Overparametrized regression under ℓ_2 adversarial attacks

Antônio H. Ribeiro, Thomas B. Schön

Uppsala University, Sweden

UPPSALA
UNIVERSITET

Workshop on the Theory of Overparameterized Machine Learning
TOPML, 2021

Overparametrized models can generalize effectively when train and test come from the same *distribution...*

Overparametrized models can generalize effectively when train and test come from the same *distribution...*

Can it also generalize effectively when there is a distribution shift?

Adversarial attacks

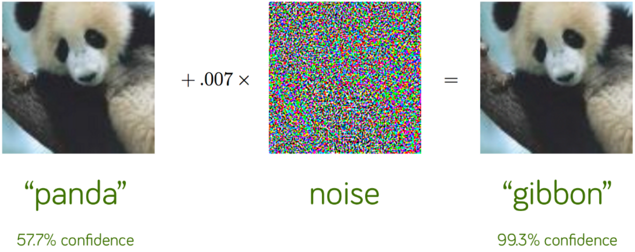


Figure: Illustration of adversarial attack. From: I.J. Goodfellow, J. Shlens, C.Szegedy , *"Explaining and Harnessing Adversarial Examples"*, ICLR 2015.

Linear regression under adversarial attacks

Model: Linear model

$$y = x^{\mathsf{T}}\beta + \epsilon$$

Linear regression under adversarial attacks

Model: Linear model

$$y = x^{\mathsf{T}}\beta + \epsilon$$

Estimated parameter: using train dataset (x_i, y_i), $i = 1, \cdots, n$:

- Underparametrized:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i} (y_i - x_i^{\mathsf{T}}\beta)^2$$

Linear regression under adversarial attacks

Model: Linear model

$$y = x^\mathsf{T}\beta + \epsilon$$

Estimated parameter: using train dataset (x_i, y_i), $i = 1, \cdots, n$:

- Underparametrized:

$$\hat{\beta} = \arg\min_{\beta} \sum_i (y_i - x_i{}^\mathsf{T}\beta)^2$$

- Overparametrized:

$$\begin{aligned} \hat{\beta} =& \arg\min_{\beta} \|\beta\|_2^2 \\ & \text{subject to } y_i = x_i{}^\mathsf{T}\beta \\ & \text{for every } i \end{aligned}$$

Linear regression under adversarial attacks

Given a data point not seen during training (x, y).

Standard risk:

$$(y - x^\mathsf{T} \hat{\beta})^2$$

Adversarial risk:

$$(y - (x + \Delta x)^\mathsf{T} \hat{\beta})^2$$

Linear regression under adversarial attacks

Given a data point not seen during training (x, y).

Standard risk:

$$(y - x^\mathsf{T}\hat{\beta})^2$$

Adversarial risk:

$$\max_{\|\Delta x\|_p \leqslant \delta} (y - (x + \Delta x)^\mathsf{T}\hat{\beta})^2$$

$\Delta x \rightsquigarrow$ Adversarially generated disturbance

Linear regression under adversarial attacks

Given a data point not seen during training (x, y).

Standard risk:

$$R = E\left\{(y - x^{\mathsf{T}}\hat{\beta})^2 \qquad\qquad\qquad \right\}$$

Adversarial risk:

$$R^{\text{adv}} = E\left\{\max_{\|\Delta x\|_p \leqslant \delta} (y - (x + \Delta x)^{\mathsf{T}}\hat{\beta})^2 \qquad\qquad\qquad \right\}$$

$\Delta x \rightsquigarrow$ Adversarially generated disturbance

Linear regression under adversarial attacks

Given a data point not seen during training (x, y).

Standard risk:

$$R = E\left\{(y - x^{\mathsf{T}}\hat{\beta})^2 | \underbrace{x_i, i = 1, \cdots, n}_{\text{training inputs}}\right\}$$

Adversarial risk:

$$R^{\text{adv}} = E\left\{\max_{\|\Delta x\|_p \leqslant \delta} (y - (x + \Delta x)^{\mathsf{T}}\hat{\beta})^2 | \underbrace{x_i, i = 1, \cdots, n}_{\text{training inputs}}\right\}$$

$\Delta x \rightsquigarrow$ Adversarially generated disturbance

Results

Bounds on the adversarial risk:

$$R + \delta^2 N_q + \sigma^2 \leqslant R^{\text{adv}} \leqslant \left(\sqrt{R} + \delta\sqrt{N_q}\right)^2 + \sigma^2.$$

Results

Bounds on the adversarial risk:

$$R + \delta^2 N_q + \sigma^2 \leqslant R^{\text{adv}} \leqslant \left(\sqrt{R} + \delta\sqrt{N_q}\right)^2 + \sigma^2.$$

$R^{\text{adv}} \rightsquigarrow$ Adversarial risk

Results

Bounds on the adversarial risk:

$$R + \delta^2 N_q + \sigma^2 \leqslant R^{\text{adv}} \leqslant \left(\sqrt{R} + \delta\sqrt{N_q}\right)^2 + \sigma^2.$$

$R^{\text{adv}} \rightsquigarrow$ Adversarial risk
$R \rightsquigarrow$ Risk

Results

Bounds on the adversarial risk:

$$R + \delta^2 N_q + \sigma^2 \leqslant R^{\text{adv}} \leqslant \left(\sqrt{R} + \delta\sqrt{N_q}\right)^2 + \sigma^2.$$

$R^{\text{adv}} \rightsquigarrow$ Adversarial risk
$R \rightsquigarrow$ Risk
$\mathrm{N}_q = E\{\|\hat{\beta}\|_q^2\}$

Results

Bounds on the adversarial risk:

$$R + \delta^2 N_q + \sigma^2 \leqslant R^{\text{adv}} \leqslant \left(\sqrt{R} + \delta\sqrt{N_q}\right)^2 + \sigma^2.$$

$R^{\text{adv}} \rightsquigarrow$ Adversarial risk
$R \rightsquigarrow$ Risk
$\mathrm{N}_q = E\{\|\hat{\beta}\|_q^2\}$
$\rightarrow$ For an ℓ_p attack $\rightsquigarrow \frac{1}{\mathrm{q}} + \frac{1}{\mathrm{p}} = 1$

Results

Bounds on the adversarial risk:

$$R + \delta^2 N_q + \sigma^2 \leqslant R^{\mathrm{adv}} \leqslant \left(\sqrt{R} + \delta\sqrt{N_q}\right)^2 + \sigma^2.$$

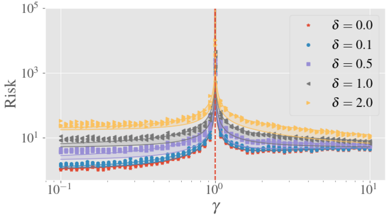
$R^{\mathrm{adv}} \rightsquigarrow$ Adversarial risk
$R \rightsquigarrow$ Risk
$\mathrm{N}_q = E\{\|\hat{\beta}\|_q^2\}$
 $\rightarrow$ For an ℓ_p attack $\rightsquigarrow \frac{1}{\mathrm{q}} + \frac{1}{\mathrm{p}} = 1$
$\delta \rightsquigarrow$ Adversarial disturbance magnitude

Results

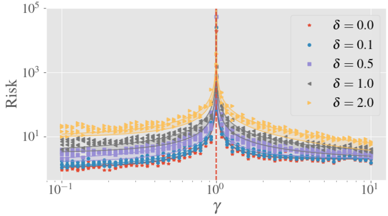
For the ℓ_2-norm we know the asymptotic for all of the above quantities!

Results

For the ℓ_2-norm we know the asymptotic for all of the above quantities!



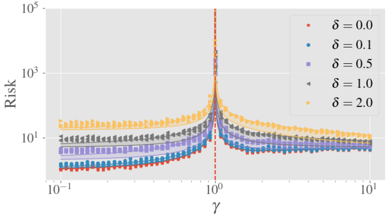
(a) Uncorrelated feature



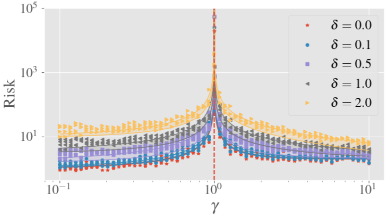
(b) Equicorrelated features ($\rho = 0.5$)

Figure: **Adversarial ℓ_2 risk**. The solid line gives the asymptotic risk lower and upper bound computed. The points correspond to the empirical adversarial risk for experiments.

For the ℓ_2-norm we know the asymptotic for all of the above quantities!



(a) Uncorrelated feature



(b) Equicorrelated features ($\rho = 0.5$)

Figure: **Adversarial ℓ_2 risk**. The solid line gives the asymptotic risk lower and upper bound computed. The points correspond to the empirical adversarial risk for experiments.

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

- This is not in line with what is observed for ℓ_∞ adversarial attacks ...

- This is not in line with what is observed for ℓ_∞ adversarial attacks ...
- There adversarial performance degrades as we increase the number of features. See, for instance:

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

- This is not in line with what is observed for ℓ_∞ adversarial attacks ...
- There adversarial performance degrades as we increase the number of features. See, for instance:
  D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Ma. Robustness May Be At Odds with Accuracy. ICLR, 2019.
- Different l_p adversarial attacks may behave *qualitatively* different as we increase the number of parameters...

Thank you!

Contact info:

antonio.horta.ribeiro@it.uu.se
thomas.schon@it.uu.se

@ahortaribeiro

antonior92.github.io
user.it.uu.se/~thosc112

github.com/antonior92
github.com/thomasschon