

Chapter 16

Robustness and adversarial examples

How sensitive is the output of a neural network to small changes in its input? Real-world observations of trained neural networks often reveal that even barely noticeable modifications of the input can lead to drastic variations in the network's predictions. This intriguing behavior was first documented in the context of image classification in [222].

Figure 16.1 illustrates this concept. The left panel shows a picture of a panda that the neural network correctly classifies as a panda. By adding an almost imperceptible amount of noise to the image, we obtain the modified image in the right panel. To a human, there is no visible difference, but the neural network classifies the perturbed image as a wombat. This phenomenon, where a correctly classified image is misclassified after a slight perturbation, is termed an *adversarial example*.

In practice, such behavior is highly undesirable. It indicates that our learning algorithm might not be very reliable and poses a potential security risk, as malicious actors could exploit it to trick the algorithm. In this chapter, we describe the basic mathematical principles behind adversarial examples and investigate simple conditions under which they might or might not occur. For simplicity, we restrict ourselves to a binary classification problem but note that the main ideas remain valid in more general situations.


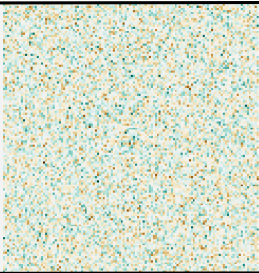

		$+ 0.01 \times$ 	$=$ 
Human:	Panda	Barely visible noise	Still a panda
NN classifier:	Panda (high confidence)	Flamingo (low confidence)	Wombat (high confidence)

Figure 16.1: Sketch of an adversarial example.

16.1 Adversarial examples

Let us start by formalizing the notion of an adversarial example. We consider the problem of assigning a label $y \in \{-1, 1\}$ to a vector $\mathbf{x} \in \mathbb{R}^d$. It is assumed that the relation between \mathbf{x} and y is described by a distribution \mathcal{D} on $\mathbb{R}^d \times \{-1, 1\}$. In particular, for a given \mathbf{x} , both values -1 and 1 could have positive probability, i.e. the label is not necessarily deterministic. Additionally, we let

$$D_{\mathbf{x}} := \{\mathbf{x} \in \mathbb{R}^d \mid \exists y \text{ s.t. } (\mathbf{x}, y) \in \text{supp}(\mathcal{D})\},$$

and refer to $D_{\mathbf{x}}$ as the **feature support**.

Throughout this chapter we denote by

$$g: \mathbb{R}^d \rightarrow \{-1, 0, 1\}$$

a fixed so-called *ground-truth classifier*, satisfying¹

$$\mathbb{P}[y = g(\mathbf{x}) | \mathbf{x}] \geq \mathbb{P}[y = -g(\mathbf{x}) | \mathbf{x}] \quad \text{for all } \mathbf{x} \in D_{\mathbf{x}}. \quad (16.1.1)$$

Note that we allow g to take the value 0 , which is to be understood as an additional label corresponding to nonrelevant or nonsensical input data \mathbf{x} . We will refer to $g^{-1}(0)$ as the **nonrelevant class**. The ground truth g is interpreted as how a human would classify the data, as the following example illustrates.

Example 16.1. We wish to classify whether an image shows a panda ($y = 1$) or a wombat ($y = -1$). Consider again Figure 16.1, and denote the three images by $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$. The first image \mathbf{x}_1 is a photograph of a panda. Together with a label y , it can be interpreted as a draw (\mathbf{x}_1, y) from \mathcal{D} , i.e. $\mathbf{x}_1 \in D_{\mathbf{x}}$ and $g(\mathbf{x}_1) = 1$. The second image \mathbf{x}_2 displays noise and corresponds to nonrelevant data as it shows neither a panda nor a wombat. In particular, $\mathbf{x}_2 \in D_{\mathbf{x}}^c$ and $g(\mathbf{x}_2) = 0$. The third (perturbed) image \mathbf{x}_3 also belongs to $D_{\mathbf{x}}^c$, as it is not a photograph but a noise corrupted version of \mathbf{x}_1 . Nonetheless, it is *not* nonrelevant, as a human would classify it as a panda. Thus $g(\mathbf{x}_3) = 1$.

Additional to the ground truth g , we denote by

$$h: \mathbb{R}^d \rightarrow \{-1, 1\}$$

some trained classifier.

Definition 16.2. Let $g: \mathbb{R}^d \rightarrow \{-1, 0, 1\}$ be the ground-truth classifier, let $h: \mathbb{R}^d \rightarrow \{-1, 1\}$ be a classifier, and let $\|\cdot\|_*$ be a norm on \mathbb{R}^d . For $\mathbf{x} \in \mathbb{R}^d$ and $\delta > 0$, we call $\mathbf{x}' \in \mathbb{R}^d$ an **adversarial example** to $\mathbf{x} \in \mathbb{R}^d$ with perturbation δ , if and only if

- (i) $\|\mathbf{x}' - \mathbf{x}\|_* \leq \delta$,
- (ii) $g(\mathbf{x})g(\mathbf{x}') > 0$,
- (iii) $h(\mathbf{x}) = g(\mathbf{x})$ and $h(\mathbf{x}') \neq g(\mathbf{x}')$.

¹To be more precise, the conditional distribution of $y | \mathbf{x}$ is only well-defined almost everywhere w.r.t. the marginal distribution of \mathbf{x} . Thus (16.1.1) can only be assumed to hold for *almost every* $\mathbf{x} \in D_{\mathbf{x}}$ w.r.t. to the marginal distribution of \mathbf{x} .

In words, \mathbf{x}' is an adversarial example to \mathbf{x} with perturbation δ , if (i) the distance of \mathbf{x} and \mathbf{x}' is at most δ , (ii) \mathbf{x} and \mathbf{x}' belong to the same (not nonrelevant) class according to the ground truth classifier, and (iii) the classifier h correctly classifies \mathbf{x} but misclassifies \mathbf{x}' .

Remark 16.3. We emphasize that the concept of a ground-truth classifier g differs from a minimizer of the Bayes risk (14.1.1) for two reasons. First, we allow for an additional label 0 corresponding to the nonrelevant class, which does not exist for the data generating distribution \mathcal{D} . Second, g should correctly classify points *outside of* $D_{\mathbf{x}}$; small perturbations of images as we find them in adversarial examples, are not regular images in $D_{\mathbf{x}}$. Nonetheless, a human classifier can still classify these images, and g models this property of human classification.

16.2 Bayes classifier

At first sight, an adversarial example seems to be no more than a misclassified sample. Naturally, these exist if the model does not generalize well. In this section we present a more nuanced view from [217].

To avoid edge cases, we assume in the following that for all $\mathbf{x} \in D_{\mathbf{x}}$

$$\text{either } \mathbb{P}[y = 1|\mathbf{x}] > \mathbb{P}[y = -1|\mathbf{x}] \quad \text{or} \quad \mathbb{P}[y = 1|\mathbf{x}] < \mathbb{P}[y = -1|\mathbf{x}] \quad (16.2.1)$$

so that (16.1.1) uniquely defines $g(\mathbf{x})$ for $\mathbf{x} \in D_{\mathbf{x}}$. We say that the distribution **exhausts the domain** if $D_{\mathbf{x}} \cup g^{-1}(0) = \mathbb{R}^d$. This means that every point is either in the feature support $D_{\mathbf{x}}$ or it belongs to the nonrelevant class. Moreover, we say that h is a **Bayes classifier** if

$$\mathbb{P}[h(\mathbf{x})|\mathbf{x}] \geq \mathbb{P}[-h(\mathbf{x})|\mathbf{x}] \quad \text{for all } \mathbf{x} \in D_{\mathbf{x}}.$$

By (16.1.1), the ground truth g is a Bayes classifier, and (16.2.1) ensures that h coincides with g on $D_{\mathbf{x}}$ if h is a Bayes classifier. It is easy to see that a Bayes classifier minimizes the Bayes risk.

With these two notions, we now distinguish between four cases.

- (i) *Bayes classifier/exhaustive distribution:* If h is a Bayes classifier and the data exhausts the domain, then there are *no adversarial examples*. This is because every $\mathbf{x} \in \mathbb{R}^d$ either belongs to the nonrelevant class or is classified the same by h and g .
- (ii) *Bayes classifier/non-exhaustive distribution:* If h is a Bayes classifier and the distribution does not exhaust the domain, then *adversarial examples can exist*. Even though the learned classifier h coincides with the ground truth g on the feature support, adversarial examples can be constructed for data points on the complement of $D_{\mathbf{x}} \cup g^{-1}(0)$, which is not empty.
- (iii) *Not a Bayes classifier/exhaustive distribution:* The set $D_{\mathbf{x}}$ can be covered by the four subdomains

$$\begin{aligned} C_1 &= h^{-1}(1) \cap g^{-1}(1), & F_1 &= h^{-1}(-1) \cap g^{-1}(1), \\ C_{-1} &= h^{-1}(-1) \cap g^{-1}(-1), & F_{-1} &= h^{-1}(1) \cap g^{-1}(-1). \end{aligned} \quad (16.2.2)$$

If $\text{dist}(C_1 \cap D_{\mathbf{x}}, F_1 \cap D_{\mathbf{x}})$ or $\text{dist}(C_{-1} \cap D_{\mathbf{x}}, F_{-1} \cap D_{\mathbf{x}})$ is smaller than δ , then there exist points $\mathbf{x}, \mathbf{x}' \in D_{\mathbf{x}}$ such that \mathbf{x}' is an adversarial example to \mathbf{x} with perturbation δ . Hence, *adversarial examples in the feature support can exist*. This is, however, not guaranteed to happen. For example, $D_{\mathbf{x}}$ does not need to be connected if $g^{-1}(0) \neq \emptyset$, see Exercise 16.18. Hence, even for classifiers that have incorrect predictions on the data, adversarial examples *do not need to exist*.

- (iv) *Not a Bayes classifier/non-exhaustive distribution:* In this case *everything is possible*. Data points and their associated adversarial examples can appear in the feature support of the distribution and adversarial examples to elements in the feature support of the distribution can be created by leaving the feature support of the distribution. We will see examples in the following section.

16.3 Affine classifiers

For linear classifiers, a simple argument outlined in [222] and [72] showcases that the high-dimensionality of the input, common in image classification problems, is a potential cause for the existence of adversarial examples.

A linear classifier is a map of the form

$$\mathbf{x} \mapsto \text{sign}(\mathbf{w}^\top \mathbf{x}) \quad \text{where } \mathbf{w}, \mathbf{x} \in \mathbb{R}^d.$$

Let

$$\mathbf{x}' := \mathbf{x} - 2|\mathbf{w}^\top \mathbf{x}| \frac{\text{sign}(\mathbf{w}^\top \mathbf{x}) \text{sign}(\mathbf{w})}{\|\mathbf{w}\|_1}$$

where $\text{sign}(\mathbf{w})$ is understood coordinate-wise. Then $\|\mathbf{x} - \mathbf{x}'\|_\infty \leq 2|\mathbf{w}^\top \mathbf{x}|/\|\mathbf{w}\|_1$ and it is not hard to see that $\text{sign}(\mathbf{w}^\top \mathbf{x}') \neq \text{sign}(\mathbf{w}^\top \mathbf{x})$.

For high-dimensional vectors \mathbf{w}, \mathbf{x} chosen at random but possibly dependent such that \mathbf{w} is uniformly distributed on a $d - 1$ dimensional sphere, it holds with high probability that

$$\frac{|\mathbf{w}^\top \mathbf{x}|}{\|\mathbf{w}\|_1} \leq \frac{\|\mathbf{x}\| \|\mathbf{w}\|}{\|\mathbf{w}\|_1} \ll \|\mathbf{x}\|.$$

This can be seen by noting that for every $c > 0$

$$\mu(\{\mathbf{w} \in \mathbb{R}^d \mid \|\mathbf{w}\|_1 > c, \|\mathbf{w}\| \leq 1\}) \rightarrow 1 \text{ for } d \rightarrow \infty, \quad (16.3.1)$$

where μ is the uniform probability measure on the d -dimensional Euclidean unit ball, see Exercise 16.17. Thus, if \mathbf{x} has a moderate Euclidean norm, the perturbation of \mathbf{x}' is likely small for large dimensions.

Below we give a sufficient condition for the existence of adversarial examples, in case both h and the ground truth g are linear classifiers.

Theorem 16.4. *Let $\mathbf{w}, \bar{\mathbf{w}} \in \mathbb{R}^d$ be nonzero. For $\mathbf{x} \in \mathbb{R}^d$, let $h(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x})$ be a classifier and let $g(\mathbf{x}) = \text{sign}(\bar{\mathbf{w}}^\top \mathbf{x})$ be the ground-truth classifier.*

For every $\mathbf{x} \in \mathbb{R}^d$ with $h(\mathbf{x})g(\mathbf{x}) > 0$ and all $\varepsilon \in (0, |\mathbf{w}^\top \mathbf{x}|)$ such that

$$\frac{|\bar{\mathbf{w}}^\top \mathbf{x}|}{\|\bar{\mathbf{w}}\|} > \frac{\varepsilon + |\mathbf{w}^\top \mathbf{x}|}{\|\mathbf{w}\|} \frac{|\mathbf{w}^\top \bar{\mathbf{w}}|}{\|\mathbf{w}\| \|\bar{\mathbf{w}}\|} \quad (16.3.2)$$

it holds that

$$\mathbf{x}' = \mathbf{x} - h(\mathbf{x}) \frac{\varepsilon + |\mathbf{w}^\top \mathbf{x}|}{\|\mathbf{w}\|^2} \mathbf{w} \quad (16.3.3)$$

is an adversarial example to \mathbf{x} with perturbation $\delta = (\varepsilon + |\mathbf{w}^\top \mathbf{x}|)/\|\mathbf{w}\|$.

Before we present the proof, we give some interpretation of this result. First, note that $\{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{w}^\top \mathbf{x} = 0\}$ is the decision boundary of h , meaning that points lying on opposite sides of this hyperplane, are classified differently by h . Due to $|\mathbf{w}^\top \bar{\mathbf{w}}| \leq \|\mathbf{w}\| \|\bar{\mathbf{w}}\|$, (16.3.2) implies that an adversarial example always exists whenever

$$\frac{|\bar{\mathbf{w}}^\top \mathbf{x}|}{\|\bar{\mathbf{w}}\|} > \frac{|\mathbf{w}^\top \mathbf{x}|}{\|\mathbf{w}\|}. \quad (16.3.4)$$

The left term is the decision margin of \mathbf{x} for g , i.e. the distance of \mathbf{x} to the decision boundary of g . Similarly, the term on the right is the decision margin of \mathbf{x} for h . Thus we conclude that adversarial examples exist if the decision margin of \mathbf{x} for the ground truth g is larger than that for the classifier h .

Second, the term $(\mathbf{w}^\top \bar{\mathbf{w}})/(\|\mathbf{w}\| \|\bar{\mathbf{w}}\|)$ describes the alignment of the two classifiers. If the classifiers are not aligned, i.e., \mathbf{w} and $\bar{\mathbf{w}}$ have a large angle between them, then adversarial examples exist even if the margin of the classifier is larger than that of the ground-truth classifier.

Finally, adversarial examples with small perturbation are possible if $|\mathbf{w}^\top \mathbf{x}| \ll \|\mathbf{w}\|$. The extreme case $\mathbf{w}^\top \mathbf{x} = 0$ means that \mathbf{x} lies on the decision boundary of h , and if $|\mathbf{w}^\top \mathbf{x}| \ll \|\mathbf{w}\|$ then \mathbf{x} is close to the decision boundary of h .

of *Theorem 16.4*. We verify that \mathbf{x}' in (16.3.3) satisfies the conditions of an adversarial example in Definition 16.2. In the following we will use that due to $h(\mathbf{x})g(\mathbf{x}) > 0$

$$g(\mathbf{x}) = \text{sign}(\bar{\mathbf{w}}^\top \mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x}) = h(\mathbf{x}) \neq 0. \quad (16.3.5)$$

First, it holds

$$\|\mathbf{x} - \mathbf{x}'\| = \left\| \frac{\varepsilon + |\mathbf{w}^\top \mathbf{x}|}{\|\mathbf{w}\|^2} \mathbf{w} \right\| = \frac{\varepsilon + |\mathbf{w}^\top \mathbf{x}|}{\|\mathbf{w}\|} = \delta.$$

Next we show $g(\mathbf{x})g(\mathbf{x}') > 0$, i.e. that $(\bar{\mathbf{w}}^\top \mathbf{x})(\bar{\mathbf{w}}^\top \mathbf{x}')$ is positive. Plugging in the definition of \mathbf{x}' , this term reads

$$\begin{aligned} \bar{\mathbf{w}}^\top \mathbf{x} \left(\bar{\mathbf{w}}^\top \mathbf{x} - h(\mathbf{x}) \frac{\varepsilon + |\mathbf{w}^\top \mathbf{x}|}{\|\mathbf{w}\|^2} \bar{\mathbf{w}}^\top \mathbf{w} \right) &= |\bar{\mathbf{w}}^\top \mathbf{x}|^2 - |\bar{\mathbf{w}}^\top \mathbf{x}| \frac{\varepsilon + |\mathbf{w}^\top \mathbf{x}|}{\|\mathbf{w}\|^2} \bar{\mathbf{w}}^\top \mathbf{w} \\ &\geq |\bar{\mathbf{w}}^\top \mathbf{x}|^2 - |\bar{\mathbf{w}}^\top \mathbf{x}| \frac{\varepsilon + |\mathbf{w}^\top \mathbf{x}|}{\|\mathbf{w}\|^2} |\bar{\mathbf{w}}^\top \mathbf{w}|, \end{aligned} \quad (16.3.6)$$

where the equality holds because $h(\mathbf{x}) = g(\mathbf{x}) = \text{sign}(\bar{\mathbf{w}}^\top \mathbf{x})$ by (16.3.5). Dividing the right-hand side of (16.3.6) by $|\bar{\mathbf{w}}^\top \mathbf{x}| \|\bar{\mathbf{w}}\|$, which is positive by (16.3.5), we obtain

$$\frac{|\bar{\mathbf{w}}^\top \mathbf{x}|}{\|\bar{\mathbf{w}}\|} - \frac{\varepsilon + |\mathbf{w}^\top \mathbf{x}|}{\|\mathbf{w}\|} \frac{|\bar{\mathbf{w}}^\top \mathbf{w}|}{\|\mathbf{w}\| \|\bar{\mathbf{w}}\|}. \quad (16.3.7)$$

The term (16.3.7) is positive thanks to (16.3.2).

Finally, we check that $0 \neq h(\mathbf{x}') \neq h(\mathbf{x})$, i.e. $(\mathbf{w}^\top \mathbf{x})(\mathbf{w}^\top \mathbf{x}') < 0$. We have that

$$\begin{aligned} (\mathbf{w}^\top \mathbf{x})(\mathbf{w}^\top \mathbf{x}') &= |\mathbf{w}^\top \mathbf{x}|^2 - \mathbf{w}^\top \mathbf{x} h(\mathbf{x}) \frac{\varepsilon + |\mathbf{w}^\top \mathbf{x}|}{\|\mathbf{w}\|^2} \mathbf{w}^\top \mathbf{w} \\ &= |\mathbf{w}^\top \mathbf{x}|^2 - |\mathbf{w}^\top \mathbf{x}| (\varepsilon + |\mathbf{w}^\top \mathbf{x}|) < 0, \end{aligned}$$

where we used that $h(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x})$. This completes the proof. \square

Theorem 16.4 readily implies the following proposition for *affine* classifiers.

Proposition 16.5. *Let $\mathbf{w}, \bar{\mathbf{w}} \in \mathbb{R}^d$ and $b, \bar{b} \in \mathbb{R}$. For $\mathbf{x} \in \mathbb{R}^d$ let $h(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b)$ be a classifier and let $g(\mathbf{x}) = \text{sign}(\bar{\mathbf{w}}^\top \mathbf{x} + \bar{b})$ be the ground-truth classifier.*

For every $\mathbf{x} \in \mathbb{R}^d$ with $\bar{\mathbf{w}}^\top \mathbf{x} \neq 0$, $h(\mathbf{x})g(\mathbf{x}) > 0$, and all $\varepsilon \in (0, |\mathbf{w}^\top \mathbf{x} + b|)$ such that

$$\frac{|\bar{\mathbf{w}}^\top \mathbf{x} + \bar{b}|^2}{\|\bar{\mathbf{w}}\|^2 + \bar{b}^2} > \frac{(\varepsilon + |\mathbf{w}^\top \mathbf{x} + b|)^2}{\|\mathbf{w}\|^2 + b^2} \frac{(\mathbf{w}^\top \bar{\mathbf{w}} + b\bar{b})^2}{(\|\mathbf{w}\|^2 + b^2)(\|\bar{\mathbf{w}}\|^2 + \bar{b}^2)}$$

it holds that

$$\mathbf{x}' = \mathbf{x} - h(\mathbf{x}) \frac{\varepsilon + |\mathbf{w}^\top \mathbf{x} + b|}{\|\mathbf{w}\|^2} \mathbf{w}$$

is an adversarial example with perturbation $\delta = (\varepsilon + |\mathbf{w}^\top \mathbf{x} + b|)/\|\mathbf{w}\|$ to \mathbf{x} .

The proof is left to the reader, see Exercise 16.19.

Let us now study two cases of linear classifiers, which allow for different types of adversarial examples. In the following two examples, the ground-truth classifier $g : \mathbb{R}^d \rightarrow \{-1, 1\}$ is given by $g(\mathbf{x}) = \text{sign}(\bar{\mathbf{w}}^\top \mathbf{x})$ for $\bar{\mathbf{w}} \in \mathbb{R}^d$ with $\|\bar{\mathbf{w}}\| = 1$.

For the first example, we construct a Bayes classifier h admitting adversarial examples in the complement of the feature support. This corresponds to case (ii) in Section 16.2.

Example 16.6. Let \mathcal{D} be the uniform distribution on

$$\{(\lambda \bar{\mathbf{w}}, g(\lambda \bar{\mathbf{w}})) \mid \lambda \in [-1, 1] \setminus \{0\}\} \subseteq \mathbb{R}^d \times \{-1, 1\}.$$

The feature support equals

$$D_{\mathbf{x}} = \{\lambda \bar{\mathbf{w}} \mid \lambda \in [-1, 1] \setminus \{0\}\} \subseteq \text{span}\{\bar{\mathbf{w}}\}.$$

Next fix $\alpha \in (0, 1)$ and set $\mathbf{w} := \alpha \bar{\mathbf{w}} + (1 - \alpha)\mathbf{v}$ for some $\mathbf{v} \in \bar{\mathbf{w}}^\perp$ with $\|\mathbf{v}\| = 1$, so that $\|\mathbf{w}\| = 1$. We let $h(\mathbf{x}) := \text{sign}(\mathbf{w}^\top \mathbf{x})$. We now show that every $\mathbf{x} \in D_{\mathbf{x}}$ satisfies the assumptions of Theorem 16.4, and therefore admits an adversarial example.

Note that $h(\mathbf{x}) = g(\mathbf{x})$ for every $\mathbf{x} \in D_{\mathbf{x}}$. Hence h is a Bayes classifier. Now fix $\mathbf{x} \in D_{\mathbf{x}}$. Then $|\mathbf{w}^\top \mathbf{x}| \leq \alpha |\bar{\mathbf{w}}^\top \mathbf{x}|$, so that (16.3.2) is satisfied. Furthermore, for every $\varepsilon > 0$ it holds that

$$\delta := \frac{\varepsilon + |\mathbf{w}^\top \mathbf{x}|}{\|\mathbf{w}\|} \leq \varepsilon + \alpha.$$

Hence, for $\varepsilon < |\mathbf{w}^\top \mathbf{x}|$ it holds by Theorem 16.4 that there exists an adversarial example with perturbation less than $\varepsilon + \alpha$. For small α , the situation is depicted in the upper panel of Figure 16.2.

For the second example, we construct a distribution with global feature support and a classifier which is not a Bayes classifier. This corresponds to case (iv) in Section 16.2.

Example 16.7. Let $\mathcal{D}_{\mathbf{x}}$ be a distribution on \mathbb{R}^d with positive Lebesgue density everywhere outside the decision boundary $\text{DB}_g = \{\mathbf{x} \mid \bar{\mathbf{w}}^\top \mathbf{x} = 0\}$ of g . We define \mathcal{D} to be the distribution of $(X, g(X))$ for $X \sim \mathcal{D}_{\mathbf{x}}$. In addition, let $\mathbf{w} \notin \{\pm \bar{\mathbf{w}}\}$, $\|\mathbf{w}\| = 1$ and $h(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x})$. We exclude $\mathbf{w} = -\bar{\mathbf{w}}$ because, in this case, every prediction of h is wrong. Thus no adversarial examples are possible.

By construction the feature support is given by $D_{\mathbf{x}} = \mathbb{R}^d$. Moreover, $h^{-1}(\{\pm 1\})$ and $g^{-1}(\{\pm 1\})$ are half spaces, which implies that, in the notation of (16.2.2) that

$$\text{dist}(C_{\pm 1} \cap D_{\mathbf{x}}, F_{\pm 1} \cap D_{\mathbf{x}}) = \text{dist}(C_{\pm 1}, F_{\pm 1}) = 0.$$

Hence, for every $\delta > 0$ there is a positive probability of observing \mathbf{x} to which an adversarial example with perturbation δ exists.

The situation is depicted in the lower panel of Figure 16.2.

16.4 ReLU neural networks

So far we discussed classification by affine classifiers. A binary classifier based on a ReLU neural network is a function $\mathbb{R}^d \ni \mathbf{x} \mapsto \text{sign}(\Phi(\mathbf{x}))$, where Φ is a ReLU neural network. As noted in [222], the arguments for affine classifiers, see Proposition 16.5, can be applied to the affine pieces of Φ , to show existence of adversarial examples.

Consider a ground-truth classifier $g: \mathbb{R}^d \rightarrow \{-1, 0, 1\}$. For each $\mathbf{x} \in \mathbb{R}^d$ we define the geometric margin of g at \mathbf{x} as

$$\mu_g(\mathbf{x}) := \text{dist}(\mathbf{x}, g^{-1}(\{g(\mathbf{x})\})^c),$$

i.e., as the distance of \mathbf{x} to the closest element that is classified differently from \mathbf{x} or the infimum over all distances to elements from other classes if no closest element exists. Additionally, we denote the distance of \mathbf{x} to the closest adjacent affine piece by

$$\nu_\Phi(\mathbf{x}) := \text{dist}(\mathbf{x}, A_{\Phi, \mathbf{x}}^c),$$

where $A_{\Phi, \mathbf{x}}$ is the largest connected region on which Φ is affine and which contains \mathbf{x} . We have the following theorem.

Theorem 16.8. *Let $\Phi: \mathbb{R}^d \rightarrow \mathbb{R}$ and for $\mathbf{x} \in \mathbb{R}^d$ let $h(\mathbf{x}) = \text{sign}(\Phi(\mathbf{x}))$. Denote by $g: \mathbb{R}^d \rightarrow \{-1, 0, 1\}$ the ground-truth classifier. Let $\mathbf{x} \in \mathbb{R}^d$ and $\varepsilon > 0$ be such that $\nu_\Phi(\mathbf{x}) > 0$, $g(\mathbf{x}) \neq 0$, $\nabla \Phi(\mathbf{x}) \neq 0$ and*

$$\mu_g(\mathbf{x}), \nu_\Phi(\mathbf{x}) > \frac{\varepsilon + |\Phi(\mathbf{x})|}{\|\nabla \Phi(\mathbf{x})\|}.$$

Then

$$\mathbf{x}' := \mathbf{x} - h(\mathbf{x}) \frac{\varepsilon + |\Phi(\mathbf{x})|}{\|\nabla \Phi(\mathbf{x})\|^2} \nabla \Phi(\mathbf{x})$$

is an adversarial example to \mathbf{x} with perturbation $\delta = (\varepsilon + |\Phi(\mathbf{x})|)/\|\nabla \Phi(\mathbf{x})\|$.

Proof. We show that \mathbf{x}' satisfies the properties in Definition 16.2.

By construction $\|\mathbf{x} - \mathbf{x}'\| \leq \delta$. Since $\mu_g(\mathbf{x}) > \delta$ it follows that $g(\mathbf{x}) = g(\mathbf{x}')$. Moreover, by assumption $g(\mathbf{x}) \neq 0$, and thus $g(\mathbf{x})g(\mathbf{x}') > 0$.

It only remains to show that $h(\mathbf{x}') \neq h(\mathbf{x})$. Since $\delta < \nu_\Phi(\mathbf{x})$, we have that $\Phi(\mathbf{x}) = \nabla\Phi(\mathbf{x})^\top \mathbf{x} + b$ and $\Phi(\mathbf{x}') = \nabla\Phi(\mathbf{x})^\top \mathbf{x}' + b$ for some $b \in \mathbb{R}$. Therefore,

$$\begin{aligned}\Phi(\mathbf{x}) - \Phi(\mathbf{x}') &= \nabla\Phi(\mathbf{x})^\top (\mathbf{x} - \mathbf{x}') = \nabla\Phi(\mathbf{x})^\top \left(h(\mathbf{x}) \frac{\varepsilon + |\Phi(\mathbf{x})|}{\|\nabla\Phi(\mathbf{x})\|^2} \nabla\Phi(\mathbf{x}) \right) \\ &= h(\mathbf{x})(\varepsilon + |\Phi(\mathbf{x})|).\end{aligned}$$

Since $h(\mathbf{x})|\Phi(\mathbf{x})| = \Phi(\mathbf{x})$ it follows that $\Phi(\mathbf{x}') = -h(\mathbf{x})\varepsilon$. Hence, $h(\mathbf{x}') = -h(\mathbf{x})$, which completes the proof. \square

Remark 16.9. We look at the key parameters in Theorem 16.8 to understand which factors facilitate adversarial examples.

- *The geometric margin of the ground-truth classifier $\mu_g(\mathbf{x})$:* To make the construction possible, we need to be sufficiently far away from points that belong to a different class than \mathbf{x} or to the nonrelevant class.
- *The distance to the next affine piece $\nu_\Phi(\mathbf{x})$:* Since we are looking for an adversarial example within the same affine piece as \mathbf{x} , we need this piece to be sufficiently large.
- *The perturbation δ :* The perturbation is given by $(\varepsilon + |\Phi(\mathbf{x})|)/\|\nabla\Phi(\mathbf{x})\|$, which depends on the classification margin $|\Phi(\mathbf{x})|$ of the ReLU classifier and its sensitivity to inputs $\|\nabla\Phi(\mathbf{x})\|$. For adversarial examples to be possible, we either want a small classification margin of Φ or a high sensitivity of Φ to its inputs.

16.5 Robustness

Having established that adversarial examples can arise in various ways under mild assumptions, we now turn our attention to conditions that prevent their existence.

16.5.1 Global Lipschitz regularity

We have repeatedly observed in the previous sections that a large value of $\|\mathbf{w}\|$ for linear classifiers $\text{sign}(\mathbf{w}^\top \mathbf{x})$, or $\|\nabla\Phi(\mathbf{x})\|$ for ReLU classifiers $\text{sign}(\Phi(\mathbf{x}))$, facilitates the occurrence of adversarial examples. Naturally, both these values are upper bounded by the Lipschitz constant of the classifier's inner functions $\mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x}$ and $\mathbf{x} \mapsto \Phi(\mathbf{x})$. Consequently, it was stipulated early on that bounding the Lipschitz constant of the inner functions could be an effective measure against adversarial examples [222].

We have the following result for general classifiers of the form $\mathbf{x} \mapsto \text{sign}(\Phi(\mathbf{x}))$.

Proposition 16.10. Let $\Phi: \mathbb{R}^d \rightarrow \mathbb{R}$ be C_L -Lipschitz with $C_L > 0$, and let $s > 0$. Let $h(\mathbf{x}) = \text{sign}(\Phi(\mathbf{x}))$ be a classifier, and let $g: \mathbb{R}^d \rightarrow \{-1, 0, 1\}$ be a ground-truth classifier. Moreover, let $\mathbf{x} \in \mathbb{R}^d$ be such that

$$\Phi(\mathbf{x})g(\mathbf{x}) \geq s. \quad (16.5.1)$$

Then there does not exist an adversarial example to \mathbf{x} of perturbation $\delta < s/C_L$.

Proof. Let $\mathbf{x} \in \mathbb{R}^d$ satisfy (16.5.1) and assume that $\|\mathbf{x}' - \mathbf{x}\| \leq \delta$. The Lipschitz continuity of Φ implies

$$|\Phi(\mathbf{x}') - \Phi(\mathbf{x})| < s.$$

Since $|\Phi(\mathbf{x})| \geq s$ we conclude that $\Phi(\mathbf{x}')$ has the same sign as $\Phi(\mathbf{x})$ which shows that \mathbf{x}' cannot be an adversarial example to \mathbf{x} . \square

Remark 16.11. As we have seen in Lemma 13.2, we can bound the Lipschitz constant of ReLU neural networks by restricting the magnitude and number of their weights and the number of layers.

There has been some criticism to results of this form, see, e.g., [98], since an assumption on the Lipschitz constant may potentially restrict the capabilities of the neural network too much. We next present a result that shows under which assumptions on the training set, there exists a neural network that classifies the training set correctly, but does not allow for adversarial examples within the training set.

Theorem 16.12. Let $m \in \mathbb{N}$, let $g: \mathbb{R}^d \rightarrow \{-1, 0, 1\}$ be a ground-truth classifier, and let $(\mathbf{x}_i, g(\mathbf{x}_i))_{i=1}^m \in (\mathbb{R}^d \times \{-1, 1\})^m$. Assume that

$$\sup_{i \neq j} \frac{|g(\mathbf{x}_i) - g(\mathbf{x}_j)|}{\|\mathbf{x}_i - \mathbf{x}_j\|} =: \widetilde{M} > 0.$$

Then there exists a ReLU neural network Φ with $\text{depth}(\Phi) = O(\log(m))$ and $\text{width}(\Phi) = O(dm)$ such that for all $i = 1, \dots, m$

$$\text{sign}(\Phi(\mathbf{x}_i)) = g(\mathbf{x}_i)$$

and there is no adversarial example of perturbation $\delta = 1/\widetilde{M}$ to \mathbf{x}_i .

Proof. The result follows directly from Theorem 9.6 and Proposition 16.10. The reader is invited to complete the argument in Exercise 16.20. \square

16.5.2 Local regularity

One issue with upper bounds involving global Lipschitz constants such as those in Proposition 16.10, is that these bounds may be quite large for deep neural networks. For example, the upper

bound given in Lemma 13.2 is

$$\|\Phi(\mathbf{x}) - \Phi(\mathbf{x}')\|_\infty \leq C_\sigma^L \cdot (Bd_{\max})^{L+1} \|\mathbf{x} - \mathbf{x}'\|_\infty$$

which grows exponentially with the depth of the neural network. However, in practice this bound may be pessimistic, and locally the neural network might have significantly smaller gradients than the global Lipschitz constant.

Because of this, it is reasonable to study results preventing adversarial examples under *local* Lipschitz bounds. Such a result together with an algorithm providing bounds on the local Lipschitz constant was proposed in [87]. We state the theorem adapted to our set-up.

Theorem 16.13. *Let $h: \mathbb{R}^d \rightarrow \{-1, 1\}$ be a classifier of the form $h(\mathbf{x}) = \text{sign}(\Phi(\mathbf{x}))$ and let $g: \mathbb{R}^d \rightarrow \{-1, 0, 1\}$ be the ground-truth classifier. Let $\mathbf{x} \in \mathbb{R}^d$ satisfy $g(\mathbf{x}) \neq 0$, and set*

$$\alpha := \max_{R>0} \min \left\{ \Phi(\mathbf{x})g(\mathbf{x}) / \sup_{\substack{\|\mathbf{y}-\mathbf{x}\|_\infty \leq R \\ \mathbf{y} \neq \mathbf{x}}} \frac{|\Phi(\mathbf{y}) - \Phi(\mathbf{x})|}{\|\mathbf{x} - \mathbf{y}\|_\infty}, R \right\}, \quad (16.5.2)$$

where the minimum is understood to be R in case the supremum is zero. Then there are no adversarial examples to \mathbf{x} with perturbation $\delta < \alpha$.

Proof. Let $\mathbf{x} \in \mathbb{R}^d$ be as in the statement of the theorem. Assume, towards a contradiction, that for $0 < \delta < \alpha$ satisfying (16.5.2), there exists an adversarial example \mathbf{x}' to \mathbf{x} with perturbation δ .

If the supremum in (16.5.2) is zero, then Φ is constant on a ball of radius R around \mathbf{x} . In particular for $\|\mathbf{x}' - \mathbf{x}\| \leq \delta < R$ holds $h(\mathbf{x}') = h(\mathbf{x})$ and \mathbf{x}' cannot be an adversarial example.

Now assume the supremum in (16.5.2) is not zero. It holds by (16.5.2), that

$$\delta < \Phi(\mathbf{x})g(\mathbf{x}) / \sup_{\substack{\|\mathbf{y}-\mathbf{x}\|_\infty \leq R \\ \mathbf{y} \neq \mathbf{x}}} \frac{|\Phi(\mathbf{y}) - \Phi(\mathbf{x})|}{\|\mathbf{x} - \mathbf{y}\|_\infty}. \quad (16.5.3)$$

Moreover,

$$\begin{aligned} |\Phi(\mathbf{x}') - \Phi(\mathbf{x})| &\leq \sup_{\substack{\|\mathbf{y}-\mathbf{x}\|_\infty \leq R \\ \mathbf{y} \neq \mathbf{x}}} \frac{|\Phi(\mathbf{y}) - \Phi(\mathbf{x})|}{\|\mathbf{x} - \mathbf{y}\|_\infty} \|\mathbf{x} - \mathbf{x}'\|_\infty \\ &\leq \sup_{\substack{\|\mathbf{y}-\mathbf{x}\|_\infty \leq R \\ \mathbf{y} \neq \mathbf{x}}} \frac{|\Phi(\mathbf{y}) - \Phi(\mathbf{x})|}{\|\mathbf{x} - \mathbf{y}\|_\infty} \delta < \Phi(\mathbf{x})g(\mathbf{x}), \end{aligned}$$

where we applied (16.5.3) in the last line. It follows that

$$\begin{aligned} g(\mathbf{x})\Phi(\mathbf{x}') &= g(\mathbf{x})\Phi(\mathbf{x}) + g(\mathbf{x})(\Phi(\mathbf{x}') - \Phi(\mathbf{x})) \\ &\geq g(\mathbf{x})\Phi(\mathbf{x}) - |\Phi(\mathbf{x}') - \Phi(\mathbf{x})| > 0. \end{aligned}$$

This rules out \mathbf{x}' as an adversarial example. □

The supremum in (16.5.2) is bounded by the Lipschitz constant of Φ on $B_R(\mathbf{x})$. Thus Theorem 16.13 depends only on the local Lipschitz constant of Φ . One obvious criticism of this result is that the computation of (16.5.2) is potentially prohibitive. We next show a different result, for which the assumptions can immediately be checked by applying a simple algorithm that we present subsequently.

To state the following proposition, for a continuous function $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ and $\delta > 0$ we define for $\mathbf{x} \in \mathbb{R}^d$ and $\delta > 0$

$$z^{\delta, \max} := \max\{\Phi(\mathbf{y}) \mid \|\mathbf{y} - \mathbf{x}\|_\infty \leq \delta\} \quad (16.5.4)$$

$$z^{\delta, \min} := \min\{\Phi(\mathbf{y}) \mid \|\mathbf{y} - \mathbf{x}\|_\infty \leq \delta\}. \quad (16.5.5)$$

Proposition 16.14. *Let $h: \mathbb{R}^d \rightarrow \{-1, 1\}$ be a classifier of the form $h(\mathbf{x}) = \text{sign}(\Phi(\mathbf{x}))$ and $g: \mathbb{R}^d \rightarrow \{-1, 0, 1\}$, let \mathbf{x} be such that $h(\mathbf{x}) = g(\mathbf{x})$. Then \mathbf{x} does not have an adversarial example of perturbation δ if $z^{\delta, \max} z^{\delta, \min} > 0$.*

Proof. The proof is immediate, since $z^{\delta, \max} z^{\delta, \min} > 0$ implies that all points in a δ neighborhood of \mathbf{x} are classified the same. \square

To apply (16.14), we only have to compute $z^{\delta, \max}$ and $z^{\delta, \min}$. It turns out that if Φ is a neural network, then $z^{\delta, \max}$, $z^{\delta, \min}$ can be approximated by a computation similar to a forward pass of Φ . Denote by $|\mathbf{A}|$ the matrix obtained by taking the absolute value of each entry of the matrix \mathbf{A} . Additionally, we define

$$\mathbf{A}^+ = (|\mathbf{A}| + \mathbf{A})/2 \text{ and } \mathbf{A}^- = (|\mathbf{A}| - \mathbf{A})/2.$$

The idea behind the Algorithm 2 is common in the area of neural network verification, see, e.g., [65, 60, 7, 237].

Remark 16.15. Up to constants, Algorithm 2 has the same computational complexity as a forward pass, also see Algorithm 1. In addition, in contrast to upper bounds based on estimating the global Lipschitz constant of Φ via its weights, the upper bounds found via Algorithm 2 include the effect of the activation function σ . For example, if σ is the ReLU, then we may often end up in a situation, where $\delta^{(\ell), \text{up}}$ or $\delta^{(\ell), \text{low}}$ can have many entries that are 0. If an entry of $\mathbf{W}^{(\ell)} \mathbf{x}^{(\ell)} + \mathbf{b}^{(\ell)}$ is nonpositive, then it is guaranteed that the associated entry in $\delta^{(\ell), \text{low}}$ will be zero. Similarly, if $\mathbf{W}^{(\ell)}$ has only few positive entries, then most of the entries of $\delta^{(\ell), \text{up}}$ are not propagated to $\delta^{(\ell+1), \text{up}}$.

Next, we prove that Algorithm 2 indeed produces sensible output.

Proposition 16.16. *Let Φ be a neural network with weight matrices $\mathbf{W}^{(\ell)} \in \mathbb{R}^{d_{\ell+1} \times d_\ell}$ and bias vectors $\mathbf{b}^{(\ell)} \in \mathbb{R}^{d_{\ell+1}}$ for $\ell = 0, \dots, L$, and a monotonically increasing activation function σ .*

Let $\mathbf{x} \in \mathbb{R}^d$. Then the output of Algorithm 2 satisfies

$$\mathbf{x}^{L+1} + \delta^{(L+1), \text{up}} > z^{\delta, \max} \text{ and } \mathbf{x}^{L+1} - \delta^{(L+1), \text{low}} < z^{\delta, \min}.$$

Algorithm 2 Compute $\Phi(\mathbf{x})$, $z^{\delta, \max}$ and $z^{\delta, \min}$ for a given neural network.

Input: weight matrices $\mathbf{W}^{(\ell)} \in \mathbb{R}^{d_{\ell+1} \times d_\ell}$ and bias vectors $\mathbf{b}^{(\ell)} \in \mathbb{R}^{d_{\ell+1}}$ for $\ell = 0, \dots, L$ with $d_{L+1} = 1$, monotonous activation function σ , input vector $\mathbf{x} \in \mathbb{R}^{d_0}$, neighborhood size $\delta > 0$

Output: Bounds for $z^{\delta, \max}$ and $z^{\delta, \min}$

```

 $\mathbf{x}^{(0)} = \mathbf{x}$ 
 $\delta^{(0), \text{up}} = \delta \mathbf{1} \in \mathbb{R}^{d_0}$ 
 $\delta^{(0), \text{low}} = \delta \mathbf{1} \in \mathbb{R}^{d_0}$ 
for  $\ell: 0$  to  $L - 1$  do
     $\mathbf{x}^{(\ell+1)} = \sigma(\mathbf{W}^{(\ell)} \mathbf{x}^{(\ell)} + \mathbf{b}^{(\ell)})$ 
     $\delta^{(\ell+1), \text{up}} = \sigma(\mathbf{W}^{(\ell)} \mathbf{x}^{(\ell)} + (\mathbf{W}^{(\ell)})^+ \delta^{(\ell), \text{up}} + (\mathbf{W}^{(\ell)})^- \delta^{(\ell), \text{low}} + \mathbf{b}^{(\ell)}) - \mathbf{x}^{(\ell+1)}$ 
     $\delta^{(\ell+1), \text{low}} = \mathbf{x}^{(\ell+1)} - \sigma(\mathbf{W}^{(\ell)} \mathbf{x}^{(\ell)} - (\mathbf{W}^{(\ell)})^+ \delta^{(\ell), \text{low}} - (\mathbf{W}^{(\ell)})^- \delta^{(\ell), \text{up}} + \mathbf{b}^{(\ell)})$ 
end for
 $\mathbf{x}^{(L+1)} = \mathbf{W}^{(L)} \mathbf{x}^{(L)} + \mathbf{b}^{(L)}$ 
 $\delta^{(L+1), \text{up}} = (\mathbf{W}^{(L)})^+ \delta^{(L), \text{up}} + (\mathbf{W}^{(L)})^- \delta^{(L), \text{low}}$ 
 $\delta^{(L+1), \text{low}} = (\mathbf{W}^{(L)})^+ \delta^{(L), \text{low}} + (\mathbf{W}^{(L)})^- \delta^{(L), \text{up}}$ 
return  $\mathbf{x}^{(L+1)}$ ,  $\mathbf{x}^{(L+1)} + \delta^{(L+1), \text{up}}$ ,  $\mathbf{x}^{(L+1)} - \delta^{(L+1), \text{low}}$ 

```

Proof. Fix $\mathbf{y}, \mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{y} - \mathbf{x}\|_\infty \leq \delta$ and let $\mathbf{y}^{(\ell)}, \mathbf{x}^{(\ell)}$ for $\ell = 0, \dots, L + 1$ be as in Algorithm 2 applied to \mathbf{y}, \mathbf{x} , respectively. Moreover, let $\delta^{\ell, \text{up}}, \delta^{\ell, \text{low}}$ for $\ell = 0, \dots, L + 1$ be as in Algorithm 2 applied to \mathbf{x} . We will prove by induction over $\ell = 0, \dots, L + 1$ that

$$\mathbf{y}^{(\ell)} - \mathbf{x}^{(\ell)} \leq \delta^{\ell, \text{up}} \quad \text{and} \quad \mathbf{x}^{(\ell)} - \mathbf{y}^{(\ell)} \leq \delta^{\ell, \text{low}}, \quad (16.5.6)$$

where the inequalities are understood entry-wise for vectors. Since \mathbf{y} was arbitrary this then proves the result.

The case $\ell = 0$ follows immediately from $\|\mathbf{y} - \mathbf{x}\|_\infty \leq \delta$. Assume now, that the statement was shown for $\ell < L$. We have that

$$\begin{aligned} \mathbf{y}^{(\ell+1)} - \mathbf{x}^{(\ell+1)} - \delta^{\ell+1, \text{up}} &= \sigma(\mathbf{W}^{(\ell)} \mathbf{y}^{(\ell)} + \mathbf{b}^{(\ell)}) \\ &\quad - \sigma(\mathbf{W}^{(\ell)} \mathbf{x}^{(\ell)} + (\mathbf{W}^{(\ell)})^+ \delta^{\ell, \text{up}} + (\mathbf{W}^{(\ell)})^- \delta^{\ell, \text{low}} + \mathbf{b}^{(\ell)}). \end{aligned}$$

The monotonicity of σ implies that

$$\mathbf{y}^{(\ell+1)} - \mathbf{x}^{(\ell+1)} \leq \delta^{\ell+1, \text{up}}$$

if

$$\mathbf{W}^{(\ell)} \mathbf{y}^{(\ell)} \leq \mathbf{W}^{(\ell)} \mathbf{x}^{(\ell)} + (\mathbf{W}^{(\ell)})^+ \delta^{\ell, \text{up}} + (\mathbf{W}^{(\ell)})^- \delta^{\ell, \text{low}}. \quad (16.5.7)$$

To prove (16.5.7), we observe that

$$\begin{aligned} \mathbf{W}^{(\ell)} (\mathbf{y}^{(\ell)} - \mathbf{x}^{(\ell)}) &= (\mathbf{W}^{(\ell)})^+ (\mathbf{y}^{(\ell)} - \mathbf{x}^{(\ell)}) - (\mathbf{W}^{(\ell)})^- (\mathbf{y}^{(\ell)} - \mathbf{x}^{(\ell)}) \\ &= (\mathbf{W}^{(\ell)})^+ (\mathbf{y}^{(\ell)} - \mathbf{x}^{(\ell)}) + (\mathbf{W}^{(\ell)})^- (\mathbf{x}^{(\ell)} - \mathbf{y}^{(\ell)}) \\ &\leq (\mathbf{W}^{(\ell)})^+ \delta^{\ell, \text{up}} + (\mathbf{W}^{(\ell)})^- \delta^{\ell, \text{low}}, \end{aligned}$$

where we used the induction assumption in the last line. This shows the first estimate in (16.5.6). Similarly,

$$\begin{aligned} \mathbf{x}^{(\ell+1)} - \mathbf{y}^{(\ell+1)} - \delta^{\ell+1, \text{low}} \\ = \sigma(\mathbf{W}^{(\ell)} \mathbf{x}^{(\ell)} - (\mathbf{W}^{(\ell)})^+ \delta^{(\ell), \text{low}} - (\mathbf{W}^{(\ell)})^- \delta^{(\ell), \text{up}} + \mathbf{b}^{(\ell)}) - \sigma(\mathbf{W}^{(\ell)} \mathbf{y}^{(\ell)} + \mathbf{b}^{(\ell)}). \end{aligned}$$

Hence, $\mathbf{x}^{(\ell+1)} - \mathbf{y}^{(\ell+1)} \leq \delta^{\ell+1, \text{low}}$ if

$$\mathbf{W}^{(\ell)} \mathbf{y}^{(\ell)} \geq \mathbf{W}^{(\ell)} \mathbf{x}^{(\ell)} - (\mathbf{W}^{(\ell)})^+ \delta^{(\ell), \text{low}} - (\mathbf{W}^{(\ell)})^- \delta^{(\ell), \text{up}}. \quad (16.5.8)$$

To prove (16.5.8), we observe that

$$\begin{aligned} \mathbf{W}^{(\ell)} (\mathbf{x}^{(\ell)} - \mathbf{y}^{(\ell)}) &= (\mathbf{W}^{(\ell)})^+ (\mathbf{x}^{(\ell)} - \mathbf{y}^{(\ell)}) - (\mathbf{W}^{(\ell)})^- (\mathbf{x}^{(\ell)} - \mathbf{y}^{(\ell)}) \\ &= (\mathbf{W}^{(\ell)})^+ (\mathbf{x}^{(\ell)} - \mathbf{y}^{(\ell)}) + (\mathbf{W}^{(\ell)})^- (\mathbf{y}^{(\ell)} - \mathbf{x}^{(\ell)}) \\ &\leq (\mathbf{W}^{(\ell)})^+ \delta^{(\ell), \text{low}} + (\mathbf{W}^{(\ell)})^- \delta^{(\ell), \text{up}}, \end{aligned}$$

where we used the induction assumption in the last line. This completes the proof of (16.5.6) for all $\ell \leq L$.

The case $\ell = L + 1$ follows by the same argument, but replacing σ by the identity. \square

Bibliography and further reading

This chapter begins with the foundational paper [222], but it should be remarked that adversarial examples for non-deep-learning models in machine learning were studied earlier in [97].

The results in this chapter are inspired by various results in the literature, though they may not be found in precisely the same form. The overall setup is inspired by [222]. The explanation based on the high-dimensionality of the data given in Section 16.3 was first formulated in [222] and [72]. The formalism reviewed in Section 16.2 is inspired by [217]. The results on robustness via local Lipschitz properties are due to [87]. Algorithm 2 is covered by results in the area of network verifiability [65, 60, 7, 237]. For a more comprehensive overview of modern approaches, we refer to the survey article [191].

Important directions not discussed in this chapter are the transferability of adversarial examples, defense mechanisms, and alternative adversarial operations. Transferability refers to the phenomenon that adversarial examples for one model often also fool other models, [168, 151]. Defense mechanisms, i.e., techniques for specifically training a neural network to prevent adversarial examples, include for example the Fast Gradient Sign Method of [72], and more sophisticated recent approaches such as [31]. Finally, adversarial examples can be generated not only through additive perturbations, but also through smooth transformations of images, as demonstrated in [1, 242].

Exercises

Exercise 16.17. Prove (16.3.1) by comparing the volume of the d -dimensional Euclidean unit ball with the volume of the d -dimensional 1-ball of radius c for a given $c > 0$.

Exercise 16.18. Fix $\delta > 0$. For a pair of classifiers h and g such that $C_1 \cup C_{-1} = \emptyset$ in (16.2.2), there trivially cannot exist any adversarial examples. Construct an example, of h , g , \mathcal{D} such that C_1 , $C_{-1} \neq \emptyset$, h is not a Bayes classifier, and g is such that no adversarial examples with a perturbation δ exist.

Is this also possible if $g^{-1}(0) = \emptyset$?

Exercise 16.19. Prove Proposition 16.5. Hint: Repeat the proof of Theorem 16.4. In the first part set $\mathbf{x}^{(\text{ext})} = (\mathbf{x}, 1)$, $\mathbf{w}^{(\text{ext})} = (\mathbf{w}, b)$ and $\bar{\mathbf{w}}^{(\text{ext})} = (\bar{\mathbf{w}}, \bar{b})$. Then show that $h(\mathbf{x}') \neq h(\mathbf{x})$ by plugging in the definition of \mathbf{x}' .

Exercise 16.20. Complete the proof of Theorem 16.12.

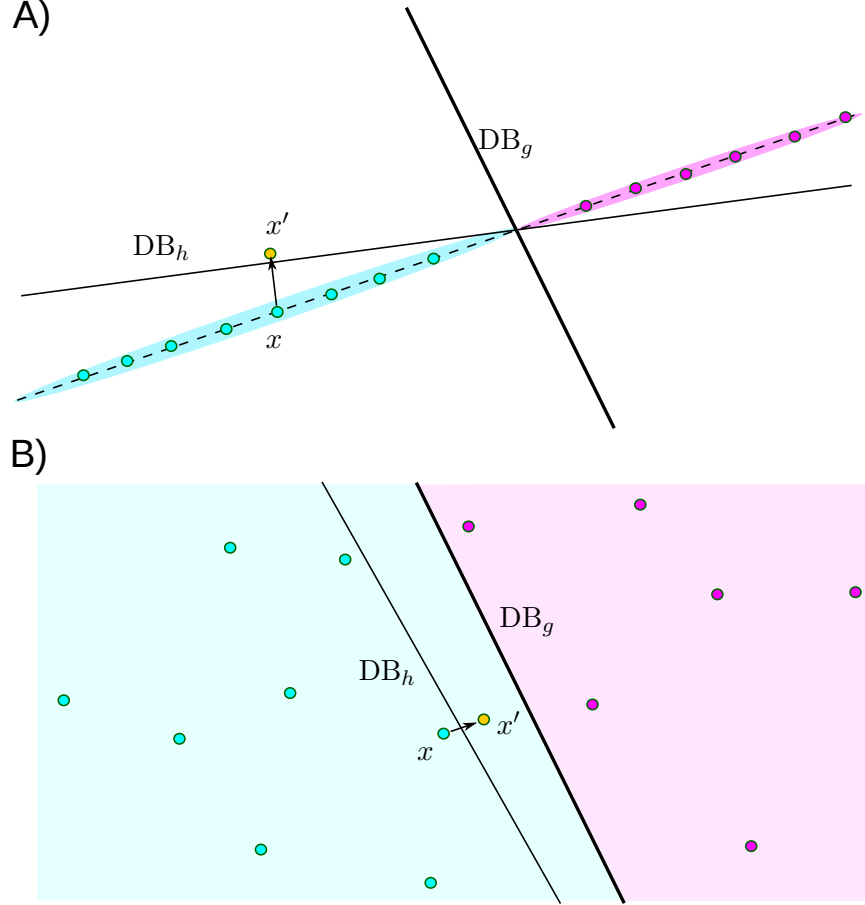


Figure 16.2: Illustration of the two types of adversarial examples in Examples 16.6 and 16.7. In panel A) the feature support $D_{\mathbf{x}}$ corresponds to the dashed line. We depict the two decision boundaries $DB_h = \{\mathbf{x} | \mathbf{w}^\top \mathbf{x} = 0\}$ of $h(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x})$ and $DB_g = \{\mathbf{x} | \bar{\mathbf{w}}^\top \mathbf{x} = 0\}$ $g(\mathbf{x}) = \text{sign}(\bar{\mathbf{w}}^\top \mathbf{x})$. Both h and g perfectly classify every data point in $D_{\mathbf{x}}$. One data point \mathbf{x} is shifted outside of the support of the distribution in a way to change its label according to h . This creates an adversarial example \mathbf{x}' . In panel B) the data distribution is globally supported. However, h and g do not coincide. Thus the decision boundaries DB_h and DB_g do not coincide. Moving data points across DB_h can create adversarial examples, as depicted by \mathbf{x} and \mathbf{x}' .