learning: A critical appraisal," *arXiv:1801.00631 [cs.AI]*, 2018.

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

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[13] P. Grindrod, "Beyond privacy and exposure: ethical issues within citizen-facing analytics," *Phil. Trans. of the Royal Society A*, vol. 374, p. 2083, 2016.

[14] J. H. Davenport, "The debate about algorithms," *Mathematics Today*, p. 162, August 2017.

[15] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[16] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 2574–2582. [Online]. Available: https://doi.org/10.1109/CVPR.2016.282

[17] A. Modas, S. Moosavi-Dezfooli, and P. Frossard, "Sparsefool: A few pixels make a big difference," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 9087–9096.

[18] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: a query-efficient black-box adversarial attack via random search," *CoRR*, vol. abs/1912.00049, 2019. [Online]. Available: http://arxiv.org/abs/1912.00049

[19] J. Lu, H. Sibai, and E. Fabry, "Adversarial examples that fool detectors," *CoRR*, vol. abs/1712.02494, 2017. [Online]. Available: http://arxiv.org/abs/1712.02494

[20] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2017, Abu Dhabi, United Arab Emirates, April 2-6, 2017*, R. Karri, O. Sinanoglu, A. Sadeghi, and X. Yi, Eds. ACM, 2017, pp. 506–519. [Online]. Available: https://doi.org/10.1145/3052973.3053009

[21] J. Su, D. V. Vargas, and S. Kouichi, "One pixel attack for fooling deep neural networks," *arXiv:1710.08864 [cs.LG]*, 2017.

[22] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, "Parseval networks: Improving robustness to adversarial examples," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 854–863.

[23] F. Croce, M. Andriushchenko, and M. Hein, "Provable robustness of ReLU networks via maximization of linear regions," *CoRR*, vol. abs/1810.07481, 2018. [Online]. Available: http://arxiv.org/abs/1810.07481

[24] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," *arXiv preprint arXiv:1802.00420*, 2018.

[25] F. Croce and M. Hein, "Provable robustness against all adversarial lp-perturbations for p greater than or equal to one," in *International Conference on Learning Representations*, 2020.

[26] I. J. Goodfellow, P. D. McDaniel, and N. Papernot, "Making machine learning robust against adversarial inputs," *Commun. ACM*, vol. 61, no. 7, pp. 56–66, 2018. [Online]. Available: https://doi.org/10.1145/3134599

[27] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[28] A. Shafahi, W. Huang, C. Studer, S. Feizi, and T. Goldstein, "Are adversarial examples inevitable?" *International Conference on Learning Representations (ICLR)*, 2019.

[29] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *5th International Conference on Learning Representations*, 2017.

[30] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE transactions on electronic computers*, no. 3, pp. 326–334, 1965.

[31] A. Gorban, A. Golubkov, B. Grechuk, E. Mirkes, and I. Tyukin, "Correction of AI systems by linear discriminants: Probabilistic foundations," *Information Sciences*, vol. 466, pp. 303–322, 2018.

[32] R. Hahnloser, R. Sarpeshkar, M. Mahowald, R. Douglas, and H. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, no. 6789, pp. 947–951, 2000.

[33] A. Gorban and D. Rossiev, *Neural networks on personal computer*. Novosibirsk: Nauka (RAN), 1996.

[34] C. F. Higham and D. J. Higham, "Deep learning: An introduction for applied mathematicians," *SIAM Review*, vol. 61, pp. 860–891, 2019.