Overparametrized regression under ℓ_2 adversarial attacks

Antônio H. Ribeiro, Thomas B. Schön
Uppsala University

We derive upper and lower bounds for the adversarial risk in linear regression. Then—using results from random matrix theory—we present the asymptotic bounds for ℓ_2 attacks. We show that the characteristic second descent in the overparametrized region is still present. The result is also confirmed experimentally.

1. Introduction

Deep learning models have achieved impressive performance in many tasks. At the same time these models turn out to be quite brittle in some situations. This is evidenced by studying adversarial examples, which show that synthesized imperceptible perturbations of the input data may cause the model to make highly confident, but erroneous predictions.

In another line of work, the good performance of these models in test data—which almost perfectly fit the training data—has motivated the study of overparametrized models and lead to new findings. A key observation that followed is the presence of a second descent in the risk as we increase the model capacity beyond the point of (almost) perfectly fitting the training data (Belkin et al., 2019).

In this work, we connect the two ideas and study the effect of adversarial attacks in overparametrized models. We consider perhaps the simplest setting where a second descent in risk has been observed: linear regression (Hastie et al., 2019; Bartlett et al., 2020). The purpose of using a simplified setting is twofold: first, to make it amenable to theoretical analysis; and, second, to make it possible to isolate the role of overparametrization in the adversarial performance. That is, deep learning systems usually contain many design parameters that interfere with each other in non-obvious ways and could conceal the role of overparametrization in the adversarial performance.

2. Setup

Data generating model. Assume the training data $(x_i, y_i) \in \mathbb{R}^m \times \mathbb{R}$ for $i = 1, \dots, n$ was generated from

$$(x_i, \epsilon_i) \sim P_x \times P_\epsilon, \tag{1a}$$

$$y_i = x_i^\mathsf{T} \beta + \epsilon_i, \tag{1b}$$

for which the random draws across $i = 1, \dots, n$ are independent. Here P_x is a distribution on $\mathbb{R}^p$, such that $E\{x\} = 0$ and $\text{Cov}\{x\} = \Sigma$. Furthermore, P_ϵ is an independent distribution in $\mathbb{R}$ such that $E\{\epsilon\} = 0$ and $\text{Var}\{\epsilon\} = \sigma$.

Notation. Here, $\|v\|_p$ denotes the p-norm of the vector v. And, given a symmetric matrix Σ, we denote $\|v\|_\Sigma = v^\mathsf{T} \Sigma v$.

Risk. The adversarial risk is

$$R_p^{\text{adv}} = E\left\{ \max_{\|\Delta x_0\|_p \le \delta} (y_0 - (x_0 + \Delta x_0)^\mathsf{T} \hat{\beta})^2 \middle| X \right\}.$$

The standard out-of-sample risk $E\left\{(y_0 - x_0^\mathsf{T}\hat{\beta})^2 \middle| X\right\}$ would correspond to the case there is no disturbance, i.e. $\delta = 0$.

Minimum norm solution. We assume the parameters have been estimated as

$$\hat{\beta} = (X^\mathsf{T} X)^\dagger X^\mathsf{T} y, \tag{2}$$

where $(X^\mathsf{T} X)^\dagger$ represents the pseudo-inverse of $X^\mathsf{T} X$. In the overparametrized $(m > n)$ case, when more then one solution is possible, this correspond to the solution for which the parameter norm $\|\beta\|_2$ is minimum.

3. Results

We first provide upper and lower bounds on the adversarial risk for adversarial attacks bounded by generic p-norms.

Lemma 1 (Bounds on the adversarial risk)**.** *For $1 < p < \infty$, let q be a positive real number for which $\frac{1}{p} + \frac{1}{q} = 1$. Let us denote $R_\Sigma = E\left\{\|\beta - \hat{\beta}\|_\Sigma^2 \middle| X\right\}$, $N_q = E\left\{\|\hat{\beta}\|_q^2 \middle| X\right\}$, then the adversarial risk is bounded,*

$$R_\Sigma + \delta^2 N_q + \sigma^2 \le R_p^{adv} \le \left(\sqrt{R_\Sigma} + \delta\sqrt{N_q}\right)^2 + \sigma^2. \tag{3}$$

The result also holds when $p = 1$ or $p = \infty$ for, respectively, $q = \infty$ and $q = 1$. Furthermore, for the case when there is no adversarial disturbance (i.e., $\delta = 0$) the equality holds.

We now limit the scope to the case $p = q = 2$ because N_2 can be treated similarly to R_Σ

in this case. The next lemma presents closed-form expressions for R_Σ and N_2 in terms of the problem matrices.

Lemma 2 (Bias-variance decomposition)**.** *We define* $\hat{\Sigma} = \frac{1}{n}X^\mathsf{T}X$, $\Phi = \hat{\Sigma}^\dagger\hat{\Sigma}$ *and* $\Pi = I - \Phi$. *Where* Φ *and* Π *are orthogonal projectors:* Π *is the projection into the null space of* X *and* Φ*, into the row space of* X*. Then:*

$$R_\Sigma = \beta^\mathsf{T}\Pi\Sigma\Pi\beta + \frac{\sigma^2}{n} tr(\hat{\Sigma}^\dagger\Sigma). \tag{4}$$

Similarly, for $q = 2$*:*

$$N_2 = \beta^\mathsf{T}\Phi\beta + \frac{\sigma^2}{n} tr(\hat{\Sigma}^\dagger). \tag{5}$$

Next we present the asymptotics for R_Σ and N_2, as $n, m \to \infty$ while keeping the ratio $m/n \to \gamma$. The risk R_Σ is extensively studied by Hastie et al. (2019) and the next result is presented in it for the case the features x_i are i.i.d. It is proved using more or less standard random matrix theory results. Other scenarios (correlated features or misspecified models) are also studied in (Hastie et al., 2019) and could be used here with minor modifications.

Lemma 3 (Asymptotics)**.** *Assume that* x_i *are i.i.d. and has a moment of order greater then 8 that is finite. Assume that* $\|\beta\|_2^2 = r^2$*. Then, as* $n, m \to \infty$ $m/n \to \gamma$*, it holds almost surely that:*

$$R_I \to \begin{cases} \sigma^2\frac{\gamma}{1-\gamma}, \gamma < 1, \\ r^2(1 - \frac{1}{\gamma}) + \sigma^2\frac{1}{\gamma-1}, \gamma > 1. \end{cases} \tag{6}$$

For $q = 2$*, it holds almost surely that:*

$$N_2 \to \begin{cases} r^2 + \sigma^2\frac{\gamma}{1-\gamma}, \gamma < 1, \\ r^2\frac{1}{\gamma} + \sigma^2\frac{1}{\gamma-1}, \gamma > 1. \end{cases} \tag{7}$$

Plugging R_Σ and N_q asymptotics back into (3) we can establish an asymptotic bounds on the adversarial risk. This bounds together with the points obtained by the sample risk obtained experimentally are displayed in Fig. 1, showing that the characteristic second descent curve after the interpolation threshold appears also for the adversarial ℓ_2 risk.

4. Discussion

Classifiers relying on many features with small correlation with the output are non-robust (Ilyas et al., 2019). They can have good normal performance, but the performance can degrade quickly under ℓ_∞ adversarial attack. Tsipras et al. (2019) show this both in a linear classification setting and, also, in more involved deep learning examples.

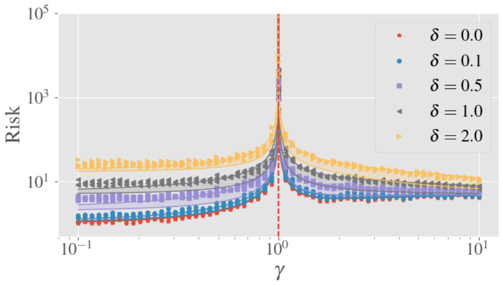

Figure 1: **Adversarial risk**. The solid lines gives the asymptotic upper and lower bounds computed using Lemmas 1 and 3. The points correspond to the empirical adversarial risk for experiments with $n = 100$. The results are for $r^2 = 2$, $\sigma^2 = 1$.

In overparametrized systems, the amplitude of individual features in the model prediction reduces with the number of features and, hence, the correlation with the output. Hence, it is expected that the ℓ_∞ adversarial performance will degrade as we increase the number of features.

The result we present here shows that this is not necessarily the case for ℓ_2 adversarial attacks and that the second descent in the risk can still be observed in the interpolation regime. The future direction of this work is to use the same framework to study the behaviour of adversarial attacks bounded by ℓ_1, ℓ_∞ and other p norms, which appear often in the study of adversarial attacks.

References

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *PNAS*, 117(48):30063–30070, 2020.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *PNAS*, 116(32):15849–15854,2019.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in High-Dimensional Ridgeless Least Squares Interpolation. *arXiv:1903.08560*, 2019.

A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial Examples Are Not Bugs, They Are Features. *Advances in Neural Information Processing Systems 32*, 2019.

D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Ma. Robustness May Be At Odds with Accuracy. *ICLR*, 2019.

Appendices

A. Proof of Lemma 1

The adversarial risk,

$$R_p^{\text{adv}} = E\left\{ \max_{\|\Delta x_0\|_p \leq \delta} (y_0 - (x_0 + \Delta x_0)^\intercal \hat{\beta})^2 \middle| X \right\}, \tag{8}$$

can be rewritten, after some algebraic manipulation, as

$$R_p^{\text{adv}} = E\left\{ \max_{\|\Delta x_0\|_p \leq \delta} \left((\Delta x_0^\intercal \hat{\beta})^2 - 2(\beta - \hat{\beta})^\intercal x_0 \Delta x_0^\intercal \hat{\beta} \right) \middle| X \right\} + \underbrace{E\left\{ \|\beta - \hat{\beta}\|_\Sigma^2 \middle| X \right\}}_{R_\Sigma} + \sigma^2. \tag{9}$$

In turn, let $r = \Delta x_0^\intercal \hat{\beta}$ and $a = (\beta - \hat{\beta})^\intercal x_0$. It follows from Hölder inequality, that $\|\Delta x_0\|_p \leq \delta$ implies that $|r| \leq \delta \|\hat{\beta}\|_q$, for q satisfying $1/p + 1/q = 1$. Since we can always choose vectors such that the equality holds, then the term inside the expectation of the second hand side is equal to: $M = \max_{|r| \leq \delta \|\hat{\beta}\|_q} (r^2 - 2ar)$. Now the maximum is attained at $r = -\delta \|\hat{\beta}\|_q$ if $a \geq 0$ and at $r = \delta \|\hat{\beta}\|_q$ if $a < 0$, hence $M = \delta^2 \|\hat{\beta}\|_q^2 + 2\delta \|\hat{\beta}\|_q |a|$. It follows that:

$$R_p^{\text{adv}} = \delta^2 \underbrace{E\left\{ \|\hat{\beta}\|_q^2 | X \right\}}_{N_q} + 2\delta E\left\{ \|\hat{\beta}\|_q |(\beta - \hat{\beta})^\intercal x_0| \middle| X \right\} + \underbrace{E\left\{ \|\beta - \hat{\beta}\|_\Sigma^2 \middle| X \right\}}_{R_\Sigma} + \sigma^2. \tag{10}$$

In turn,

$$0 \leq E\left\{ \|\hat{\beta}\|_q |(\beta - \hat{\beta})^\intercal x_0| \middle| X \right\} \leq \sqrt{E\left\{ \|\hat{\beta}\|_q^2 \middle| X \right\} E\left\{ \|\beta - \hat{\beta}\|_\Sigma^2 \middle| X \right\}} = \sqrt{N_q R_\Sigma}. \tag{11}$$

where the second inequality is a direct application of the Cauchy-Schwartz inequality. The results follows.

B. Proof of Lemma 2

Proof for N_2. From Eq. (1b) and Eq. (2) it follows that:

$$\hat{\beta} = \underbrace{(X^\intercal X)^\dagger X^\intercal X}_{\Phi} \beta + \underbrace{(X^\intercal X)^\dagger}_{\frac{1}{n}\hat{\Sigma}^\dagger} X^\intercal \epsilon. \tag{12}$$

Hence, since $\hat{\Sigma}$ is symmetric:

$$\hat{\beta}^\intercal \hat{\beta} = \beta^\intercal \Phi^\intercal \Phi \beta + \frac{1}{n^2} \epsilon^\intercal X \hat{\Sigma}^\dagger \hat{\Sigma}^\dagger X^\intercal \epsilon, \tag{13}$$

where the first term is equal to $\beta^\intercal \Phi \beta$ since, Φ is symmetric i.e., $\Phi^T = \Phi$, and since it is a projector i.e., $\Phi\Phi = \Phi$. Matrix that satisfy these two properties are called orthogonal projectors.

Now, since the second term is a scalar it is equal to its trace. Using the fact that the trace is invariant over cyclic permutations,

$$\epsilon^{\mathsf{T}} X \hat{\Sigma}^{\dagger} \hat{\Sigma}^{\dagger} X^{\mathsf{T}} \epsilon = \mathrm{tr}\left\{\hat{\Sigma}^{\dagger} X^{\mathsf{T}} \epsilon \epsilon^{\mathsf{T}} X \hat{\Sigma}^{\dagger}\right\}. \tag{14}$$

From the assumption the noise samples are independent and have variance σ^2, we have $E\{\epsilon\epsilon^{\mathsf{T}}\} = \sigma^2 I$, where I is the identity matrix. Since we can swap the trace and the expectation we obtain

$$E\left\{\hat{\beta}^{\mathsf{T}}\hat{\beta}\middle| X\right\} = \beta^{\mathsf{T}}\Phi\beta + \frac{1}{n^2}\mathrm{tr}\left\{\hat{\Sigma}^{\dagger} X^{\mathsf{T}} \underbrace{E\{\epsilon\epsilon^{\mathsf{T}}\}}_{\sigma^2 I} X \hat{\Sigma}^{\dagger}\right\}. \tag{15}$$

And the results follow from the definition of $\hat{\Sigma}^{\dagger}$ and the property of pseudoinverse: $\hat{\Sigma}^{\dagger}\hat{\Sigma}\hat{\Sigma}^{\dagger} = \hat{\Sigma}^{\dagger}$
Proof for R_{Σ} . From (12), it follows that:

$$\beta - \hat{\beta} = \underbrace{(I - \Phi)}_{\Pi}\beta + \frac{1}{n}\hat{\Sigma}^{\dagger} X^{\mathsf{T}}\epsilon. \tag{16}$$

where, again Π is a orthogonal projector i.e., $\Pi^T = \Pi$ and $\Pi\,\Pi = \Pi$. We can then compute the close formula (4)for R_{Σ} using exact the same procedure as we did for N_2.

C. Proof of Lemma 3

Lemma 3 proof is given in (Hastie et al., 2019) Section 2.4 and Section 3.