Explaining Adversarial Vulnerability with Differential Geometry

Loïc Shi-Garrier[1], Nidhal C. Bouaynaya[1,2], Daniel Delahaye[1]
loic.shi-garrier@enac.fr, bouaynaya@rowan.edu, daniel.delahaye@enac.fr
[1]Ecole Nationale de l'Aviation Civile, Université de Toulouse, France.
[2]Dept. of Electrical and Computer Engineering, Rowan University, New Jersey, USA.

Introduction

Adversarial vulnerability can be defined as the extreme sensitivity of learning models' predictions to small perturbations of their inputs.
Despite outstanding practical achievements, a consensus is yet to emerge to fully explain this phenomenon.
In this work, we explore how methods borrowed from differential geometry could shed light on adversarial vulnerability.

Framework

Multi-class classification:

$$f : \mathcal{X} \subseteq \mathbb{R}^d \to \Delta^{c-1}, \quad x \mapsto \theta.$$

The probability simplex Δ^{c-1} is the parameter space of the family of categorical distributions.
Fisher information metric on Δ^{c-1}:

$$g_\theta = \sum_{i,j=1}^{c} \frac{1}{\theta^i} \delta_{ij} d\theta^i d\theta^j.$$

Pullback metric $\widetilde{g} = f^* g$. In coordinates:

$$\widetilde{G}_x = J_x^T G_{f(x)} J_x.$$

Adversarial robustness

Let $x \in \mathcal{X}$.
Let $\mathcal{D}$ be the decision boundary in Δ^{c-1}, and $\delta(x) = d(f(x), \mathcal{D})$ be the distance between $f(x)$ and $\mathcal{D}$.
Let $\mathcal{B}(x, \epsilon)$ be the Euclidean ball, and let $\widetilde{\mathcal{B}}(x, \delta(x))$ be the geodesic ball induced by $\widetilde{g}$.
Robustness criteria. If:

$$\mathcal{B}(x, \epsilon) \subseteq \widetilde{\mathcal{B}}(x, \delta(x)),$$

then the model f is adversarially robust at x.

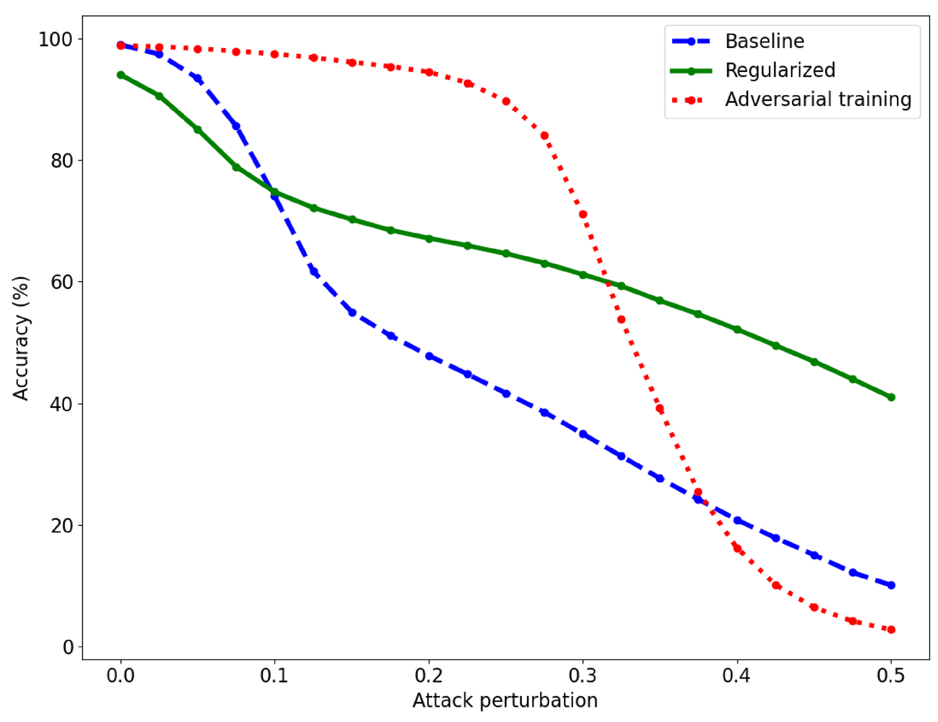
Isometry regularization

Let $P_x = (\ker \widetilde{g}_x)^\perp$, and let $\overline{g}$ be the Euclidean metric. The isometry condition is:

$$\widetilde{g}_x|_{P_x} = \frac{\delta(x)^2}{\epsilon^2} \overline{g}_x|_{P_x}.$$

We define the following regularization:

$$\alpha(x, f) = \left\| \widetilde{J}_x \widetilde{J}_x^T - \frac{\delta(x)^2 \rho(x)^2}{\epsilon^2} I_{c-1} \right\|_F.$$



Accuracy of the baseline (blue), regularized (green), and adversarially trained (red) models for various attack perturbations on the MNIST dataset. The perturbations are obtained with PGD using l_∞ norm.

Jacobian regularization

The robustness condition is, for all $X \in T_x\mathcal{X}$:

$$\widetilde{g}_x(X, X) \leq \frac{\delta(x)^2}{\epsilon^2} \overline{g}_x(X, X).$$

We define the following regularization:

$$\alpha(x, f) = h\left(\|\widetilde{J}_x\|_2^2 - \frac{\delta(x)^2}{\kappa(x)^2 \epsilon^2} \right),$$

where h is a "soft barrier function".

Research directions

1. **Certified defense**
   Derive a certified defense by strongly enforcing the robustness criteria on a chosen proportion of the training examples. Can we prove that the accuracy is maximized under the constraint of a chosen robustness level?
2. **Extensions**
   The proposed regularizations focus on l_2 white-box attacks for multi-class classification. It can be extended to regression tasks (e.g., using the family of multivariate normal distributions) as well as to other attacks (e.g., l_∞ attacks or unrestricted attacks such as spatial attacks).
3. **Other metric**
   Find another metric or another family of distributions such that the robustness criteria is optimal (i.e., $\widetilde{\mathcal{B}}(x, \delta(x))$ is exactly the set of points connected to x with the same class than x).
4. **Exact robustness criteria**
   Derive an exact robustness criteria by taking into account the curvature of $\widetilde{g}$. Is there a formulation of this exact robustness criteria that is computationally tractable?
5. **Data leaf**
   Consider the distribution $P : x \mapsto (\ker \widetilde{g}_x)^\perp$. Under mild assumptions [1], the distribution P is integrable. Moreover, the underlying data distribution may be supported on a unique leaf of the foliation associated to P. The data leaf framework may explain why the generalization and robustness properties are dependent on the training set distribution.

References

[1] Luca Grementieri and Rita Fioresi. Model-centric data manifold: The data through the eyes of the model. *SIAM Journal on Imaging Sciences*, 15(3):1140–1156, 2022.

[2] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional smoothing with virtual adversarial training, 2015.

[3] Aran Nayebi and Surya Ganguli. Biologically inspired protection of deep networks from adversarial attacks, 2017.

[4] Chenxiao Zhao, P. Thomas Fletcher, Mixue Yu, Yaxin Peng, Guixu Zhang, and Chaomin Shen. The Adversarial Attack and Detection under the Fisher Information Metric. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1):5869–5876, 2019.

[5] Jörg Martin and Clemens Elster. Inspecting adversarial examples using the Fisher information. *Neurocomputing*, 382:80–86, 2020.

[6] Yujun Shi, Benben Liao, Guangyong Chen, Yun Liu, Ming-Ming Cheng, and Jiashi Feng. Understanding adversarial behavior of dnns by disentangling non-robust and robust components in performance metric, 2019.

[7] Chaomin Shen, Yaxin Peng, Guixu Zhang, and Jinsong Fan. Defending against adversarial attacks by suppressing the largest eigenvalue of fisher information matrix, 2019.

[8] Marine Picot, Francisco Messina, Malik Boudiaf, Fabrice Labeau, Ismail Ben Ayed, and Pablo Piantanida. Adversarial robustness via fisher-rao regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[9] Eliot Tron, Nicolas Couellan, and Stéphane Puechmorel. Canonical foliations of neural networks: application to robustness, 2022.