Machine Learning Course - CS-433

Adversarial ML

Nov 10, 2020

Small changes by Nicolas Flammarion, small changes by Martin Jaggi 2019;

Last updated on: November 23, 2020

Introduction

Some ML tasks are inherently difficult. E.g., consider the examples in Figure 1. Even humans might not be able to classify all examples correctly. So we would not be surprised to see NNs struggle in such cases. But typically this is where they shine, having a performance on par or perhaps even surpassing humans. But to date, NNs (or other classifiers for that matter) are not as *robust* in their decisions as humans and can be easily tricked by adversarially chosen perturbations of the input, even if the perturbations are small. This can lead to problems.

Figure 1: Dog or mop?

In the sequel it might be good to have a concrete example in mind. E.g., think of self-driving cars. And we will assume that we are using NNs. But the basic principle applies to a much wider setting.

So consider a self-driving car. Such a device hopefully will be able to recognize and correctly interpret street signs with very high probability. If the ML algorithm that performs

this task can be easily tricked by slightly manipulating the input then insurance companies might not be happy![1]

The Basic Attack using Back Propagation

So let us look at the simplest instance. Assume that we have trained a binary classifier $f : \mathcal{X} \to \{\pm 1\}$ given some training set $\mathcal{S} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$. There is an underlying data distribution $\mathcal{D}$ that is unknown to us. Assume that we have trained the network well and that it has a small true risk, i.e.,

$$\mathcal{L}(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}_{\{f(\mathbf{x})\neq y\}}] \leq \delta,$$

where $\delta > 0$ is a small number. We are happy.
But assume now that an adversary were allowed to manipulate, i.e., change, the input $\mathbf{x}$ slightly. How much worse can the risk be made? If we put no restriction on the "power" of the adversary then clearly she can increase the risk at will. So let us say that the adversary can change the given input $\mathbf{x}$ to $\tilde{\mathbf{x}}$, but that we require that $\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq \varepsilon$ for some small $\varepsilon > 0$. What norm shall we pick? This depends on the

[1] Just to clear up any confusion. There are also GANs – which confusingly stands for generative adversarial networks. Even though there is also the name "adversarial" in this topic, this has nothing to do with the current set-up. In GANs the adversarial view-point helps to train a network. E.g., assume that the task of the network is to to create realistically looking faces given "random-like" input. Here the network is used in a "generative" way. It has at it's disposal a set of human faces. In order to train this network to perform its task better we use a second network that tries to distinguish between faces generated by the network and real faces. This is the adversary. Training now proceeds in two phases that are interlaced. Given a generative network we train a powerful adversary to perform the distinction and given an adversary we train a hopeful better generative network.

application and it is part of what is called the *threat model*. We will get back to this point later. It is customary in the literature to pick either ℓ_1, ℓ_2, or ℓ_∞. We can now define the adversarial risk, call it $\mathcal{R}(f, \varepsilon)$, as

$$\mathcal{R}(f, \varepsilon) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\Big[\max_{\tilde{\mathbf{x}}:\|\mathbf{x}-\tilde{\mathbf{x}}\|\leq\varepsilon} \mathbb{1}_{\{f(\tilde{\mathbf{x}})\neq y\}}\Big].$$

In words, for every input $\mathbf{x}$ the adversary can find the worst perturbation allowed by the norm constraint.
Depending on whether you are interested in "breaking" the classifier or try to make it robust we are faced with numerous questions. Here are some in no particular order.

1. How do we find adversarial perturbations efficiently?
2. In order to find those perturbations, what access to the classifier do we need?
3. By how much worse can we make the risk?
4. Are there measures we can take to make a given classifier more robust?
5. Are there particular ways to train the classifier to make it robust?

We will explore some of these questions in the following sections.

White Box Attacks

So assume that we are given a binary classifier f. To be concrete let us assume that it is implemented by a NN. We

have complete access to the NN and are given an input $\mathbf{x}$. How do we find an adversarial perturbation $\tilde{\mathbf{x}}$ so that $\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq \varepsilon$? You will explore this in the exercise session. Here is a the basic idea.

If f does not classify $\mathbf{x}$ correctly then we are done – we can just set $\tilde{\mathbf{x}} = \mathbf{x}$. So we might as well assume that it does indeed classify $\mathbf{x}$ correctly. To be specific, assume that the data is separable in principle, i.e., that there exists a "ground truth" encoded by the function $h : \mathcal{X} \to \{\pm 1\}$. In other words, the correct label is non-probabilistic and is given by $h(\mathbf{x})$. Further assume that $f(\mathbf{x})$ has the form

$$f(\mathbf{x}) = \begin{cases} 1, & 0 \leq \frac{1}{2} \leq g(\mathbf{x}) \leq 1, \\ -1, & 0 \leq g(\mathbf{x}) < \frac{1}{2}, \end{cases} \tag{1}$$

where $g : \mathcal{X} \to [0, 1]$ represents the probability that $y = 1$ given the input. I.e., we assume that, like in logistic regression, we first compute a probability and then we quantize.

Of course $g(\mathbf{x})$ depends on all the weights and bias terms within the NN but we omit this dependence in our notation since in the present context we are interested in variations due to changes in the input rather than changes in the parameters.

Why do we assume that f has this form? We will use gradient descent in order to find adversarial perturbations. For this to work we want the objective function to be smooth. We will therefore use g rather than the original f.

Using the backpropagation algorithm, we can efficiently compute $\nabla_{\mathbf{x}} g(\mathbf{x})$. Note that this is the gradient with respect to changes in the input rather than changes in the parameters.

Hence

$$h(\mathbf{x})\nabla_{\mathbf{x}} g(\mathbf{x})$$

is a vector of length D (the dimension of the input space) which is positive in positions where an increase in that input makes the prediction more correct (increases the probability of the correct label) and negative where an increase in that dimension makes the prediction less correct.

Assume that we are allowed to move the point $\mathbf{x}$ to a new point $\tilde{\mathbf{x}}$, with the restriction that $\|\tilde{\mathbf{x}} - \mathbf{x}\|_2 \leq \varepsilon$. Note that we have assumed an ℓ_2 constraint. Our aim is to make the prediction as bad as possible. This means that we would like to decrease the probability of the correct label as much as possible.

If ε is small then it makes sense to assume that the change in the probability is given by the first order change, i.e., that it is well predicted by the gradient. Given our "budget" in terms of ℓ_2 the optimum move is then to move in the opposite direction of the gradient, i.e., to define[2]

$$\tilde{\mathbf{x}} = \mathbf{x} - \varepsilon h(\mathbf{x}) \frac{\nabla_{\mathbf{x}} g(\mathbf{x})}{\|\nabla_{\mathbf{x}} g(\mathbf{x})\|_2}.$$

If, for a particular $\mathbf{x}$, we manage to "flip" the probability using a move that is bounded by ε then we have found an adversarial example. In general, we might not want to take a single step but rather only move partially in this direction and then iterate this process.

[2]Mathematically speaking this corresponds to the following. We are given a fixed vector $\mathbf{w}$ of unit norm (in the ℓ_2 sense). Then the inner product $\mathbf{w}^\top \mathbf{v}$ conditioned on $\|\mathbf{w}\|_2 \leq \varepsilon$ is maximized by chosing $\mathbf{v} = \varepsilon \mathbf{w}$. This can be seen e.g. by the Cauchy Schwartz inequality $\mathbf{w}^\top \mathbf{v} \leq \|\mathbf{w}\|_2 \|\mathbf{v}\|_2$.

The above attack is known as a "white box" attack – we have access to the details of the algorithm.
Here is a good problem to think about. Assume that we are given as above the gradient $\nabla_{\mathbf{x}} g(\mathbf{x})$. What is the locally optimal move if our constraint is $\|\tilde{\mathbf{x}}-\mathbf{x}\|_1 \leq \varepsilon$ or $\|\tilde{\mathbf{x}}-\mathbf{x}\|_\infty \leq \varepsilon$ instead of the ℓ_2 constraint we used above?

Black Box Attacks

In the above attack we needed access to the NN that implemented the prediction in order to compute the gradients. In general it is a good idea in anything involving security to assume that the adversary has this level of access. Assume e.g., autonomous cars. Typically those are sold in large quantities and it will not be difficult for an adversary to get a hold of one of those boxes.
But even if we limit the adversary it turns out that similar attacks can still be carried out. Those are typically known as "black box" attacks. In such attacks we assume that we can observe the input out relationship but we cannot look inside the box.
Early on in the development of adversarial machine learning it was assumed that adversarial attacks could be prevented by preventing access to the algorithm or by obfuscating or masking the gradients (e.g., by adding some small amount of noise to the input before passing it through the network or by adding noise to the output). But it has been shown that all such approaches can be easily broken and that black box access is enough.
In the simplest case assume the previous setting and assume

further that we have access to the "unquantized" output $g(\mathbf{x})$. Assume that we want to compute the derivative with respect to the i-th input component. We can do this numerically by asking for the input output pairs for both $\mathbf{x}$ and $\tilde{\mathbf{x}}$ which is equal to $\mathbf{x}$, except in the i-th component, where it is slightly perturbed.
But what can we do if we only see the quantized output, i.e., the decision. In this case we cannot compute the derivative. An ingenious approach for this case is the following. Given black box access to f create a new set of samples $\mathcal{S} = \{(\mathbf{x}, f(\mathbf{x}))\}$ by computing many input-output pairs. Given $\mathcal{S}$ train a new NN. This new NN does not have to have the same structure (depth, width, activation functions etc) as the original one. Now run a white box attack on the newly-trained NN. Those adversarial direction that are found in this way often also cause trouble for the original network. If we assume that we have a sufficiently large number of examples so that we can learn f perfectly then this is perhaps not so surprising. But this approach seems to work also in cases where the number of samples we have at our disposal is quite reasonable.

Adversarial Attacks on Physical Objects

So far in our discussion we have assumed that we are given an input $\mathbf{x}$ and that we are finding an adversarial perturbation. But there exist much more interesting and potentially more dangerous variants. Let us go back to the example of a self-driving car. The car presumably has a camera (or multiple cameras). Assume that the car approaches an in-

tersection and sees a traffic sign. Perhaps this traffic sign is a stop sign. How can the adversary perturb the input. Perhaps the adversary can "hack" the car itself. But a much simpler attack is to perturb the actual physical stop sign. Is it possible to change the stop sign slightly so that a human will still recognize it as a stop sign but that the car will likely mistake it for a different sign? Note that this is quite more complicated than our original set-up. We are not given a particular input $\mathbf{x}$ but a whole family of such inputs – this family comes about since the care might approach the same stop sign from various slightly different angles, distances, and under slightly different ambient lighting conditions. And we would like the same attack (the same physical manipulation of this stop sign) to work under a broad set of those conditions. Even in this cases it has been shown that adversarial changes can be found! This is quite non-trivial! Figure 2 shows what a perturbed stop sign might look like. Note that for a human the change is visible but it is unlikely that it will cause confusion.

Why Some ML Algorithms Might Not Be Robust – Non-Robust Features

The following example illustrates one reason why a ML algorithm might learn a classification rule that has low standard risk but high adversarial risk.
This is a toy example, but the basic idea is sound the ML algorithm might rely on a large set of "non-robust" features that can be easily tricked by perturbations.

Figure 2: To stop or not to stop.

Consider a binary classification task. We have $y \in \{\pm 1\}$. Let $\hat{\mathbf{x}}$ be the input vector. Assume that after applying a suitable transform we get the new input vector $\mathbf{x}$ that has the following simple structure.

We have $\mathbf{x} = (x_1, \cdots, x_D)$, where $x_i = a_i y + Z_i$, $i = 1, \cdots, D$, where Z_i is Gaussian zero-mean and unit-variance noise which is independent for each component. Further, $a_1 = 1$ and for $i = 2, \cdots, D$, we have $a_i = \sqrt{\frac{\log(D)}{D-1}}$. The exact values for a_i are not so important and are chosen simply for convenience. What is important is that the first component contains a strong signal component, whereas the other features have a very weak such component. Strong versus weak is with respect to the strength of the noise that is added (zero-mean Gaussian of unit variance). We will say that the first feature is *robust* whereas the other features are not robust.

To summarize, each of the D components is a scaled and

noisy version of the label and all components represent conditionally independent observations. Further, the first component contains a strong signal component. The remaining $D-1$ features contain extremely weak signal components but there are many of them (we assume that D is large). Finally, assume that we are given the prior on the label $p(y)$ and that it is uniform.

Assume at first that we are interested in the best classifier without adversarial perturbations, i.e., the classifier with the smallest possible risk (error probability). This is the Bayes classifier, i.e., we should compute the posterior probability and then choose that label that maximizes the posterior:

$$\begin{aligned}\text{argmax}_{\hat{y}\in\{\pm 1\}} p(\hat{y} \mid \mathbf{x}) &= \text{argmax}_{\hat{y}\in\{\pm 1\}} \frac{p(\mathbf{x} \mid \hat{y})p(\hat{y})}{p(\mathbf{x})} \\ &= \text{argmax}_{\hat{y}\in\{\pm 1\}} \prod_{i=1}^{d} p(\mathbf{x}_i \mid \hat{y}).\end{aligned}$$

In the last step we have used the fact that under our model the observations are conditionally independent so that we get a product of the probabilities and that we have a uniform

prior. This can further be simplified to

$$\begin{aligned}
\text{argmax}_{\hat{y}\in\{\pm 1\}} p(\hat{y} \mid \mathbf{x}) &= \text{argmax}_{\hat{y}\in\{\pm 1\}} \prod_{i=1}^{d} p(\mathbf{x}_i \mid \hat{y}) \\
&= \text{argmax}_{\hat{y}\in\{\pm 1\}} \log \prod_{i=1}^{d} p(\mathbf{x}_i \mid \hat{y}) \\
&= \text{argmax}_{\hat{y}\in\{\pm 1\}} \sum_{i=1}^{d} \log p(\mathbf{x}_i \mid \hat{y}) \\
&= \text{argmax}_{\hat{y}\in\{\pm 1\}} \sum_{i=1}^{d} \log \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\mathbf{x}_i - \hat{y}a_i)^2} \\
&= \text{argmin}_{\hat{y}\in\{\pm 1\}} \sum_{i=1}^{d} (\mathbf{x}_i - \hat{y}a_i)^2 \\
&= \text{argmin}_{\hat{y}\in\{\pm 1\}} \sum_{i=1}^{d} (\mathbf{x}_i^2 - 2\mathbf{x}_i\hat{y}a_i + \hat{y}^2 a_i^2) \\
&= \text{argmax}_{\hat{y}\in\{\pm 1\}} \hat{y} \sum_{i=1}^{d} \mathbf{x}_i a_i.
\end{aligned}$$

Recall that we observe $\mathbf{x}$ which is the vector $y\big(a_1 = 1, a_2 = \sqrt{\frac{\log(D)}{D-1}}, \cdots, a_D = \sqrt{\frac{\log(D)}{D-1}}\big)$ under Gaussian noise with iid zero-mean components and unit variance in each dimension. Therefore the expression that we maximize over $\hat{y}$ above is equal to

$$\hat{y}y(\sum_{i=1}^{D} a_i^2) + \hat{y}\sum_{i=1}^{D} a_i Z_i = \hat{y}y(1 + \log(D)) + \hat{y}Z,$$

where Z is a Gaussian noise with variance $(\sum_{i=1}^{D} a_i^2) = 1 + \log(D)$. Scaling everything by $1/(1 + \log(D))$, we see that this is equivalent to observing the signal $y = \pm 1$ under a zero-mean Gaussian noise with variance $1/(1 + \log(D))$. Hence as D grows the variance tends to zero and our error probability will go to zero as well. In other words, if we train our ML algorithm well then we can hope to get close to zero standard risk when the dimension D grows.

But assume now that we allow the adversary to move the point $\mathbf{x}$ into $\tilde{\mathbf{x}}$ and that our norm is ℓ_∞ with $\varepsilon = 2\sqrt{\frac{\log D}{D-1}}$. In this case the adversary can do the following. She can first make an optimal decision based on the observation. As we have seen, with high probability she will know the correct label. She can then can move each of the non-robust features $i = 2, \cdots, D$ into the wrong direction by an amount $2\sqrt{\frac{\log D}{D-1}}$. This means that she can in effect flip the non-robust features. She can also move the first component somewhat but this has almost no effect. If we ignore the effect on the first component we see that with this change the classifier will misclassify every single sample with high probability!

In summary, we have seen an example where the optimal standard risk is essentially zero but the adversarial risk of the same classifier is almost 1. This is as bad as it gets. And we have seen that this is due to the fact that the classifier relied heavily on many very weak features that were easy to perturb.

Could we have constructed a more robust classifier? Yes, certainly, but at a price. Assume that we build a classifier based on the first feature only. The best classifier in this case

is again a Bayes classifier. A little bit of thought shows that it is given by simply taking the sign of $\mathbf{x}_1$. What is the risk of this classifier? We have a label that is either $+1$ or -1. We add a zero-mean unit-variance Gaussian random variable to it and ask what is the probability that this noise changes the sign. A little bit of thought shows that the incurred error probability in this case is equal to

$$\frac{1}{\sqrt{2\pi}} \int_1^\infty e^{-\frac{1}{2}x^2} dx \sim 0.16.$$

In words, this classifier incurs a standard risk of about sixteen percent. This is much higher than the standard risk of essentially zero we saw above. But in this case the risk almost does not change if we consider the adversarial setting since in our threat model we allow the adversary to move each component by only a very small amount. So we see in this model a trade-off between a small standard risk and a small adversarial risk. We cannot get both.

The above example might look a little construed and also it might not be clear how much of this is due to the fact that we allow the adversary to change the components in an ℓ_∞ sense. Indeed, it is more challenging to find simple examples if the threat model consist of changes in ℓ_2. But similar ideas still apply.

The idea for the above model and further details come from the paper "Robustness May Be at Odds with Accuracy," by Tsipras, Dimitris, Santurkar, Shibani, Engstrom, Logan, Turner, Alexander, and Madry, Aleksander.

Why Some ML Algorithms Might Not Be Robust – The Curse of Dimensionality – Again!

Let us talk about a simple setting to see why in high dimensions adversarial examples cannot be completely avoided.
We consider a binary classification example. Let us assume that $\mathcal{X} = \mathbb{R}^{500}$, i.e., $D = 500$. Our data is perfectly separable. For $y = -1$ the data is distributed uniformly on the surface of a sphere of radius 1 and for $y = 1$ the data is uniformly distributed on the surface of a sphere of radius 1.3. Both classes have equal size.
In the paper "The Relationship Between High-Dimensional Geometry and Adversarial Examples," by Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S. Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow the authors performed the following experiment.
They took a 2-hidden layer NN with ReLU activation functions and 500 nodes per hidden layer. They then trained with 50 million samples using SGD. The network trained very well, giving no error on 20 million test samples. Despite this they were easily able to find adversarial examples by moving test samples by a small amount (on the order of $1/\sqrt{D}$)!
Let us now discuss how this is an essentially unavoidable consequence of working in very high dimension.
Consider a classifier f that has a small but non-zero error probability. Consider lets say only those errors where $(\mathbf{x}, y)$ is such that $\|\mathbf{x}\|_2 = 1$ and $y = -1$ but $f(\mathbf{x}) = 1$. I.e., these are points lying on the inner sphere but that are classified

as belonging to the outer sphere. Let E be the set of such points that are misclassified.

For a reason that will become clear hopefully soon, assume that this set forms a spherical cap in the e_1 direction, i.e., these are the points $\mathbf{x}$ of the form $\|\mathbf{x}\|_2 = 1$ and $\mathbf{x}_1 \geq \alpha$ for some appropriate constant α. Without loss of generality we can assume that the error probability, call it p, fulfills $0 \leq p \leq \frac{1}{2}$ so that $\alpha \geq 0$.

Given p what is the value of α that gives us p (recall that we assume a uniform distribution of the data on the sphere)? This amounts to computing the surface area of the spherical cap, dividing it by the surface area of the whole sphere and equating this ratio to p. Recall that the sphere has radius 1. If we are using our low-D intuition we might think that as p varies from 0 to $\frac{1}{2}$, α will vary from 1 to 0. But in fact, we will see now that α is of order $1/\sqrt{D}$ and that p only modulates the constant in front of this expression! This means that as D becomes large, the spherical cap extends almost completely down to the equator. In other words, almost all the mass sits at the equator. As a consequence, adversarial examples are essentially unavoidable – if we randomly pick a point it is likely very close to the equator. And such a point is hence very likely $1/\sqrt{D}$-close to a point that f misclassifies. In other words, it is easy to find small adversarial moves that will cause the point to be misclassified!

It remains to clarify two things. First, let us check for the example above that indeed the spherical cap almost extends all the way down to the equator when D is large and the error probability is non-zero. Second, why did we assume

that the error set forms a spherical cap?
To answer the first question we will take a convenient shortcut. In principle the area of a spherical cap is explicitly known. But the expression is somewhat unwieldy. But note the following. Picking a point uniformly at random on a sphere of radius 1 in D-dimensional real space is almost the same as picking a Gaussian vector with iid zero-mean components of variance $\sigma^2 = 1/D$. Such a vector will be spherically symmetric and it will have a norm very close to 1.
But for the latter model it is very easy to assess the probability that this vector extends beyond a positive number β in the first component. This probability is given by

$$\int_\beta^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x_1^2}{2\sigma^2}} dx_1 = p.$$

One way to express this is to define

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy.$$

E.g., $Q(1) = 0.16$, i.e., the probability that a zero-mean unit-variance Gaussian has a value larger than 1 is about 16 percent. With this function we get the expression

$$Q(\beta/\sigma) = p.$$

Since $\sigma = \frac{1}{\sqrt{D}}$, we see from this expression that β must be of the form $\alpha/\sqrt{D}$, as claimed. More precisely,

$$\beta = Q^{-1}(p)/\sqrt{D}.$$

To answer the second question consider the following. You likely know that, in any dimension and for a fixed volume, the body that has the smallest surface area is the sphere. This is a famous so-called isoperimetric inequalities. Many such isoperimetric inequalities exist. The version we need is the following. Consider a sphere in D dimension and and a subset E on the surface of this sphere of a fixed size. Now add to this subset all points on the sphere that have distance no more than ϵ away from this set E. What is the smallest this set can be? It turns out that the answer is that the resulting set is the smallest if we start with a spherical cap! Applied to our problem this means that by assuming that the misclassified set was a spherical cap we in fact computed a lower bound on the adversarial misclassification rate!

Constructing a Robust Classifier from Scratch – Adversarial Training

So far we have assumed that we are given a classifier and we discussed what might happen if we allow adversarial changes to the input. But we can take a more pro-active approach. First, we can include the aim for adversarial robustness in the training phase. This is what we discuss now. Second, given a classifier we can ask if we can modify it to make it more robust. We will briefly discuss this second approach in the next section. In practice we might want to apply both techniques. We will be *very* brief.

The approach taken in the paper "Towards Deep Learning Models Resistant to Adversarial Attacks" by Madry, Alek-

sander, Makelov, Aleksandar, Schmidt, Ludwig, Tsipras, Dimitris, and Vladu is the following perhaps most natural idea. Rather than minimizing the usual cost function

$$\min_{\Theta} \ \big[\mathcal{L}(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}_{\{f_\Theta(\mathbf{x})\neq y\}}]\big]$$

over all parameters Θ of the model, why not directly minimize what matters, namely

$$\min_{\Theta} \ \big[\mathcal{R}(f,\varepsilon) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\max_{\tilde{\mathbf{x}}:\|\mathbf{x}-\tilde{\mathbf{x}}\|\leq\varepsilon} \mathbb{1}_{\{f(\tilde{\mathbf{x}})\neq y\}}]\big].$$

This leads to a min-max formulation. As written, this function is not easy to optimize for several reasons. The first is that it is not smooth (due to the indicator function). Therefore let us instead look at the problem

$$\min_{\Theta} \ \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\max_{\tilde{\mathbf{x}}:\|\mathbf{x}-\tilde{\mathbf{x}}\|\leq\varepsilon} (\tfrac{1}{2} - y(g(\tilde{\mathbf{x}}) - \tfrac{1}{2}))\big].$$

Here we assumed that we deal with a binary classification problem, i.e., $y \in \{\pm 1\}$ and that the classifier $f(\mathbf{x})$ has the form (1) so that $(\frac{1}{2} - y(g(\tilde{\mathbf{x}}) - \frac{1}{2}))$ is the predicted probability of the incorrect label. Now we are dealing with a smooth function.
Next, we do not have access to the distribution but instead we have a sample $\mathcal{S}$. Therefore the equivalent problem is

$$\min_{\Theta} \frac{1}{N}\sum_{n=1}^{N} \max_{\tilde{\mathbf{x}_n}:\|\mathbf{x}_n-\tilde{\mathbf{x}}_n\|\leq\epsilon} (\frac{1}{2} - y_n(g(\tilde{\mathbf{x}}_n) - \tfrac{1}{2})).$$

It is still not completely obvious how to minimize this. It turns out that the following is correct. Compute for each $\mathbf{x}_n$

the worst perturbation $\tilde{\mathbf{x}}_n$. Then take the gradient of this expression and move towards the negative gradient direction. Of course, this is computationally intensive since for each sample and each iteration we need to find the worst-case perturbation.

Let us apply this algorithm to our example with one robust and many non-robust features. To keep things simple let us assume that we even know that the optimum classifier is of the form as discussed, i.e., we first form the sum $\sum_{i=1}^{D} \mathbf{x}_i a_i$ and then take the sign, but we do not know what the best constants a_i are that we should use. Hence, we are led to the optimization task

$$\max_a \frac{1}{N} \sum_{n=1}^{N} \min_{\tilde{\mathbf{x}_n}: \|\mathbf{x}_n - \tilde{\mathbf{x}}_n\|_\infty \leq \epsilon} y_n \sum_{i=1}^{D} (\tilde{\mathbf{x}}_n)_i a_i + \lambda(\sum_{i=1}^{D} a_i^2 - 1 - \log(D)),$$

where $\epsilon = 2\sqrt{\frac{\log(D)}{D-1}}$ and where we added the term $\lambda(\sum_{i=1}^{D} a_i^2 - 1 - \log(D))$ in order to limit the scale of the coefficients.

Assume at first that we started with the optimum choice for the non-adversarial setting, i.e., $a_1 = 1$ and $a_i = \alpha = \sqrt{\frac{\log(D)}{D-1}}$. And now let us do one gradient step. The adversary will move all features in the "incorrect direction" by an amount 2α. This will leave the robust feature essentially unchanged but will "flip" all non-robust features. The gradient will have the form

$$\begin{aligned} (1 - 2\alpha + 2\lambda) + \mathcal{N}(0, \sigma^2 = \frac{1}{N}), \quad i = 1, \\ -\alpha(1 - 2\lambda) + \mathcal{N}(0, \sigma^2 = \frac{1}{N}), \quad i = 2, \cdots, D, \end{aligned}$$

and we will walk a little bit into the direction of this gradient (in the current setting we want to maximize and not to minimize the given expression). If N is large then the additional noise is negligible and the direction will become deterministic. Further, think of λ as small. We see that the non-robust features will move towards zero.
Indeed, if we consider as a second example the case where we set $a_i = 0$, $i = 2, \cdots, D$, and pick a_1 to be positive, we see that this is a fixed point of the adversarial training algorithm. This training algorithm is called *adversarial training*.

Making an Existing Classifier Robust – Randomized Smoothing

In some instances we might be handed a classifier that has small standard risk but might be prone to adversarial errors and are asked to make it more robust rather than learning a new classifier from scratch. To date the perhaps most promising idea of accomplishing this is *randomized smoothing*.
Let f be the given classifier. Assume that it maps $\mathbb{R}^D$ to $\mathcal{Y}$, the set of labels. From this we derive the *smoothed* classifier, call it g. This new smoothed classifier g maps an input $\mathbf{x}$ to that class c that is most likely returned by f if presented the input $\mathbf{x} + \mathbf{z}$, where $\mathbf{z}$ is a vector of iid zero-mean random Gaussian variables with a variance of σ^2 per component. This idea was introduced by Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D.,and Jana, S. in the paper entitled "Certified robustness to adversarial examples with

differential privacy". It was shown to work well for various test data sets like ImageNet. There are several variants of this. E.g., rather than using this probabilistic definition using Gaussians we could average over a ball of radius $\sqrt{D}\sigma$ (with a uniform distribution). The effect would be more or less the same. This resulting training algorithm variants are called *robust training*.

The basic idea why this adds robustness is the following. Consider e.g. the binary case. Take a point $\mathbf{x}$ and consider the original classifier f. Lets say that the label $y = 1$ is much more likely to be returned by f if we give it the point $\mathbf{x} + \mathbf{z}$ than the label $y = -1$. Now consider what happens if we give f the label $\tilde{\mathbf{x}} + \mathbf{z}$ instead, where $\|\tilde{\mathbf{x}} - \mathbf{x}\|_2 \leq \epsilon$ for some not too large ϵ. Since the resulting points in the two cases are picked with a probability that is not too different it is intuitive that in average we are still more likely to return $y = 1$ rather than $y = -1$. I.e., we will return the same label for points not too far away. Let us make this precise.

Consider first the case of $D = 1$, i.e., we are operating on a line. On the line each point x has a label $f(x) \in \{\pm 1\}$ attached to it. For a point x on this line define $p = \mathbb{E}[\mathbb{1}_{\{f(x+z)=1\}}]$. Let us assume that $\frac{1}{2} < p \leq 1$. This means that "more" of the points in the neighborhood of x are given the label $y = 1$ than the label $y = -1$. Consider now $\tilde{p} = \mathbb{E}[\mathbb{1}_{\{f(\tilde{x}+z)=1\}}]$ where $\|\tilde{\mathbf{x}} - \mathbf{x}\|_2 \leq \epsilon$. Assume e.g. that the new point $\tilde{x}$ is moved ϵ to the right from x. How small can $\tilde{p}$ be. A little bit of thought shows that the worst case is if all the points that were originally labeled $y = -1$ were on the right of the point x in the tail of the Gaussian.

More precisely, the worst case happens if all the point to the left of $x+\sigma Q^{-1}(1-p)$ are labeled $y=1$ and all points to the right of this point are labeled $y=-1$. How far can we then move $\tilde{x}$ to the right away from x so that still the majority of the points is labeled $y=1$. We see that we can move it at most by $\sigma Q^{-1}(1-p)$. This is pleasing. The larger p, i.e., the more biased the original average was, the more we are adversarially robust.

This was in 1-D. But exactly the same happens in any dimension since the noise is Gaussian with iid components. Hence, no matter what direction we are considering, in this direction we are dealing with a Gaussian with zero mean and variance σ^2.

So why not choose a very large σ so that we get a very large robustness radius? The averaging (smoothing) will in general increase the standard risk and it can do so considerably. I.e., there might be a considerable price to pay for the added robustness.