Notes on Maximization of Inner Products over Norm Balls

David Meyer

dmm@{1-4-5.net,uoregon.edu,...}

March 16, 2017

Abstract

This document started life as an exploration into the surprising discovery of *adversarial examples* by Szegedy et al. [3]. This discovery has, among other things, led to new ways of thinking about unsupervised training of deep models [1] while at the same time causing confusion and concern about the nature of learning in such (deep) models. These notes explore the analysis of adversarial examples given in [2] and elsewhere. In particular, I couldn't understand why for linear models, the perturbation $\mathbf{w}^T\boldsymbol{\eta}$ was maximized by setting $\boldsymbol{\eta} = \text{sign}(\mathbf{w})$. The answer is related to what are called Norm Balls and the relationship between norms and inner products.

1 Introduction

The surprising discovery of *adversarial examples* by Szegedy et al. [3] has led to new ways of thinking about unsupervised training of deep models [1] while at the same time causing confusion and concern about the nature of learning in such (deep) models. These notes explore the analysis of adversarial examples given in [2] and elsewhere.

What Szegedy et al. [3] discovered was that several machine learning models, including state-of-the-art neural networks, are vulnerable to adversarial examples. That is, these machine learning models can misclassify examples that are only slightly different (imperceptibly so in many cases) from correctly classified examples drawn from the data distribution. In many cases, a wide variety of models with different architectures trained on different subsets of the training data misclassify the same adversarial example (this is kind of shocking). The implication is that adversarial examples expose fundamental problems in popular training algorithms.

2 Linear Explanation of Adversarial Examples

In [2], a discussion of the "Linear Explanation of Adversarial Examples" explains the existence of adversarial examples for linear models. The section starts with a description of the precision of the sensor or storage media. Here they use the common example digital images, which often use only eight bits per pixel (gray scales) and as a result they discard all information below $\frac{1}{255}$ of the dynamic range. The point here is that since the precision of the features is limited, it makes little sense for a classifier to respond differently to an input $\boldsymbol{x}$ than to an adversarial input $\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\eta}$, if of course every element of the perturbation $\boldsymbol{\eta}$ is smaller than the precision of the features. Let's say that ϵ is the largest value *below* the resolution of the sensor (or storage media). Then for problems with well-separated classes, we expect a classifier to assign the same class to $\mathbf{x}$ and $\hat{\mathbf{x}}$, so long as $||\boldsymbol{\eta}||_\infty < \epsilon$ (where $||\boldsymbol{x}||_\infty$ is the *max* or *infinity* norm).

Next, consider the activation induced by an adversarial example $\hat{\mathbf{x}}$:

$$\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\eta} \qquad \text{\# adversarial example } \hat{\mathbf{x}} \text{ given perturbation } \boldsymbol{\eta} \tag{1}$$

$$\mathbf{w}^{\mathrm{T}}\hat{\boldsymbol{x}} = \mathbf{w}^{\mathrm{T}}\boldsymbol{x} + \mathbf{w}^{\mathrm{T}}\boldsymbol{\eta} \qquad \text{\# activation } \mathbf{w}^{\mathrm{T}}\hat{\boldsymbol{x}} \text{ induced by } \boldsymbol{\eta} \tag{2}$$

Notes:

- The adversarial perturbation causes the activation to grow by $\mathbf{w}^{\mathrm{T}}\boldsymbol{\eta}$
- $\mathbf{w}^{\mathrm{T}}\boldsymbol{\eta}$ is maximized, subject to $||\boldsymbol{\eta}||_\infty < \epsilon$, by assigning $\boldsymbol{\eta} = \text{sign}(\boldsymbol{w})$

What I couldn't understand about Goodfellow's description is why, in the linear case, the perturbation $\mathbf{w}^{\mathrm{T}}\boldsymbol{\eta}$ is maximized, subject to $||\boldsymbol{\eta}||_\infty < \epsilon$, by assigning $\boldsymbol{\eta} = \text{sign}(\boldsymbol{w})$, where the sign or signum function is defined at follows:

$$\text{sign}(x) := \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$

The remainder of this note contains a brief overview of the *Maximization of Inner Products over Norm Balls*, which contains the parts of the answer to my question.

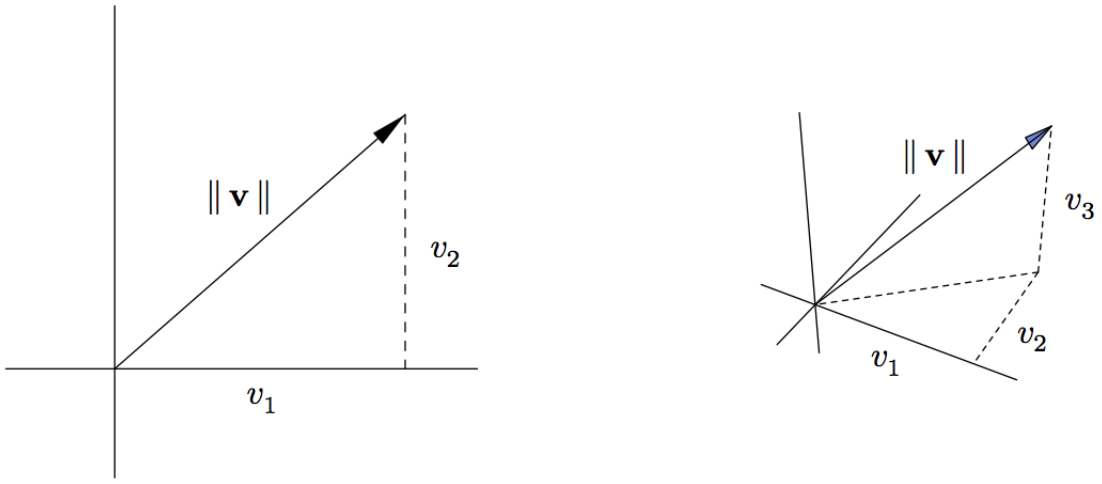

Figure 1: Euclidean Norms in $\mathbb{R}^2$ and $\mathbb{R}^3$

3 The Maximization of Inner Products over Norm Balls

First, what is a norm? The most common definition involves a function $||.|| : \mathbb{R}^n \to \mathbb{R}$, the *vector norm*, which has the following properties:

1. $||\mathbf{x}|| \geq 0$ for any vector $\mathbf{x} \in \mathbb{R}^n$, and $||\mathbf{x}|| = 0$ iff $\mathbf{x} = 0$
2. $||\alpha\mathbf{x}|| = |\alpha|||\mathbf{x}||$ for any vector $\mathbf{x} \in \mathbb{R}^n$ and any scalar α
3. $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$ for any vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

The last property is called the *triangle inequality*. Note also that when $n = 1$, the absolute value function is a vector norm.

The most commonly used vector norms belong to the family of l-norms, or sometimes l_p-norms, which are defined as

$$||\mathbf{x}||_p = \Big(\sum_{i=1}^{n} |x_i|^p \Big)^{\frac{1}{p}} \tag{3}$$

Now, it can be shown that for any $p > 0$, $||.||_p$ defines a vector norm. The vector norms of particular interest include:

- $p = 1$: The ℓ_1-norm $||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|$
- $p = 2$: The ℓ_2-norm or *Euclidean* norm $||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\mathbf{x}^{\mathrm{T}}\mathbf{x}}$
- $p = \infty$: The ℓ_∞-norm $||\mathbf{x}||_\infty = \max_{1 \le i \le n} |x_i|$

Figure 1 shows the Euclidean Norms in $\mathbb{R}^2$ and $\mathbb{R}^3$.

There are many (many) useful theorems one can prove here, but I want to jump ahead to norm balls.

3.1 What is a Norm Ball?

Again there are many theorems one can prove about norm balls, but suffice it to say that in Cartesian space $\mathbb{R}^n$ with the p-norm, ℓ_p, an open ball is the set

$$B(r) = \left\{ \mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^{n} |x_i|^p < r^p \right\} \qquad \# \, r \text{ is the radius of ball } B \tag{4}$$

For $n = 2$, in particular, the balls of ℓ_1 (often called the taxicab or Manhattan metric) are squares with the diagonals parallel to the coordinate axes; those of ℓ_∞ (the Chebyshev metric) are squares with the sides parallel to the coordinate axes. For other values of p, the balls are bounded by Lam curves (hypoellipses or hyperellipses). For $n = 3$, the balls of ℓ_1 are octahedra with axis-aligned body diagonals, those of ℓ_∞ are cubes with axis-aligned edges, and those of ℓ_p with $p > 2$ are superellipsoids. See Figure 2 for a few examples.

4 Maximization of Inner Products over Norm Balls

Finally we're getting to the heart of the matter. There is an interesting relationship between inner products and norms, but I'm going to skip that here. Rather, consider the following: Given a nonzero vector $\mathbf{y} \in \mathbb{R}^n$,, consider the problem of finding some vector $\mathbf{x} \in \mathcal{B}_p$ (i.e., the unit ball in ℓ_p norm) that maximizes the inner product $\mathbf{x}^{\mathrm{T}}\mathbf{y}$. That is, given some nonzero vector $\mathbf{y}$, we want to solve for $\mathbf{x}$ that satisfies $\max_{||\mathbf{x}||_p \le 1} \mathbf{x}^{\mathrm{T}}\mathbf{y}$.

For $p = 2$, the solution is straight forward since $\mathbf{x}^{\mathrm{T}}\mathbf{y} = \cos\theta ||\mathbf{x}||_2 ||\mathbf{y}||_2$; $\mathbf{x}$ should be parallel to $\mathbf{y}$ (i.e., the angle between $\mathbf{x}$ and $\mathbf{y}$ is zero) so that the norm is as large as possible, that is, one. The unique solution is here is $\mathbf{x}_2^* = \frac{\mathbf{y}}{||\mathbf{y}||_2}$.

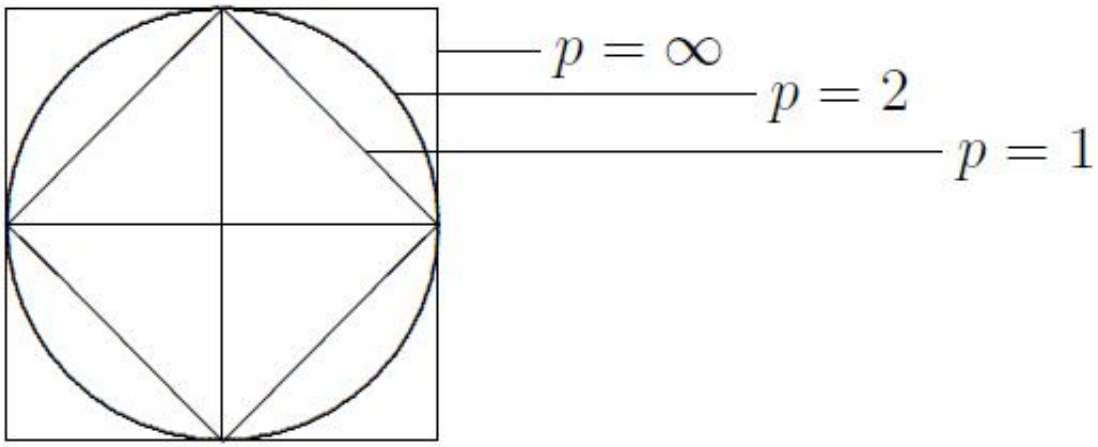

Figure 2: Norm Balls in $\mathbb{R}^2$

This is where things get (more) interesting and relevant to my question as to why $\boldsymbol{\eta}$ is set to sign($\mathbf{w}$) in Goodfellow's work. So consider the case where $p = \infty$, i.e., the max norm. Now, since $\mathbf{x}^{\mathrm{T}}\mathbf{y} = \sum_{i=1}^{n} x_i y_i$, where each element x_i is such that $|x_i| \leq 1$, we can see that the sum is maximized by setting $x_i = \text{sign}(y_i)$. Further, we can see that $\max_{||\mathbf{x}||_\infty \leq 1} \mathbf{x}^{\mathrm{T}}\mathbf{y} = \sum_{i=1}^{n} |y_i| = ||\mathbf{y}||_1$. So here we see that the optimal solution may not be unique, since any $x_i \in [-1, 1]$, corresponding to $y_i = 0$, could be selected without modifying the optimal value.

Finally, for completeness, consider the case where $p = 1$. Here the inner product $\mathbf{x}^{\mathrm{T}}\mathbf{y} = \sum_{i=1}^{n} x_i y_i$ can be interpreted as a weighted average of the y_i's where the x_i's are the weights whose absolute values must sum up to one. The maximum of the weighted average is achieved by first finding the y_i having the largest absolute value, that is, by finding one index m such that $|y_i| \leq |y_m|$ for all $i = 1, 2, \cdots, n$, and then setting

$$|x_1^*|_i = \begin{cases} \text{sign}(y_i) & \text{if } i = m \\ 0 & \text{otherwise} \end{cases}$$

What we wind up with there is that $\max_{||\mathbf{x}||_1 \leq 1} \mathbf{x}^{\mathrm{T}}\mathbf{y} = \max_i |y_i| = ||\mathbf{y}||_\infty$. Here again the optimal solution may not be unique since in the case when $\mathbf{y}$ has more than one entry

with maximum absolute value we can choose m to be any of the indicies corresponding to these maxima. This shows the goal of maximizing inner products over norm balls, namely to create a n-dimensional unit vector that goes from the origin to the surface of the norm ball.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680.

[2] Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *https://arxiv.org/abs/1412.6572* (2014).

[3] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *https://arxiv.org/abs/1312.6199* (2013).