6.883 Science of Deep Learning – Spring 2018 March 21, 2018

Lecture 13: Adversarially Robust Generalization

Lecturer: Aleksander Mądry

Scribe: Sayeri Lala, Yueqi Sheng, Tristan Bepler
(Revised by Andrew Ilyas and Dimitris Tsipras)

1 Introduction

Recall the min-max problem corresponding to adversarial training (from Lecture 11):

$$\min_{\theta} \sum_{i=1}^{N} \max_{\delta \in \Delta} loss(\theta, x_i + \delta, y_i).$$

In previous lectures, we noted that despite the non-convexity/non-concavity of the minimization and the maximization subproblems respectively, this problem is solveable in practice for standard datasets. That is, one can train a classifier that correctly classifies each training example (i.e. any $x \in \{x_i\}$), as well as all the points in its Δ neighborhood. However, when it comes robustness on the test set ($x \notin \{x_i\}$) a measure of robust *generalization*), the results can vary wildly. For instance, adversarial training has good test-set performance on the MNIST dataset, but rather poor performance on the CIFAR10 dataset (ref. Figure 1).

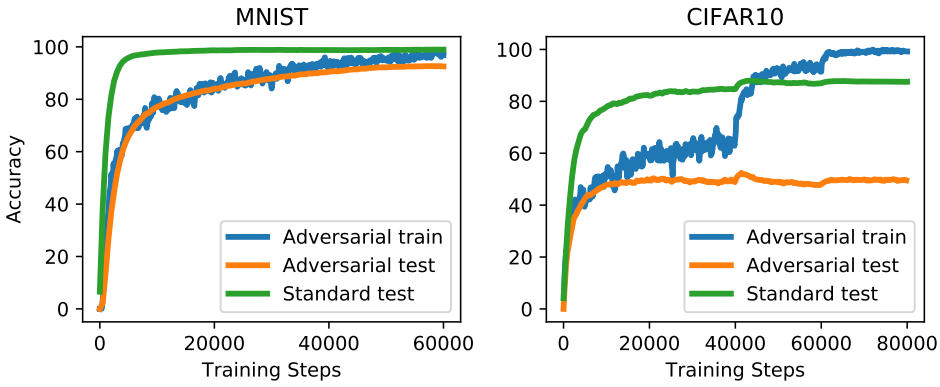

Figure 1: Adversarially robust generalization on the MNIST and CIFAR10 datasets. While the model generalizes well on MNIST, there is a large gap in the adversarial accuracy of the CIFAR10 training and test set [1].

What could explain the performance difference between the two datasets? To gain more insight, we need to first refine our notion of generalization.

Standard Generalization Recall (Lecture 3) that, when solving a classification problem, our goal is to minimize the *distributional loss*, given by

$$\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim D} \left[loss(\theta, x, y) \right],$$

where D is the true data distribution, (x, y) is a sampled input-label pair, and θ is the vector of classifier parameters.

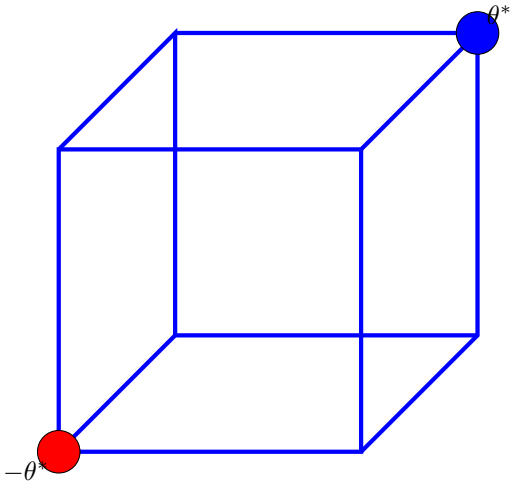

Figure 2: Data generated according to Bernoulli distribution. For this distribution, linear classifiers generalize in the standard and robust sense given a single example training point.

Since we don't have access to the true distribution, we compute the empirical mean of the loss based on the samples obtained from the sampling distribution $\hat{D}$. We thus focus on minimizing the *empirical loss*,

$$\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim\widehat{D}} \Big[loss(\theta, x, y) \Big].$$

The gap between the distributional and the empirical losses is called *generalization gap*.

Adversarially Robust Generalization Our goal when training robust classifiers to solve the problem

$$\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim D} \Big[\max_{\delta\in\Delta} loss(\theta, x+\delta, y) \Big].$$

Again, we don't have access to D, so we compute the empirical mean of the loss based on the samples obtained from the sampling distribution $\hat{D}$. So the expression becomes:

$$\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim\widehat{D}} \Big[\max_{\delta\in\Delta} loss(\theta, x+\delta, y) \Big].$$

The difference between these two quantities is defined as the *adversarially robust generalization gap*.

2 Adversarially Robust Generalization (Lower) Bounds

In Lecture 3, we established upper bounds on the expected error term in the standard setting. In order to understand how classifiers generalize differently in the standard and adversarial sense, we will study the two examples presented in [2], where our goal is to learn a linear classifier between two copies of a distribution. In both cases, we will establish that for linear classifier, *a single* sample suffices to generalize in the standard sense— it turns out, however, that at least $\sqrt{d}$ samples are needed to generalize robustly. In one case (somewhat akin to MNIST), this can be resolved by learning a non-linear classifier; in the other, we show an information-theoretic barrier to learning a robust classifier with any less than $\Omega(\sqrt{d})$ samples.

2.1 "Bernoulli" case

First, we consider the case where the data comes from one of two Bernoulli-like distributions; our goal is to learn a linear classifier which distinguishes between the two distributions. The sampling procedure is as follows:

1. We sample a random point θ u.a.r. from the hypercube, i.e. $\theta \in \{+1, -1\}^d$.

2. Choose y according to the following distribution:

$$y = \begin{cases} +1, & \text{with probability} \frac{1}{2}, \\ -1, & \text{with probability} \frac{1}{2}. \end{cases}$$

3. Generate each coordinate of x according to the following distribution:

$$x_i = \begin{cases} y\theta_i, & \text{with probability } \frac{1}{2} + \tau, \\ -y\theta_i, & \text{with probability } \frac{1}{2} - \tau. \end{cases}$$

We say these data samples (x, y) come from a Bernoulli distribution $D_B(\tau)$ using the above scheme. This distribution is visualized in Figure 2. We will fix

$$\tau = \frac{C}{d^{\frac{1}{4}}},$$

where d is the number of dimensions and C some constant. Under this setup, we give and prove the following claims:

Claim 1 *Given a single sample (x_1, y_1) from $D_B(\tau)$, the linear classifier $f(x) = sign((y_1 x_1)^\top x)$ generalizes in the standard sense.*

Proof For $f(x)$ to generalize, we first require $\mathbb{E}[y_2 f(x_2)] > 0$ where $\mathbb{E}$ is the expectation over draws of θ, (x_1, y_1) and (x_2, y_2) from $D_B(\tau)$. Using the definition of $f(x)$ above, it suffices to ensure that the following quantity is > 0

$$\begin{aligned} \mathbb{E}[y_2 \left(y_1 x_1\right)^\top x_2] &= \sum_{i=1}^{d} \mathbb{E}\left[y_1 y_2 x_1^{(i)} x_2^{(i)}\right] \\ &= \sum_{i=1}^{d} \left(\frac{1}{2} + \tau\right)\left(\frac{1}{2} + \tau\right) - 2\left(\frac{1}{2} + \tau\right)\left(\frac{1}{2} - \tau\right) + \left(\frac{1}{2} - \tau\right)\left(\frac{1}{2} - \tau\right) \quad (1) \\ &= 4d\tau^2 \quad (2) \\ &= 4C^2\sqrt{d} \quad (3) \end{aligned}$$

where (1) is simply writing out the expectation and (3) uses the definition of τ above. Therefore, $\mathbb{E}[y_2 f(x_2)] > 0$. It remains to show that $y_2 f(x_2)$ is close to this expectation with high-probability. We can show that since $\sum_i y_1 y_2 x_1^{(i)} x_2^{(i)}$ is sub-Gaussian,

$$(y_1 x_1)^T (y_2 x_2) \approx 4C^2\sqrt{d} \pm O\left(\sqrt{d} \log \frac{1}{2}\right).$$

■

Robust linear classifiers need more samples. We have shown that a single sample suffices for standard generalization. We now prove the previously claimed statement, that we need $\sqrt{d}$ samples to learn a robust linear classifier.

Claim 2 *Given $\{(x_1, y_1), \cdots, (x_m, y_m)\} \sim_{iid} D_B(\tau)$. For any **linear** classifier $f_w(x) = \langle w, x \rangle$ with $w \in \mathbb{R}^d$, f_w doesn't generalize robustly to adversarial examples unless $m = \Omega(\sqrt{d})$. Here the set of possible δ is $\Delta = \{\delta : \|\delta\|_\infty \leq \epsilon\}$ and $\epsilon = \frac{C_2}{d^{\frac{1}{4}}}$.*

Proof (sketch) First, some intuition: Suppose $\langle w, x\rangle \geq 0$. Then, the worst-case $\delta \in \Delta$ (which again, is the ℓ_∞ ball) for the current classifier is given by

$$\min_{\delta\in\Delta} \langle w, x+\delta\rangle \equiv \langle w, x\rangle + \min_{\delta\in\Delta} \langle w, \delta\rangle = \langle w, x\rangle - \epsilon \left\|\mathrm{w}\right\|_1 .$$

where the last step follows from the properties of the ℓ_∞ norm. Thus intuitively, we need the effect of $f(x)$ to outweight the adversarial additive $\epsilon \left\|\mathrm{w}\right\|_1$ in order to get a good error rate.

Suppose f_w is the classifier learned given $\{(x_i, y_i)\}$. From Claim 1, for f_w to be robust we need $\langle w, y(x+\delta)\rangle \geq 0$ with high probability. We begin with a few observations about our problem:

- By definition of $D_B(\tau)$, given y, x_i has the same distribution as

  $$x_i = F_i y \theta_i,$$

  where F_i has distribution $Ber(\frac{1}{2}+\tau)$.

- We can write w_i as follows:

  $$w_i = sign(w_i)|w_i| = S_i\theta_i|w_i|$$

  where $|w_i|, sign(w_i)$ is a function of $\{(x_i, y_i)\}$ and $S_i = \theta_i sign(w_i)$ represents whether w_i has the same sign as θ_i.

Now, we start by rewriting $\langle w, y(x+\delta)\rangle$ as follows:

$$\langle w, y(x+\delta)\rangle = \sum_i w_i y(x_i+\delta_i) = \sum_i |w_i| S_i\theta_i y x_i - \epsilon|w_i| = \sum_i |w_i|(S_iF_i - \epsilon) \tag{4}$$

where the first inequality follows from $\left\|\delta\right\|_\infty \leq \epsilon$, and the last line follows from

$$x_i y\theta_i = (F_i y\theta_i)y\theta_i = F_i.$$

We claim (without full proof, see [2]) that the only way f_w is a robust classifier is when it learns θ. This is captured by the following Lemma:

Lemma 3 *For all i, we have that $\theta_i \cdot sign(w_i) = 1$ with high probability.*

Proof (sketch) If this is not the case, i.e. if the algorithm does not learn the sign of θ_i, then we show $\langle w, yx\rangle$ may not have enough magnitude to cancel $\epsilon \left\|\mathrm{w}\right\|_1$.

$$\mathbb{E}[\langle w, y(x+\delta)\rangle] = \sum_i \mathbb{E}[|w_i|S_iF_i] - \epsilon \left\|\mathrm{w}\right\|_1$$

$\mathbb{E}[|w_i|S_iF_i] = 2\tau|w_i|\mathbb{E}[S_i]$. It can be shown (as in [2]) that $|E[S_i]| \leq C'\tau$ for some constant C'. This implies

$$\mathbb{E}[\langle w, y(x+\delta)\rangle] \leq (C'\tau^2 - \epsilon)\left\|\mathrm{w}\right\|_1$$

Since $\epsilon = \frac{C_2}{d^{\frac{1}{4}}}$, $\tau = \frac{C_1}{d^{\frac{1}{4}}}$, we must have that $\mathbb{E}[\langle w, y(x+\delta)\rangle] < 0$. ■

To get a lower bound on number of samples needed to get such w, first observe that given Lemma 3, f_w can distinguish a τ-biased coin from a fair coin. This is because $\langle w, y(x+\delta)\rangle = \sum_i |w_i|(F_i - \epsilon) \geq 0$ but if $\sum_i |w_i|(P_i - \epsilon) \leq 0$ w.h.p. for $\epsilon = \frac{C_2}{d^{\frac{1}{4}}}$.

As we have seen in previous lectures, however, to distinguish τ biased coin we need $\Omega(\frac{1}{\tau^2})$ samples. Pluging in $\tau = \frac{C_1}{d^{\frac{1}{4}}}$ reveals that the number of sample needed is $\Omega(\sqrt{d})$. ■

A robust non-linear classifier. If we were to do some pre-processing on $(x+\delta)$, then using the fact that $\epsilon = O(\tau)$, one could recover (x, y). Observe that $x \in \{1, -1\}^d$, and for all i we have

$$|(x+\delta)_i - x_i| = |\delta_i| \leq \epsilon.$$

Thus, one can threshold/quantize $x+\delta$ to go back to standard case (where one sample suffices). Therefore, there exist non-linear classifiers for this setting that generalize robustly with a single sample. This will not be the case for the next setting.

2.2 Gaussian distribution

We present a Gaussian distribution that exhibits an *information-theoretic* (and thus classifier-independent) gap between the number of samples needed in the standard case and in the adversarial case.

Our Gaussian distribution $D_G(\sigma)$ is defined as follows:

- Sample $\theta \sim \mathcal{N}(0, I)$ where $\theta \in \mathbb{R}^d$
- Sample (x, y) from the distribution:,

$$y \sim Unif(\{1, -1\})$$

$$x|\theta \sim \mathcal{N}(y\theta, \sigma^2 I),$$

  where $\sigma = c_1 d^{\frac{1}{4}}$.

Under this setup, we can once again show that a single sample suffices for standard generalization:

Claim 4 *Given a single sample (x_1, y_1) from $D_G(\sigma)$, the linear classifier $f(x) = sgn(\langle y_1 \frac{x_1}{\|x_1\|_2}, x\rangle)$ generalizes in the standard sense. In particular, with high probability over a single first sample (x_1, y_1), we get a classifier that generalizes almost perfectly, such that that*

$$\mathbb{P}_{(x,y)\sim D_G(\sigma)}[yf(x) > 0] \approx 1$$

Proof Given (x_1, y_1), we define $w = y_1 \frac{x_1}{\|x_1\|_2}$. Since we have that $yf(x) = y \cdot sgn(\langle y_1 \frac{x_1}{\|x_1\|_2}, x\rangle) = sgn(\langle w, yx\rangle)$, we can rewrite the above probability as

$$\mathbb{P}_{(x,y)\sim D_G(\sigma)}[yf(x) > 0] = \mathbb{P}_{z\sim\mathcal{N}(\theta,\sigma^2 I)}[\langle w, z\rangle > 0].$$

Thus, to prove the claim it suffices to upper bound $\mathbb{P}_{z\sim\mathcal{N}(\theta,\sigma^2 I)}[\langle w, z\rangle \leq 0]$.

Now, let $z' \sim \mathcal{N}(0, \sigma^2 I)$. Observe that $\langle w, z'\rangle$ has distribution $\mathcal{N}(0, \sigma^2)$, and z has the same distribution as $\theta + z'$.

$$\mathbb{P}[\langle w, z\rangle \leq 0] = \mathbb{P}[\langle w, \theta\rangle + \langle w, z'\rangle \leq 0] = \mathbb{P}[\langle w, z'\rangle \leq -\langle w, \theta\rangle] \tag{5}$$

We wish to upper bound the above probability. Observe that

- For LHS: $\langle w, z'\rangle$ has the same distribution as $\mathcal{N}(0, \sigma^2)$ where we use the fact that $\langle w, w\rangle = 1$.
- For RHS:

$$\langle w, \theta\rangle = \frac{1}{\|x_1\|_2} \|x_1\|_2 \langle w, \theta\rangle .$$

  Recall that $\|x_1\|_2 \langle w, \theta\rangle = \langle y_1 x_1, \theta\rangle$, which has the same distribution as $\langle \theta, \theta\rangle + \langle z', \theta\rangle$.

- For $\langle x_1, x_1\rangle$: x_1 has the same distribution as $\theta + z'$. $\|x_1\|_2 \sim \|\theta + z'\|_2$

Now, recall the following standard tail bound for Gaussian:

Theorem 5 (Hoeffding's Inequality) *For $X \sim \mathcal{N}(0, \sigma^2 I)$ and $w \in \mathbb{R}^n$ s.t. $\|w\|_2 = 1$, for any $t \geq 0$,*

$$\mathbb{P}[|\langle w, X\rangle| \geq t] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Using this tail bound, we can rewrite the probability given in 5 as:

$$\mathbb{P}_{z'}[\langle w, z'\rangle \leq -\langle w, \theta\rangle] \leq 2\exp\left(-\frac{\langle w, \theta\rangle^2}{2\sigma^2}\right). \tag{6}$$

What remains is just to bound the range of $\langle w, \theta\rangle$. We take advantage of the following concentration bound for chi-squared distributions:

Theorem 6 (Concentration of χ^2 random variables) *For $X \in \mathcal{N}(0, I)$, $\delta \in (0, 1)$*

$$\mathbb{P}[\langle X, X\rangle \geq (1 \pm \delta)d] \leq \exp\left(-\frac{d\delta^2}{8}\right)$$

We will first bound the value of $\langle \theta, \theta\rangle$. This follows directly from the concentration given by Theorem 6: $\langle \theta, \theta\rangle = (1 \pm \frac{1}{\sqrt{d}})d$ w.p. at least $(1 - 2exp(-\frac{\sqrt{d}}{8}))$.

Assume $\langle \theta, \theta\rangle \in (1 \pm \frac{1}{\sqrt{d}})d$ from now on. Based on observation above (that $\langle x_1, x_1\rangle \sim \langle \theta + z', \theta + z'\rangle$), we can bound the value of $\|\mathrm{x}_1\|_2$:

$$\mathbb{P}\left[\|\mathrm{x}_1\|_2 \geq (1 + 2\sigma)\sqrt{d}\right] \leq \mathbb{P}\left[\|\theta\|_2 + \|\mathrm{z'}\|_2 \geq (1 + 2\sigma)\sqrt{d}\right] = \mathbb{P}\left[\langle z', z'\rangle \geq 4\sigma^2 d\right] \leq \exp\left(-\frac{d}{2}\right),$$

where the last inequality follows from Theorem 6 with $\delta = \frac{1}{\sqrt{d}}$. We now turn to show the concentration of $\langle y_1 x_1, \theta\rangle$:

$$\mathbb{P}[|\langle y_1 x_1, \theta\rangle| \geq (1 \pm 2\sigma)d] \leq \mathbb{P}[|\langle \theta, z'\rangle| \geq \sigma d] \leq 2\exp\left(-\frac{d^2}{2\langle \theta, \theta\rangle}\right)$$

where the last inequality follows from Theorem 5.

Combining the above bounds gives $\langle w, \theta\rangle \approx \sqrt{(1 \pm 2\sigma)d}$. Plugging this into equation 6 with $t = \sqrt{(1 - 2\sigma)d}$ completes the proof. ■

One can extend the above proof to the case where linear classifier is generated using n samples $\{(x_i, y_i)\} \sim D_G(\sigma)$ instead of one. In this case, we let $w' = \frac{1}{n}\sum_i y_i x_i$ and $f(x) = sgn(\langle \frac{1}{\|\mathrm{w'}\|_2} w', x\rangle)$. Observe that for a fixed θ, $w' \sim \mathcal{N}(\theta, \frac{\sigma^2}{n} I)$, and also that $\|\mathrm{w'}\|_2 \sim \|\theta + \frac{1}{\sqrt{\mathrm{n}}}\mathrm{z'}\|_2$. We can obtain then concentration results in a similar fashion.

Claim 7 *For any binary classifier f_n, the expected l_∞^ϵ-robust classification error of f_n is at least $(1 - \frac{1}{d})\frac{1}{2}$ if $n \leq c_2 \frac{\epsilon^2\sqrt{d}}{\log d}$, where $\epsilon \geq 0$.*

Proof To prove this claim, we first show that for any algorithm g_n that learns a binary classifier f_n, the expected l_∞^ϵ-robust classification error of f_n is at least

$$\frac{1}{2}\mathop{\mathbb{P}}_{v \sim \mathcal{N}(0,I)}\left[\sqrt{\frac{n}{\sigma^2 + n}}\|v\|_\infty \leq \epsilon\right]$$

.

where $\sigma > 0, \epsilon \geq 0$, n samples are drawn according to the Gaussian model.

We first state the expected l_∞^ϵ-robust classification error of f_n:

$$\mathop{\mathbb{E}}_{\theta \sim \mathcal{N}(0,I)}\left[\mathop{\mathbb{E}}_{y_1,\ldots,y_n \sim \mathcal{R}}\left[\mathop{\mathbb{E}}_{x1,\ldots,x_n \sim \mathcal{N}(y_i\theta, \sigma^2 I)}\left[\mathop{\mathbb{E}}_{y \sim \mathcal{R}}\left[\mathop{\mathbb{P}}_{x \sim \mathcal{N}(y\theta, \sigma^2 I)}\left[\exists x' \in B_\infty^\epsilon(x) : f_n(x') \neq y\right]\right]\right]\right]\right]$$

We re-express this expectation in terms of a new variable $z \sim \mathcal{N}(\theta, \sigma^2 I)$, where $f_n = g_n((y_1 z_1, y_1), ..., (y_n z_n, y_n)))$. We'll soon see that doing this allows us to study the margin of the learned classifier and establish lower bounds on the l_∞^ϵ-robust classification error.

The expectation now becomes

$$\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(0,I)}\left[\mathop{\mathbb{E}}_{y_1,\ldots,y_n\sim\mathcal{R}}\left[\mathop{\mathbb{E}}_{z_1,\ldots,z_n\sim\mathcal{N}(\theta,\sigma^2 I)}\left[\mathop{\mathbb{E}}_{y\sim\mathcal{R}}\left[\mathop{\mathbb{P}}_{x\sim\mathcal{N}(y\theta,\sigma^2 I)}\left[\exists x' \in B^{\epsilon}_{\infty}(x) : f_n(x') \neq y\right]\right]\right]\right]\right] \tag{7}$$

$$= \mathop{\mathbb{E}}_{y_1,\ldots,y_n\sim\mathcal{R}}\left[\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(0,I)}\left[\mathop{\mathbb{E}}_{z_1,\ldots,z_n\sim\mathcal{N}(\theta,\sigma^2 I)}\left[\mathop{\mathbb{E}}_{y\sim\mathcal{R}}\left[\mathop{\mathbb{P}}_{x\sim\mathcal{N}(y\theta,\sigma^2 I)}\left[\exists x' \in B^{\epsilon}_{\infty}(x) : f_n(x') \neq y\right]\right]\right]\right]\right] \tag{8}$$

Swapping expectations over z and θ means the posterior for θ is given by $\theta \sim \mathcal{N}(\mu', \Sigma')$, where

$$\mu' = \frac{n}{\sigma^2 + n}\bar{z}, \qquad \Sigma' = \frac{\sigma^2}{\sigma^2 + n} I, \qquad \bar{z} = \frac{\sum_{i=1}^{n} z_i}{n}.$$

Hence, the expectation becomes

$$\mathop{\mathbb{E}}_{y_1,\ldots,y_n\sim\mathcal{R}}\left[\mathop{\mathbb{E}}_{z_1,\ldots,z_n\sim\mathcal{M}}\left[\mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\mu',\Sigma')}\left[\mathop{\mathbb{E}}_{y\sim\mathcal{R}}\left[\mathop{\mathbb{P}}_{x\sim\mathcal{N}(y\theta,\sigma^2 I)}\left[\exists x' \in B^{\epsilon}_{\infty}(x) : f_n(x') \neq y\right]\right]\right]\right]\right] \tag{9}$$

where $\mathcal{M}$ is some distribution. Let ψ be l^{ϵ}_{∞}-robust classification error on a single sample, i.e.

$$\psi = \mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\mu',\Sigma')}\left[\mathop{\mathbb{E}}_{y\sim\mathcal{R}}\left[\mathop{\mathbb{P}}_{x\sim\mathcal{N}(y\theta,\sigma^2 I)}\left[\exists x' \in B^{\epsilon}_{\infty}(x) : f_n(x') \neq y\right]\right.\right].$$

We now provide a lower bound on ψ. Let $A_+ = \{x | f_n(x) = +1\}$, $A_- = \{x | f_n(x) = -1\}$. We define a perturbation set B in terms of A_- as This set contains the positive labeled points that, when perturbed by some perturbation in B, are labeled negatively by f_n:

$$B^{\epsilon}_{\infty}(A_+) = \left\{x \in A_+ | \exists\ x' \in A_- : ||x - x'||_{\infty} \leq \epsilon\right\}.$$

This set is non-empty whenever $||\mu^{'}||_{\infty} \leq \epsilon$. We define $B^{\epsilon}_{\infty}(A_-)$ analogously; these definitions are visualized in Figure 3. The posterior distribution on θ, which represents the posterior distribution on the positive and negative examples, $\mu^{'}_{+}, \mu^{'}_{-}$ respectively, determines the margin of the classifier hyperplane.

As σ^2 becomes larger than n, μ'_+, μ'_- approach $\bar{0}$ and Σ' approaches I. This corresponds to the margin of the classifier hyperplane decreasing, meaning the classifier cannot separate the two classes well. In particular, the size of the sets $B^{\epsilon}_{\infty}(A_-), B^{\epsilon}_{\infty}(A_+)$ increases as the classifier margin decreases. In Figure 3, this corresponds to the size of the overlap between the two distributions increasing.

In the extreme case, where $\mu'_+ = \bar{0}, \mu'_- = \bar{0}$ and $\Sigma' = I$, the posterior distributions completely overlap. In this case, the l^{ϵ}_{∞}-robust classification error is $\frac{1}{2}$, which is realizable by using even a constant classifier (a classifier that always predicts 1 class label).

Hence, the lower bound on ψ is $\frac{1}{2}$, when $||\mu'||_{\infty} \leq \epsilon$. Plugging this lower bound into the expected l^{ϵ}_{∞}-robust classification error gives (continuing from 9):

$$\geq \mathop{\mathbb{E}}_{y_1,\ldots,y_n\sim\mathcal{R}}\left[\mathop{\mathbb{E}}_{z_1,\ldots,z_n\sim\mathcal{M}}\left[\frac{1}{2}\mathbb{I}\left[||\mu'||_{\infty} \leq \epsilon\right]\right]\right] \tag{10}$$

$$= \frac{1}{2}\mathop{\mathbb{E}}_{z_1,\ldots,z_n\sim\mathcal{M}}\left[\mathbb{I}\left[||\mu'||_{\infty} \leq \epsilon\right]\right] \tag{11}$$

$$= \frac{1}{2}\mathop{\mathbb{P}}_{z_1,\ldots,z_n\sim\mathcal{M}}\left[\frac{n}{n+\sigma^2}||\bar{z}||_{\infty} \leq \epsilon\right] \tag{12}$$

By analyzing the distribution of z wrt θ (ref. A.2 in [2]), we can simplify the expression to

$$\geq \frac{1}{2}\mathop{\mathbb{P}}_{\theta\sim\mathcal{N}(0,I)}\left[\sqrt{\frac{n}{\sigma^2+n}}||\theta||_{\infty} \leq \epsilon\right].$$

■

The above theorem shows that the l_∞^ϵ-robust classification error depends on the margin of the learned classifier. In particular, to reduce the error, we need to increase the probability that the margin is greater than ϵ, i.e.,

$$\sqrt{\frac{n}{\sigma^2+n}}||\theta||_\infty > \epsilon$$

.

This means n needs to be much larger than σ^2. Using the above theorem, we show that the expected l_∞^ϵ-robust classification error of f_n is at least $(1-\frac{1}{d})\frac{1}{2}$ if

$$n \leq \frac{\epsilon^2\sigma^2}{8\log d}$$

Proof We give a lower bound on the following quantity:

$$\mathop{\mathbb{P}}_{v\sim\mathcal{N}(0,I)}\left[\sqrt{\frac{n}{\sigma^2+n}}\|v\|_\infty \leq \epsilon\right]$$

by upper bounding the inner expression given $n \leq \frac{\epsilon^2\sigma^2}{8\log d}$ i.e.,

$$\sqrt{\frac{n}{\sigma^2+n}} \leq \sqrt{\frac{\epsilon^2\sigma^2}{8\sigma^2\log d}} = \frac{\epsilon}{2\sqrt{2\log d}}.$$

Then,

$$\mathop{\mathbb{P}}_{v\sim\mathcal{N}(0,I)}\left[\sqrt{\frac{n}{\sigma^2+n}}\|v\|_\infty \leq \epsilon\right] \geq \mathop{\mathbb{P}}_{v\sim\mathcal{N}(0,I)}\left[\sqrt{\frac{\epsilon}{2\sqrt{2\log d}}}\|v\|_\infty \leq \epsilon\right] = \mathop{\mathbb{P}}_{v\sim\mathcal{N}(0,I)}\left[\|v\|_\infty \leq 2\sqrt{2\log d}\right]$$

This probability term is at least $(1-\frac{1}{d})$ (for proof, see Theorem 5.8 in Section 6 of [2]). Plugging this lower bound into the above theorem proves this corollary. ■

In particular, plugging in for $\sigma = cd^{\frac{1}{4}}$ gives $n \leq \frac{c\epsilon^2\sqrt{d}}{\log d}$. This shows that $n \geq \sqrt{d}$ in order for the classifier to generalize robustly and achieve error less than $\frac{1}{2}$. For very high dimensional problems, this means n needs to be very large. This could explain why for problems on high dimensional datasets e.g., CIFAR10, adversarial robustness requires significantly more samples than required for standard generalization.

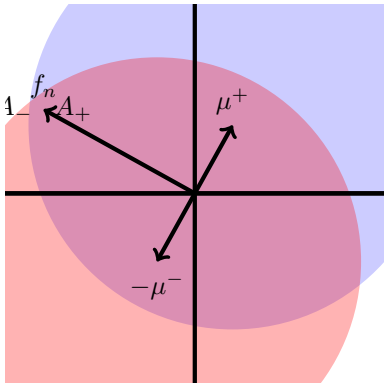

Figure 3: Example of the Gaussian distribution. When the number of samples n is much smaller than the variance σ^2, the classifier f_n cannot separate the two classes well and has poor robust generalization.

References

[1] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[2] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Mądry. Adversarially robust generalization requires more data. In *International Conference on Machine Learning (ICML)*, 2018.