Explaining Adversarial Vulnerability with Information Geometry

Anonymous authors

I. Introduction and context

Adversarial vulnerability can be informally described as the extreme sensitivity of learning models' predictions to small perturbations of their inputs. Two surprising facts about adversarial examples are their ubiquity and their tendency to be transferable [1]. Besides hurting the intuition, adversarial vulnerability is a symptom of a broader and very concerning issue: the lack of robustness of deep learning models. This issue impedes the further developments and implementations of deep learning models in critical domains, such as air transportation, where the behavior of the model must be certified [2].

For the last decade, the research community has developed increasingly powerful attacks and increasingly efficient defenses [3]. Despite these practical achievements, a consensus is yet to emerge to fully explain the phenomenon of adversarial vulnerability [4], [5], [6], [7].

In this work, we extend an under-explored yet promising direction to explain adversarial vulnerability of learning models. Following works such as [8], [9] [10], [11], [12], [13], [14], [15], [16], we see the inputs of a supervised learning model as parameters of a family of probability distributions corresponding to the predictions of the model. It is then possible to define a low-dimensional Riemannian submanifold (called the *data leaf*) on the data space. The data leaf is a geometrical description of the data as seen by the model for the task it has been trained for. In particular, it encodes important properties of the model such as its generalization capabilities as well as its robustness. We study how the properties of the data leaf provide an unified explanation of adversarial vulnerability and highlight potential applications of this framework.

II. A geometrical framework for the robustness of learning models

A. The data leaf

Let $\mathcal{F} = \{p(y|\theta)\}$ be a family of probability density functions. For example, the family $\mathcal{F}$ can be chosen as the probability simplex for classification tasks and the family of normal distributions for regression tasks. The parameter θ belongs to a subset $\Theta \subseteq \mathbb{R}^m$ called the parameter space. $\mathcal{F}$ can be seen as a topological manifold with a global coordinate system θ. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be the data space (i.e., the set containing the data). A supervised learning model N is defined as a smooth application

$$\begin{aligned} N : \mathcal{X} &\to \Theta \\ x &\mapsto \theta. \end{aligned}$$

Equipped with the Fisher Information metric (FIM) G_θ defined by

$$G_\theta = \mathbb{E}_{y\sim p(y|\theta)}[\nabla_\theta \log p(y|\theta)(\nabla_\theta \log p(y|\theta))^T],$$

the family $\mathcal{F}$ becomes a Riemannian manifold. The data space $\mathcal{X}$ can be equipped with a pseudo-metric G_x defined as the pullback of the FIM by the model. In coordinate, we get $G_x = J_x^T G_\theta J_x$ where J_x is the Jacobian matrix of $\theta = N(x)$ with respect to x. Following [14], we call G_x the local data matrix. Consider the distribution $D : x \mapsto (\ker G_x)^\perp$ where the orthogonal $\perp$ is taken according to the Euclidean metric. In [14], the authors show that, under mild assumptions, the distribution D is integrable. Moreover, they claim that the training set as well as the test set belong to a unique leaf of the foliation associated to D. Following them, we call this leaf the *data leaf*.

B. Properties of the data leaf related to adversarial vulnerability

Following [9], we claim that adversarial vulnerability is due to a discrepancy between the distance used in the data space $\mathcal{X}$ to distinguish similar inputs (e.g., any l_p norm) and the distance induced by the local data matrix G_x on $\mathcal{X}$. In this work, we consider only the l_2 norm (i.e., the Euclidean metric) to measure the attack budget. This discrepancy can be further split into two components:

1) A component tangent to the data leaf that can be quantified as the discrepancy between the induced Euclidean metric on the data leaf and the restriction of G_x to the data leaf, which is non-singular.
2) A component orthogonal to the data leaf that corresponds to the extrinsic curvature of the data leaf when $\mathcal{X}$ is equipped with the Euclidean metric. This component was investigated in [16].

C. Existence of the data leaf

For this framework to be useful, the data leaf must exist, i.e., the distribution D must be integrable. The distribution D is integrable into a regular foliation (i.e., a foliation whose leaves are regular immersed submanifolds) if and only if D has constant rank and is involutive [14]. A sufficient condition is given by the following result [16]:

Proposition 1. *If there exists a torsion-free metric-compatible connection ∇ on $T\mathcal{X}$, then the distribution D is integrable.*

Two directions can be followed:

1) What are the conditions on the model for the distribution D to have constant rank? If the distribution has not constant rank, is it integrable as a singular foliation? What are the consequences for the geometry of the data leaf?
2) What are the conditions on the model for the distribution D to be involutive (which is a necessary condition for the distribution to be integrable)?

III. Potential applications

The data leaf framework can help design regularization methods to improve adversarial robustness. It may explain the dependence of the generalization capabilities and robustness of a model to the training set distribution ("robust" vs "non-robust" dataset [7]). Besides adversarial vulnerability, the data leaf may be used as generative model and may explain other phenomenon such as catastrophic forgetting.

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014.

[2] G. Vidot, C. Gabreau, I. Ober, and I. Ober, "Certification of embedded systems based on machine learning: A survey," 2021. [Online]. Available: https://arxiv.org/abs/2106.07221

[3] H. Liang, E. He, Y. Zhao, Z. Jia, and H. Li, "Adversarial Attack and Defense: A Survey," *Electronics*, vol. 11, no. 8, p. 1283, 2022.

[4] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[5] T. Tanay and L. Griffin, "A boundary tilting persepective on the phenomenon of adversarial examples," 2016. [Online]. Available: https://arxiv.org/abs/1608.07690

[6] J. Gilmer, L. Metz, F. Faghri, S. S. Schoenholz, M. Raghu, M. Wattenberg, and I. J. Goodfellow, "Adversarial spheres," in *International Conference on Learning Representations*, 2018.

[7] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," in *Advances in Neural Information Processing Systems*, 2019.

[8] T. Miyato, S.-i. Maeda, M. Koyama, K. Nakae, and S. Ishii, "Distributional smoothing with virtual adversarial training," 2015. [Online]. Available: https://arxiv.org/abs/1507.00677

[9] A. Nayebi and S. Ganguli, "Biologically inspired protection of deep networks from adversarial attacks," 2017. [Online]. Available: https://arxiv.org/abs/1703.09202

[10] C. Zhao, P. T. Fletcher, M. Yu, Y. Peng, G. Zhang, and C. Shen, "The Adversarial Attack and Detection under the Fisher Information Metric," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 5869–5876, 2019.

[11] J. Martin and C. Elster, "Inspecting adversarial examples using the Fisher information," *Neurocomputing*, vol. 382, pp. 80–86, 2020.

[12] Y. Shi, B. Liao, G. Chen, Y. Liu, M.-M. Cheng, and J. Feng, "Understanding adversarial behavior of dnns by disentangling non-robust and robust components in performance metric," 2019. [Online]. Available: https://arxiv.org/abs/1906.02494

[13] C. Shen, Y. Peng, G. Zhang, and J. Fan, "Defending against adversarial attacks by suppressing the largest eigenvalue of fisher information matrix," 2019. [Online]. Available: https://arxiv.org/abs/1909.06137

[14] L. Grementieri and R. Fioresi, "Model-centric data manifold: The data through the eyes of the model," *SIAM Journal on Imaging Sciences*, vol. 15, no. 3, pp. 1140–1156, 2022.

[15] M. Picot, F. Messina, M. Boudiaf, F. Labeau, I. Ben Ayed, and P. Piantanida, "Adversarial robustness via fisher-rao regularization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[16] E. Tron, N. Couellan, and S. Puechmorel, "Canonical foliations of neural networks: application to robustness," 2022. [Online]. Available: https://arxiv.org/abs/2203.00922