In this lecture, we’ll try to address the following ill-formed question:

Why do adversarial examples for samples from natural distributions exist?

In this class, we’ll see a number of explanations of why adversarial examples exist, of varying plausibility. Underlying all of these settings are different toy generative models which exhibit different behaviors under adversarial perturbations. Roughly speaking, these toy models live in one of four worlds:

1. **Adversarial examples are statistically inevitable.** This first world is the most pessimistic: it says that we should never exist that we can ever get adversarially robust estimators.

2. **Robust classifiers require more data.** This second world suggests another, but less fatal, statistical bottleneck: that we can get adversarially robust estimators, but we are currently limited by the amount of data we have.

3. **Robust classification is computationally intractable.** The third possibility suggests a different problem: that there exist classifiers that are robust, but that it may be hard to find them efficiently.

4. **Efficient robust classifiers exist.** The final world suggests that we are just dumb, and there exist robust classifiers, but we just haven’t found the right algorithm.

In the rest of this lecture, we’ll go into depth into these settings. Since there’s not much to say about the fourth world, we’ll omit it, but in the rest of the lecture we’ll see examples for the first three. All of these settings will be a huge oversimplification of the actual generative model we care about (i.e. the distribution over natural images). The question is whether they still shed any light on the real-world phenomena.

**Notation** Let us first recall some notation from last lecture. Recall that given a distribution \( \mathcal{D} \) over labeled examples \((X, y)\), a set of defined perturbations \( \mathcal{P}(x) \) for each \( x \in \mathbb{R}^d \), a classifier \( f \), and a loss function \( \ell \), we defined

\[
R(f) = R(f, \mathcal{D}, \mathcal{P}, \ell) = \mathbb{E}_{(X, y) \sim \mathcal{D}} [\ell(f(X), y)]
\]

\[
R_{\text{rob}}(f, \mathcal{P}) = R_{\text{rob}}(f, \mathcal{D}, \mathcal{P}, \ell) = \mathbb{E}_{(X, y) \sim \mathcal{D}} \left[ \sup_{X' \in \mathcal{P}(X)} \ell(f(X'), y) \right].
\]

A special case of this which we will consider extensively throughout this class is the case where \( \ell = \ell_{0/1} \) is the zero-one loss. We will also focus in this class on the case where the perturbation sets \( \mathcal{P} \) are bounded \( \ell_p \) balls. Recall that for every \( x \in \mathbb{R}^d \), we let \( \mathcal{P}_{p, \varepsilon}(x) \) to be the \( \ell_p \) ball of radius \( \varepsilon \) around \( x \).
1 Adversarial examples are statistically inevitable

Here we’ll give some examples where it is impossible to avoid adversarial examples. The simplest setting is one we’ll revisit, so we’ll give it a name:

**Definition 1.1** (The \((\theta^*, \sigma)\)-Gaussian model). Let \(\theta \in \mathbb{R}^d\), and let \(\sigma \in (0, \infty]\). We say a sample \((X, y) \in \mathbb{R}^d \times \{-1, 1\}\) is drawn from the \((\theta, \sigma)\)-Gaussian model, denoted \(\mathcal{F}(\theta, \sigma)\), if it is produced as follows: first, draw \(y \in \{-1, 1\}\) uniformly at random, then let \(X \sim \mathcal{N}(y\theta, \sigma^2I)\).

Then, we have:

**Theorem 1.1.** Let \(\theta \in \mathbb{R}^d\) be the vector whose coordinates are \(\theta_i = \frac{\sqrt{\log d}}{\sqrt{d}}\) for all \(i \in [d]\). Let \(S = \{(X_1, y_1), \ldots, (X_n, y_n)\}\) be a set of \(n\) i.i.d. samples from \(\mathcal{F}(\theta, 1)\). Then:

(a) if \(n \gtrsim d\), there exists an estimator \(f_S(x)\) so that \(R(f_S, \mathcal{F}, \ell_0/1) \leq 1/poly(d)\).

(b) for any estimator \(f\), and any \(\delta > 0\) we have

\[
R_{rob}(f, \mathcal{F}, \mathcal{P}_{\infty, 2, \frac{\sigma}{\sqrt{d}}, \ell_0/1}) \geq \frac{1}{2}.
\]

Roughly speaking, this theorem states that in this setting, given enough data, there is an estimator that succeeds at the non-robust version of the classification task with extremely high probability. However, there is no classifier that can succeed at the robust version of the classification task, even given a vanishing amount of coordinate-wise noise. Note that the trivial estimator (i.e., guess \(\pm 1\) deterministically) achieves risk \(1/2\), so the lower bound says beating the trivial bound is impossible.

**Proof of Theorem 1.1.** We will first prove (a). The estimator is simple: given \(S\), let \(\hat{\theta} = \frac{1}{n} \sum y_i X_i\), and let \(f_S(x) = \text{sgn} \left( \langle \hat{\theta}, x \rangle \right)\). That is, simply learn the direction of \(\theta\), project onto this direction, and check if the projection is positive or negative. We now prove that this satisfies the desired properties. Note that \(\hat{\theta} \sim \mathcal{N}(\theta, \frac{1}{n} I)\). Hence, by standard Gaussian concentration, we have that

\[
\text{Pr} \left[ \|\hat{\theta} - \theta\|_2^2 \geq t \cdot \frac{\sigma^2}{n} \right] \lesssim \exp(-\Omega(t))
\]

for all \(t \geq 1\). By our choice of \(n\), and since \(\|\theta\|_2^2 = \log d\), we have that

\[
\text{Pr} \left[ \|\hat{\theta} - \theta\|_2^2 \geq \frac{1}{10} \cdot \|\theta\|_2^2 \right] \leq \frac{1}{\text{poly}(d)}.
\]

Condition on the complement of the event in (1) holding for the rest of the proof. Denote this good event \(E\). Now, given a fresh sample \((X, y) \sim \mathcal{F}(\theta, 1)\), we wish to demonstrate that conditioned on this event, \(f_S\) classifies \((X, y)\) correctly with high probability. Let’s consider the case where \(y = 1\) (the case where \(y = -1\) is symmetric). Then we can assume that \(X = \theta + Z\), where \(Z \sim \mathcal{N}(0, I)\). We have

\[
\left\langle \hat{\theta}, X \right\rangle = \left\langle \hat{\theta}, \theta \right\rangle + \left\langle \hat{\theta}, Z \right\rangle \\
\geq \frac{9}{10} \cdot \|\theta\|_2^2 + \left\langle \hat{\theta}, Z \right\rangle = \frac{9}{10} \log d + \left\langle \hat{\theta}, Z \right\rangle.
\]

Further notice that \(\left\langle \hat{\theta}, Z \right\rangle \sim \mathcal{N}(0, \|\theta\|_2^2)\). Conditioned on \(E\), this random variable is a Gaussian with variance at most \(2 \log d\). Hence, by standard Gaussian concentration, the probability that this exceeds \(9/10 \log d\) is at most \(1/\text{poly}(d)\). Hence this random variable is positive with probability \(1 - 1/\text{poly}(d)\), conditioned on \(E\). Combining these two bounds yields the desired result.

We now prove (b). It suffices to show the following: no algorithm can distinguish between the following two scenarios: (1) \(X\) is an i.i.d. draw from \(\mathcal{N}(\theta, I)\), or (2) \(X \in \mathcal{P}_{\infty, d-1/2+\epsilon}(Z)\) for \(Z \sim \mathcal{N}(-\theta, I)\). To do so, it suffices to construct a coupling \((X, Z)\) between \(\mathcal{N}(\theta, I)\) and \(\mathcal{N}(-\theta, I)\) so that \(\|X - Z\|_\infty < 2 \frac{\sqrt{\log d}}{\sqrt{d}}\) almost surely (why?). But this is straightforward: the coupling is simply \((X, X - 2\theta)\) for \(X \sim \mathcal{N}(\theta, I)\). \(\square\)
This is of course a very simple example for a relatively contrived problem. One can actually prove relatively strong bounds of this form, at least under some constraints on the smoothness of the data distribution. Specifically, assume that the data distribution over $X \mathbb{R}^d$ is given by $g(Z)$, where $Z \sim \mathcal{N}(0, I)$ is an $d'$-dimensional Gaussian. That is, we assume that the data is some transformation of a Gaussian, which is a typical assumption for generative models. Additionally, let $\omega : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be monotone and invertible, and suppose that the PDF of the marginal distribution over $X$, denoted $p(x)$, satisfies
\begin{equation}
\|g(z) - g(z')\| \leq \omega (\|z - z'\|_2),
\end{equation}
where the norm of the LHS is arbitrary. Then, we have the following theorem:

**Theorem 1.2** (Special case of Theorem 1 in [1]). Let $\mathcal{D}$ be a distribution over inputs satisfying (2) with respect to a norm $\|\cdot\|$. Let $f : \mathbb{R}^d \to [K]$ be an arbitrary classification function, and suppose that $\Pr[f(X) = i] \leq 1/2$ for all $i \in [K]$. Then, for all $\eta > 0$, we have that
\begin{equation}
R_{\text{rob}}(f, \mathcal{D}, \mathcal{P}, \ell_{0/1}) \leq \sqrt{\frac{\pi}{2}} \exp \left( -\omega^{-1}(\eta)^2 / 2 \right),
\end{equation}
where $\mathcal{P}(z) = \{z' \in \mathbb{R}^d : \|z' - z\| < \eta\}$.

In other words, if the data comes from a relatively smooth transformation of a Gaussian, and the classification is relatively balanced, then there are fundamental limits on the robust accuracy of any classifier for the data.

The main counter-argument against this setting is that we tend to believe there do exist robust classifiers for the settings we care about. Specifically, we tend to believe that humans are the robust classifier we seek, almost by definition.

# 2 Robust classifiers require more data

The second possibility is that robust classification is possible, but our datasets are still too small for us to achieve it: that getting accurate and robust classifiers will require more data than getting accurate non-robust classifiers. Here is one setting in which this holds:

**Theorem 2.1** (Theorems 5 and 6 in [2]). Let $(X_1, y_1), \ldots, (X_n, y_n) \sim \mathcal{F}(\theta, \sigma)$ where $\theta \sim \mathcal{N}(0, I)$, and $\sigma = c d^{1/4}$. Then, for an appropriate choice of $c$, we have:

(a) If $n \geq 1$, there exists an estimator $f_n$ that achieves $R(f) \leq 0.01$.

(b) for any $\epsilon > 0$, and any estimator $f_n$, we have that $R_{\text{rob}}(f_n, \mathcal{F}, \mathcal{P}_{\infty, \epsilon}, \ell_{0/1}) \geq \frac{1}{2}(1 - 1/d)$, so long as
\begin{equation}
n \lesssim \frac{\epsilon^2 \sqrt{d}}{\log d}.
\end{equation}

In particular, if $\epsilon = \Theta(1)$, then this theorem says that non-robust classification with high probability can be achieved with even a single sample, and robust classification requires polynomially many samples.

**Proof.** The proof of (a) is straightforward and left to the reader (and possibly homework). In this lecture we’ll focus on the proof of (b). The way to do this will be to play some tricks with order of integration. Let $f_n$ be any classifier that depends on $(X_1, y_1), \ldots, (X_n, y_n)$. Observe that $X_i = y_i Z_i$ where $y_i \sim \{\pm 1\}$ uniformly at random at $Z_n \sim \mathcal{N}(\theta, I)$. Then we can write the robust loss of $f_n$ as:
\begin{equation}
R_{\text{rob}}(f) = \mathop{\mathbb{E}}_{\theta \sim \mathcal{N}(0, I)} \mathop{\mathbb{E}}_{y_1, \ldots, y_n \sim \{\pm 1\}} \mathop{\mathbb{E}}_{Z_1, \ldots, Z_n \sim \mathcal{N}(\theta, \sigma^2 I)} \mathop{\Pr}_{z_1, \ldots, z_n \sim \mathcal{N}(\theta, \sigma^2 I)} \left[ \exists x' \in \mathcal{P}(x) : f_n(x') \neq f(x) \right],
\end{equation}
where $f_n$ depends on $Z_1, \ldots, Z_n$ and $y_1, \ldots, y_n$, but nothing else. We now wish to pull the expectation over $\theta$ inside. Note that $\theta$ is clearly independent of the $y_i$, so that is not hard. The main difficulty is switching the order between the $Z_i$ and the $\theta$. To do so, we will use the fact that posterior distributions of Gaussians are Gaussians. Formally, a straightforward calculation shows:
Fact 2.2. An equivalent way to generate $\theta, Z_1, \ldots, Z_n$ is to draw some $\theta_2 \sim \mathcal{N}(0,1)$, let $Z_1, \ldots, Z_n \sim \mathcal{N}(\theta_2, \sigma^2 I)$, and the define $\theta$ given $Z_1, \ldots, Z_n$ to be a Gaussian with mean and covariance given by

$$
\mu' = \frac{n}{\sigma^2 + n} Z, \quad \Sigma' = \frac{\sigma^2}{\sigma^2 + n} I,
$$

where $Z = \sum_{i=1}^n Z_i$.

Thus, for some distribution $D'$, we have that the above is equal to

$$
\mathbb{E}_{y_1, \ldots, y_n \sim \{\pm 1\}} \mathbb{E}_{Z_1, \ldots, Z_n \sim D'} \left[ \mathbb{E}_{\theta \sim \mathcal{N}(\mu', \Sigma')} \mathbb{E}_{y \sim \{\pm 1\}} \mathbb{E}_{x \sim \mathcal{N}(y\theta, \sigma^2 I)} \left[ \exists x' \in \mathcal{P}(x) : f_n(x') \neq y \right] \right].
$$

We can further simplify the expression inside the brackets to

$$
\mathbb{E}_{y \sim \{\pm 1\}} \mathbb{E}_{x \sim \mathcal{N}(y\mu', \sigma^2 I + \Sigma')} \left[ \exists x' \in \mathcal{P}(x) : f_n(x') \neq y \right].
$$

The point of this is now that the expectations inside of the brackets are independent of $f_n$, so that while bounding this quantity, we can treat it as a fixed function. In particular, suppose that $y = 1$, and let $A^- = \{x \in \mathbb{R}^d : f_n(x) = -1\}$ be the set of points where $f_n$ decides $-1$, and let $A^+ = (A^-)^c$ be the set where $f_n$ decides $1$. Here is the key point: observe that if $y = 1$, then there exists an adversarial perturbation of $x$ if and only if $x$ has $\ell_\infty$ distance at most $\varepsilon$ to $A^-$. Now if $||\mu'||_\infty \leq \varepsilon$, then $\mu'$ itself is a valid $\ell_\infty$ perturbation, and hence

$$
\mathbb{P}_{x \sim \mathcal{N}(\mu', \sigma^2 I + \Sigma')} \left[ \exists x' \in \mathcal{P}(x) : f_n(x') \neq y \right] \geq \mathbb{I} ||\mu'||_\infty \leq \varepsilon \cdot \mathbb{P}_{x \sim \mathcal{N}(0, \sigma^2 I + \Sigma')} \left[ x \in A^- \right].
$$

Similarly,

$$
\mathbb{P}_{x \sim \mathcal{N}(-\mu', \sigma^2 I + \Sigma')} \left[ \exists x' \in \mathcal{P}(x) : f_n(x') \neq y \right] \geq \mathbb{I} ||\mu'||_\infty \leq \varepsilon \cdot \mathbb{P}_{x \sim \mathcal{N}(0, \sigma^2 I + \Sigma')} \left[ x \in A^+ \right].
$$

Combining these two, and using that $A^- \cup A^+ = \mathbb{R}^d$, we obtain that

$$
\mathbb{E}_{y \sim \{\pm 1\}} \mathbb{P}_{x \sim \mathcal{N}(y\mu', \sigma^2 I + \Sigma')} \left[ \exists x' \in \mathcal{P}(x) : f_n(x') \neq y \right] \geq \frac{1}{2} \cdot \mathbb{I} \|\mu'||_\infty \leq \varepsilon \cdot \mathbb{P}_{x \sim \mathcal{N}(0, \sigma^2 I + \Sigma')} \left[ \frac{n}{\sigma^2 + n} \|Z\|_\infty \leq \varepsilon \right].
$$

Thus all we need to do is understand the distribution of $Z$. However, Fact 2.2 immediately implies that $Z$ is distributed as $\mathcal{N}((1 + \sigma^2/n)I)$. Thus, simplifying we get that

$$
R_{\text{rob}}(f) \geq \frac{1}{2} \mathbb{P}_{\theta_2 \sim \mathcal{N}(0, I)} \left[ \sqrt{n} \frac{\|\theta_2\|_\infty}{\sigma^2 + n} \leq \varepsilon \right].
$$

Combining this with a standard tail bound on Gaussians yields the desired theorem.
3 Robust classification is computationally intractable

In this section, we’ll give an example of a learning problem which is easy to do non-robustly, easy to do statistically robustly, but likely impossible to do efficiently, robustly. In fact, all we really need to produce the most basic example is the following assumption:

**Assumption 3.1.** There exist two classes of distributions $D_0, D_1$ over distributions in $\mathbb{R}^d$ so that if $D_0 \in D_0$ and $D_1 \in D_1$ so that the following is true: given $n$ samples $X_1, \ldots, X_n$ from some $D \in D_0$ for $\alpha \in \{0, 1\}$ there exists an estimator $f_n$ so that $f_n(X_1, \ldots, X_n) = \alpha$ with probability $99/100$ so long as $n = \text{poly}(d)$, however, no efficiently computable $f_n$ can achieve $\Pr[f_n(X_1, \ldots, X_n) = \alpha] \geq 1/2 + o(1)$ unless $n = d^{o(1)}$.

Moreover, for any $D_0 \in D_0$ and $D_1 \in D_1$, there exist sets $S_0, S_1$ so that $\Pr_{D_1}[X \in S_1] \geq 1 - d^{-\omega(1)}$ for $i \in \{0, 1\}$, and $\|x_0 - x_1\|_\infty = \omega(1)$ for all $x_0 \in S_0$ and $x_1 \in S_1$.

There are a couple of ways one can go about giving evidence to this (quite reasonable) assumption: [3], building on work of [4], shows a learning problem which is hard for any SQ algorithm, which is a restricted class of efficient algorithms. However, we remark that the resulting classes of distributions are always quite pathological in nature, and it is not clear how much they reflect real-world issues.

Given this assumption, we have:

**Theorem 3.1.** Grant Assumption 3.1. Then there exists a distribution $F$ so that there exist classifiers $f_n, f_n'$ so that $R(f_n) \leq 0.01$, and for which $R_{\text{rob}}(f_n', \mathcal{P}_{\infty, \epsilon}) \leq 0.01$ for some $\epsilon = \omega(1)$, so long as $n = \text{poly}(d)$, but for which $R_{\text{rob}}(g_n, \mathcal{P}_{\infty, \epsilon}) \geq 1 - 2 - o(1)$ for any efficiently computable $g_n$.

**Proof.** The reduction is straightforward: we take an instance of the distinguishing problem in the Assumption, so let $D_0 \in D_0$ and $D_1 \in D_1$. To form our classification problem, we let $\beta_1, \ldots, \beta_n \sim \{\pm 1\}$, and our samples are given as $(X_i, y_i)$, where $X_i \in \mathbb{R}^{d+1}$, so that the first $d$ coordinates of $X_i$ are drawn from $D_{\beta_i}$, and the last coordinate of $X_i$ is $\beta_i\epsilon$. Similarly let $y_i = \beta_i$.

We now verify that this has the desired properties. It is clear that there is a good non-robust classifier: simply ignore the first $d$ coordinates of $X_i$, and use the last coordinate as a perfect classifier. However, in the robust setting, the adversary can always set the last coordinate of $X_i$ to 0. The information-theoretic estimator does not care: it can simply use the information from the first $d-1$ coordinates. It can use this to learn $D_0$ and $D_1$, and form the sets $S_0$ and $S_1$. Given a potentially corrupted test sample $X$, it simply takes the first $d$ coordinates of $X$, and finds which set they are closer to, and outputs this solution. It is clear that this estimator succeeds with high probability. However, it is similarly clear that no efficient algorithm can succeed at this problem, by Assumption.

Later on, [3, 5] also construct another type of classification task with similar (but quantitatively stronger) bounds, from standard cryptographic primitives.

It is hard to argue that this scenario is impossible. One could argue that human brains have had significantly more time (and in a different model of computation) to get good at image classification than say a neural network. However, at the same time, the relative artificial-ness of the examples is a bit worrying. Moreover, it is essentially impossible to check this hypothesis, so it is not clear how to validate this hypothesis.

4 Efficient robust classifiers exist

To be determined...?

References


