Article

ADVERSARIAL ROBUSTNESS WITH PARTIAL ISOMETRY

Loïc Shi-Garrier[1,‡] , Nidhal Carla Bouaynaya[2], Daniel Delahaye3

[1] ENAC, Université de Toulouse, 7, Avenue Edouard Belin, Toulouse, France; loic.shi-garrier@enac.fr
[2] Dept. of Electrical and Computer Engineering, Rowan University, New Jersey, USA ; bouaynaya@rowan.edu
[3] ENAC, Université de Toulouse, 7, Avenue Edouard Belin, Toulouse, France; loic.shi-garrier@enac.fr
* Correspondence: loic.shi-garrier@enac.fr

Abstract: Despite their outstanding performance, deep learning models still lack robustness guarantees, notably in the face of adversarial examples. This major weakness hinders their trustworthiness and prevents the introduction of these learning systems in critical domains where a certain level of robustness must be certified. In this paper, we present an information geometric framework to derive a precise robustness criteria for l_2 white-box attacks in a multi-class classification setting. We endow the output space with the Fisher information metric and derive a criteria on the input-output Jacobian to ensure robustness. We show that model robustness can be achieved by constraining the model to be partially isometric around the training points. The approach is tested on MNIST and CIFAR-10 against adversarial attacks. It is shown to be significantly more robust than defensive distillation and Jacobian regularization for medium-sized perturbations and more robust than adversarial training for large perturbations, while still maintaining desired accuracy.

Keywords: Adversarial robustness, Information geometry, Fisher information metric, Multi-class classification.

1. Introduction

One of the motivations for studying machine learning robustness is the high sensitivity of neural networks to adversarial attacks, i.e., small perturbations in the input data that are able to fool a network. Adversarial attacks have been shown to be both ubiquitous and transferable [1], [2], [3]. Beyond the security threat, adversarial attacks are the evidence of the dramatic lack of robustness in machine learning models [4], [5]. Without robustness, trustworthiness in machine learning is impossible [6].

In this paper, we shed an information geometric perspective to adversarial robustness in machine learning models. We show that robustness can be achieved by encouraging the model to be isometric in the orthogonal space of the kernel of the pullback Fisher metric. We subsequently formulate a regularization defense method for adversarial robustness. We focus on l_2 white-box attacks against multi-class classification tasks; but the approach could be extended to more general settings, e.g., unrestricted attacks and black-box attacks, as well as to other supervised learning tasks. The regularized model is evaluated on MNIST and CIFAR-10 against PGD l_∞ attacks and AutoAttack [7] with l_∞ and l_2 norms. Comparisons with unregularized model, defensive distillation [8], Jacobian regularization [9], and Fisher information regularization [10] show significant improvement in robustness. Moreover, the regularized model is able to ensure robustness for larger perturbations compared to adversarial training.

The remaining of this paper is organized as follows. Section 2 introduces the notations and definitions. Then, a sufficient condition for adversarial robustness at a sample point is derived. Section 3 presents our method to approximate the robustness condition. The method relies on encouraging the model to be isometric in the orthogonal complement of the kernel of the pullback of the FIM. Section 4 presents several experiments to evaluate the proposed method. Section 5 discusses the results in the lights of related work on adversarial

Citation: Shi-Garrier, L., Bouaynaya, N., Delahaye, D. . *Entropy* **2023**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

defense. Finally, section 6 concludes the paper and suggests potential extensions of this work. Section 7 provides the proofs of the results stated in the earlier sections.

2. Notations and definitions

2.1. Notations

Let $d, c \in \mathbb{N}^*$ such that $d \geq c > 1$. Let $m = c - 1$. In the learning framework, d will be the dimension of the input space, while c will be the number of classes.
The range of a matrix M is denoted $\mathrm{rg}(M)$, its rank is denoted $\mathrm{rk}(M)$.
The Euclidean norm (i.e., l_2 norm) is denoted $\| \cdot \|$.
We use the notation $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.
We denote the components of a vector v by $v^i \in \mathbb{R}$ with a superscript. Smooth means C^∞.

2.2. Definitions

Consider a multi-class classification task. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the *input domain*, and let $\mathcal{Y} = \{1, \ldots, c\} \subset \mathbb{N}$ be the set of labels for the classification task. For example, in MNIST, we have $\mathcal{X} = [0,1]^d$ (with $d = 784$) and $c = 10$. We assume that $\mathcal{X}$ is an d-dimensional embedded smooth connected submanifold of $\mathbb{R}^d$.

Definition 1 (Probability simplex)**.** Define the *probability simplex* of dimension m by:

$$\Delta^m = \left\{ \theta \in \mathbb{R}^c : \forall k \in \{1, \ldots, c\}, \theta^k > 0 \text{ and } \sum_{i=1}^{c} \theta^i = 1 \right\}.$$

Δ^m is a smooth submanifold of $\mathbb{R}^c$ of dimension $m = c - 1$. When we write $\theta \in \Delta^m$, we see θ as having m coordinates: $\theta = (\theta^1, \ldots, \theta^m)$. Then, we define $\theta^c = 1 - \sum_{i=1}^m \theta^i$.

A machine learning model (e.g., a neural network) is often seen as assigning a label $y \in \mathcal{Y}$ to a given input $x \in \mathcal{X}$. Instead, in this work, we see a model as assigning the *parameters* of a random variable Y to a given input $x \in \mathcal{X}$. The random variable Y has a probability density function p_θ belonging to the *family of c-dimensional categorical distributions* $\mathcal{S} = \{p_\theta : \theta \in \Delta^m\}$.

$\mathcal{S}$ can be endowed with a differentiable structure by using $p_\theta \in \mathcal{S} \mapsto (\theta^1, \ldots, \theta^m) \in \mathbb{R}^m$ as a global coordinate system. Hence, $\mathcal{S}$ becomes a smooth manifold of dimension m (more details on this construction can be found in [11], Chapter 2). We can identify p_θ with $(\theta^1, \ldots, \theta^m)$.

We see any machine learning model as a smooth map $f : \mathcal{X} \to \Delta^m$ that assigns to an input $x \in \mathcal{X}$, the parameters $\theta = f(x) \in \Delta^m$ of a c-dimensional categorical distribution $p_\theta \in \mathcal{S}$. In practice, a neural network produces a vector of logits $s(x)$. Then, these logits are transformed into the parameters θ with the softmax function: $\theta = \mathrm{softmax}(s(x))$.

In order to study the sensitivity of the predicted $f(x) \in \Delta^m$ with respect to the input $x \in \mathcal{X}$, we need to be able to measure distances both in $\mathcal{X}$ and in Δ^m. In order to measure distances on smooth manifolds, we need to equip each manifold with a Riemannian metric.

First, we consider Δ^m. As described above, we see Δ^m as the family of categorical distributions. A natural Riemannian metric for Δ^m (i.e., a metric that reflects the statistical properties of Δ^m) is the *Fisher information metric* (FIM).

Definition 2 (Fisher information metric)**.** For each $\theta \in \Delta^m$, the *Fisher information metric* (FIM) g defines a *symmetric positive-definite bilinear form* g_θ over the tangent space $T_\theta \Delta^m$. In the *standard coordinates* of $\mathbb{R}^c$, we have, for all $\theta \in \Delta^m$ and for all *tangent vectors* $v, w \in T_\theta \Delta^m$:

$$g_\theta(v, w) = v^T G_\theta w,$$

where G_θ is the *Fisher information matrix* for parameter $\theta \in \Delta^m$ defined by:

$$G_{\theta,ij} = \frac{\delta_{ij}}{\theta^i} + \frac{1}{\theta^c}. \tag{1}$$

For any $\theta \in \Delta^m$, the matrix G_θ is *symmetric positive-definite and non-singular* (Proposition 1.6.2 in [12]). The FIM induces a distance on Δ^m called the *Fisher-Rao distance* denoted $d(\theta_1, \theta_2)$ for any $\theta_1, \theta_2 \in \Delta^m$.

The FIM has two remarkable property. First, it is the "infinitesimal distance" of the the *relative entropy* (Theorem 4.4.5 in [12]), which is the loss function used to train a multi-class classification model. The other remarkable property of the FIM is Chentsov's theorem [13] claiming that the FIM is the *unique* Riemannian metric on Δ^m that is invariant under sufficient statistics (up to a multiplicative constant).

Now, we consider $\mathcal{X}$. Since we are studying adversarial robustness, we need a metric that formalizes the idea that two close data points must be "indistinguishable" from a human perspective (or any other relevant perspective). A natural choice is the *Euclidean metric* induced from $\mathbb{R}^d$ on $\mathcal{X}$.

Definition 3 (Euclidean metric)**.** We consider the *Euclidean space* $\mathbb{R}^d$ endowed with the *Euclidean metric* $\bar{g}$. It is defined in the standard coordinates of $\mathbb{R}^d$ for all $x \in \mathbb{R}^d$ and for all tangent vectors $v, w \in T_x\mathbb{R}^d$ by:

$$\bar{g}_x(v, w) = v^T w,$$

thus its matrix is the identity matrix of dimension d denoted I_d. The Euclidean metric induces a distance on $\mathbb{R}^d$ that we will denote with the l_2-norm: $\|x_1 - x_2\|$ for any $x_1, x_2 \in \mathbb{R}^d$.

From now on, we fix:

- a smooth map $f : (\mathcal{X}, \bar{g}) \to (\Delta^m, g)$. We denote by f^i the i-th component of f in the standard coordinates of $\mathbb{R}^c$.
- a point $x \in \mathcal{X}$.
- a positive real number $\epsilon > 0$.

Definition 4 (Euclidean ball)**.** Define the Euclidean open ball centered at x with radius ϵ by:

$$\bar{b}(x, \epsilon) = \left\{ z \in \mathbb{R}^d : \|z - x\| < \epsilon \right\}.$$

Definition 5. Define the set (Figure 1):

$$\mathcal{A} = \left\{ \theta \in \Delta^m : \arg\max_i \theta^i = \arg\max_i f^i(x) \right\}.$$

For simplicity, assume that $f(x)$ is not on the "boundary" of $\mathcal{A}$, such that $\arg\max_i f^i(x)$ is well defined.

Definition 6 (Geodesic ball of the FIM)**.** Let $\delta > 0$ be the Fisher-Rao distance between $f(x)$ and $\Delta^m \setminus \mathcal{A}$ (Figure 2).
Define the geodesic ball centered at $f(x) \in \Delta^m$ with radius δ by:

$$b(f(x), \delta) = \{\theta \in \Delta^m : d(f(x), \theta) \leq \delta\}.$$

In section 3.3, we propose a efficient approximation of δ.

Definition 7 (Pullback metric)**.** On $\mathcal{X}$, define the *pullback metric* $\tilde{g}$ of g by f. In the standard coordinates of $\mathbb{R}^d$, $\tilde{g}$ is defined for all tangent vectors $v, w \in T_x\mathcal{X}$ by:

$$\tilde{g}_x(v, w) = v^T J_x^T G_{f(x)} J_x w,$$

where J_x is the Jacobian matrix of f at x (in the standard coordinates of $\mathbb{R}^d$ and $\mathbb{R}^c$). Define the matrix of $\tilde{g}_x$ in the standard coordinates of $\mathbb{R}^d$ by:

$$\widetilde{G}_x = J_x^T G_{f(x)} J_x. \tag{2}$$

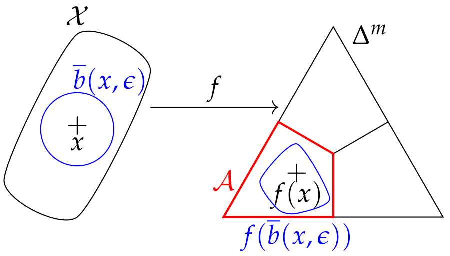

Figure 1. ϵ-robustness at x is enforced if and only if $f(\bar{b}(x,\epsilon)) \subseteq \mathcal{A}$.

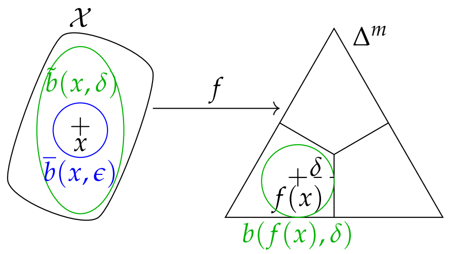

Figure 2. ϵ-robustness at x is enforced if $\bar{b}(x,\epsilon) \subseteq \tilde{b}(x,\delta)$.

Definition 8 (Geodesic ball of the pullback metric)**.** Let $\tilde{d}$ be the distance induced by the pullback metric $\tilde{g}$ on $\mathbb{R}^d$. We can define the geodesic ball centered at x with radius δ by:

$$\tilde{b}(x,\delta) = \left\{ z \in \mathbb{R}^d : \tilde{d}(x,z) \le \delta \right\}.$$

2.3. Robustness condition

Definition 9 (Robustness)**.** We say that f is ϵ-robust at x if:

$$\forall z \in \mathbb{R}^d, \|z - x\| < \epsilon \Rightarrow f(z) \in \mathcal{A}. \tag{3}$$

Equivalently, we can write (Figure 1):

$$f(\bar{b}(x,\epsilon)) \subseteq \mathcal{A}. \tag{4}$$

Proposition 10 (Sufficient condition for robustness)**.** *If $\bar{b}(x,\epsilon) \subseteq \tilde{b}(x,\delta)$, then f is ϵ-robust at x (Figure 2).*

Our goal is to start from Proposition 10 and make several assumptions in order to derive a condition that can be efficiently implemented.

Working with geodesic balls $\bar{b}(x,\epsilon)$ and $\tilde{b}(x,\delta)$ is intractable, so our first assumption consists in using an "infinitesimal" condition by restating Proposition 10 in the tangent space $T_x\mathcal{X}$ instead of working directly on $\mathcal{X}$.

Definition 11. In $T_x\mathcal{X}$, define the Euclidean ball of radius ϵ by:

$$\overline{\mathcal{B}}_x(0,\epsilon) = \left\{ v \in T_x\mathcal{X} : \bar{g}_x(v,v) = v^T v \le \epsilon^2 \right\}.$$

Definition 12. In $T_x\mathcal{X}$, define the $\tilde{g}_x$-ball of radius δ by:

$$\widetilde{\mathcal{B}}_x(0,\delta) = \left\{ v \in T_x\mathcal{X} : \tilde{g}_x(v,v) = v^T \widetilde{G}_x v \le \delta^2 \right\}.$$

Assumption 1. We replace Proposition 10 by:

$$\overline{\mathcal{B}}_x(0,\epsilon) \subseteq \widetilde{\mathcal{B}}_x(0,\delta). \tag{5}$$

For small enough δ, Eq. (5) implies ϵ-robustness at x. However, contrary to Proposition 10, Eq. (5) does not offer any guarantee on the ϵ-robustness at x for arbitrary δ.

Proposition 13. *Eq. (5) is equivalent to:*

$$\forall v \in T_x\mathcal{X}, \quad \tilde{g}_x(v,v) \le \frac{\delta^2}{\epsilon^2} \bar{g}_x(v,v). \tag{6}$$

Since $m < d$, the Jacobian matrix J_x has rank smaller or equal to m. Thus, since $G_{f(x)}$ has full rank, $\widetilde{G}_x = J_x^T G_{f(x)} J_x$ has rank at most m (when J_x has rank m).

Assumption 2. The Jacobian matrix J_x has full rank equal to m.

3. Derivation of the regularization method

3.1. The partial isometry condition

In order to simplify the notations, we replace:

- J_x by J which is a full-rank $m \times d$ real matrix.
- $G_{f(x)}$ by G which is a $m \times m$ symmetric positive definite real matrix.
- $\widetilde{G}_x$ by $\widetilde{G}$ which is a $d \times d$ symmetric positive semidefinite real matrix.

We define $D = (\ker(\widetilde{G}))^{\perp}$. We will use the two following facts.

Fact 14.

$$D = \mathrm{rg}(J^T) = (\ker(J))^{\perp} = \left(\ker(J^T G J)\right)^{\perp}$$

Fact 15. $J^T G J$ is symmetric positive semidefinite. Thus, by the spectral theorem, the eigenvectors associated to its nonzero eigenvalues are all in $D = \mathrm{rg}(J^T)$.
In particular, since $\mathrm{rk}(J) = m$, there exists an orthonormal basis of $T_x\mathcal{X}$, denoted $\mathcal{B} = (e_1, \dots, e_m, e_{m+1}, \dots, e_d)$, such that each e_i is an eigenvector of $J^T G J$ and such that $(e_1, \dots, e_m)$ is a basis of $D = \mathrm{rg}(J^T)$ and $(e_{m+1}, \dots, e_d)$ is a basis of $\ker(J)$.

The set $D = \mathrm{rg}(J^T)$ is a m-dimensional subspace of $T_x\mathcal{X}$. $\tilde{g}_x$ does not define an inner product[1] on $T_x\mathcal{X}$ because $\widetilde{G}$ has a nontrivial kernel of dimension $d - m$. However, when restricted to D, $\tilde{g}_x|_D$ defines an inner product.

Definition 16. We define the restriction of $\widetilde{\mathcal{B}}_x(0, \delta)$ to D:

$$\widetilde{\mathcal{B}}_D(0, \delta) = \left\{ v \in D : v^T \widetilde{G} v \leq \delta \right\}$$

Definition 17. We define the restriction of $\overline{\mathcal{B}}_x(0, \epsilon)$ to D:

$$\overline{\mathcal{B}}_D(0, \epsilon) = \left\{ v \in D : v^T v \leq \epsilon^2 \right\}.$$

Assumption 3. We replace Eq. (5) with:

$$\overline{\mathcal{B}}_D(0, \epsilon) = \widetilde{\mathcal{B}}_D(0, \delta). \tag{7}$$

Eq. (7) is the limit case of Eq. (5), in the sense that if Eq. (7) holds, then $\widetilde{\mathcal{B}}_x(0, \delta)$ is the smallest possible $\tilde{g}_x$-ball (for the inclusion) such that Eq. (5) holds.

Proposition 18. *Eq. (7) is equivalent to:*

$$\forall v \in D, \quad \tilde{g}_x(v, v) = \frac{\delta^2}{\epsilon^2} \bar{g}_x(v, v). \tag{8}$$

We can rewrite Eq. (8) in a matrix form:

$$\forall v \in D, \quad v^T \widetilde{G} v = \frac{\delta^2}{\epsilon^2} v^T v. \tag{9}$$

In section 3.2, we will show how to exploit the properties of the FIM to derive a closed-form expression for a matrix $P \in \mathrm{GL}_m(\mathbb{R})$ such that $G = P^T P$. For now, we assume that we can easily access such a P and we are looking for a condition on P and J that is equivalent with Eq. (9).

[1] In particular, the set $\widetilde{\mathcal{B}}_x(0, \delta)$ is not bounded, i.e., it is a cylinder rather than a ball.

Proposition 19. *The following statements are equivalent:*

$$(i) \qquad \forall u \in D, \quad u^T J^T G J u = \frac{\delta^2}{\epsilon^2} u^T u,$$
$$(ii) \qquad PJJ^T P^T = \frac{\delta^2}{\epsilon^2} I_m,$$

where I_m is the identity matrix of dimension $m \times m$.

Proposition 19 constrains the Jacobian matrix PJ to be a *semi-orthogonal matrix* (multiplied by a homothety matrix). With this condition, f becomes a *partial isometry*, at least in the neighborhood of the training points.
Finally, we can define a regularization term:

$$\alpha(x, \epsilon, f) = \frac{1}{m^2} \left|\left|\left| PJJ^T P^T - \frac{\delta^2}{\epsilon^2} I_m \right|\right|\right|, \tag{10}$$

where $|||\cdot|||$ is any matrix norm, such as the Frobenius norm or the spectral norm. We use the Frobenius norm in the experiments of section 4. To compute $\alpha(x, \epsilon, f)$, we only need to compute the Jacobian matrix J which can be efficiently achieved with backpropagation. The loss function is:

$$L(y, x, \epsilon, f) = l(y, f(x)) + \lambda\, \alpha(x, \epsilon, f), \tag{11}$$

where l is the cross-entropy loss and $\lambda > 0$.

3.2. Coordinate change

In this subsection, we show how to compute the matrix P that was introduced in Proposition 19. To this end, we isometrically embed Δ^m into the Euclidean space $\mathbb{R}^c$ using the following inclusion map:

$$\begin{aligned} \mu : \Delta^m &\longrightarrow \mathbb{R}^c \\ \left(\theta^1, \dots, \theta^m\right) &\longmapsto 2\left(\sqrt{\theta^1}, \dots, \sqrt{\theta^m}, \sqrt{1 - \sum_{i=1}^m \theta^i}\right) \end{aligned}$$

We can easily see that μ is an embedding. If $\mathcal{S}^m(2)$ is the sphere of radius 2 centered at the origin in $\mathbb{R}^c$, then $\mu(\Delta^m)$ is the subset of $\mathcal{S}^m(2)$ where all coordinates are strictly positive (using the standard coordinates of $\mathbb{R}^c$).

Proposition 20. *Let g be the Fisher information metric on Δ^m (Definition 2), and $\overline{g}$ be the Euclidean metric on $\mathbb{R}^c$. Then μ is an isometric embedding of (Δ^m, g) into $(\mathbb{R}^c, \overline{g})$.*

Now, we use the stereographic projection to embed Δ^m into $\mathbb{R}^m$:

$$\begin{aligned} \tau : \mu(\Delta^m) &\longrightarrow \mathbb{R}^m \\ (\mu^1, \dots, \mu^m, \mu^c) &\longmapsto 2\left(\frac{\mu^1}{2 - \mu^c}, \dots, \frac{\mu^m}{2 - \mu^c}\right), \end{aligned}$$

with $\mu^c = 2\sqrt{1 - \sum_{i=1}^m \theta^i}$.

Proposition 21. *In the coordinates τ, the FIM is:*

$$G_{\tau,ij} = \frac{4}{\left(1 + \|\tau/2\|^2\right)^2} \delta_{ij}. \tag{12}$$

Let $\widetilde{J}$ be the Jacobian matrix of $\tau \circ \mu : \Delta^m \to \mathbb{R}^m$ at $f(x)$. Then we have:

$$G = \widetilde{J}^T G_\tau \widetilde{J} = \frac{4}{\left(1 + \|\tau/2\|^2\right)^2} \widetilde{J}^T \widetilde{J}. \tag{13}$$

Thus, we can choose:

$$P = \frac{2}{1+\|\tau/2\|^2}\widetilde{J}. \tag{14}$$

Write $f(x) = \theta = (\theta^1, \dots, \theta^m)$ and $\theta_c = 1 - \sum_{i=1}^{m} \theta^i$. For simplicity, write $\tau^i(\theta) = \tau^i(\mu(\theta)) = 2\sqrt{\theta^i}/(1-\sqrt{\theta^c})$ for $i = 1, \dots, m$. More explicitly, we have:

Proposition 22. *For $i, j = 1, \dots, m$:*

$$P_{ij} = \frac{\delta_{ij}}{\sqrt{\theta^i}} - \frac{\tau^i(\theta)}{2\sqrt{\theta^c}}. \tag{15}$$

3.3. The Fisher-Rao distance

In this subsection, we derive a simple upper-bound for δ (i.e., the Fisher-Rao distance between $f(x)$ and $\Delta^m \setminus \mathcal{A}$). In Proposition 20, we have shown that the probability simplex Δ^m endowed with the FIM can be isometrically embedded into the m-sphere of radius 2. Thus, the angle β between two distributions of coordinates θ_1 and θ_2 in Δ^m with $\mu_1 = \mu(\theta_1)$ and $\mu_2 = \mu(\theta_2)$ is:

$$\cos(\beta) = \frac{1}{4}\sum_{i=1}^{c} \mu_1^i \mu_2^i = \sum_{i=1}^{c} \sqrt{\theta_1^i \theta_2^i}.$$

The Riemannian distance between these two points is the arc length on the sphere:

$$d(\theta_1, \theta_2) = 2 \arccos \sum_{i=1}^{c} \sqrt{\theta_1^i \theta_2^i}.$$

In the regularization term defined in Eq. (10), we replace δ by the following upper bound:

$$\delta = d(f(x), \Delta^m \setminus \mathcal{A}) \leq d(f(x), O),$$

where $O = \frac{1}{c}(1, \dots, 1)$ is the center of the simplex Δ^m. Thus:

$$\delta \leq 2 \arccos \sum_{i=1}^{c} \sqrt{\frac{f(x)^i}{c}}. \tag{16}$$

4. Experiments

The regularization method introduced in Section 3 is evaluated on MNIST and CIFAR-10 datasets.

4.1. Experiments on MNIST dataset

4.1.1. Experimental Setup

For the MNIST dataset[2], we implemented a LeNet model with two convolutional layers of 32 and 64 channels respectively, followed by one hidden layer with 128 neurons. We train three models: one regularized model, one baseline unregularized model, and one model trained with adversarial training. All three models are trained with Adam optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) for 30 epochs, with a batch size of 64, and a learning rate of 10^{-3}. For the regularization term, we use a budget of $\epsilon = 5.6$, which is chosen to contain the l_∞ ball of radius 0.2. The adversarial training is conducted with 10 iterations of PGD with a budget $\epsilon_{adv} = 0.2$ using l_∞ norm. We found that $\lambda = 10^{-6}$ yields the best performance in terms of robustness-accuracy tradeoff; this value is small because we did not attempt to normalize the regularization term.

The models are trained on the 60,000 images of MNIST's training set, then tested on the 10,000 images of the test set. The baseline model achieves an accuracy of 98.9% (9893/10000), the regularized model achieves an accuracy of 94.0% (9403/10000), and the

[2] Code available here: https://github.com/lshigarrier/geometric_robustness.git

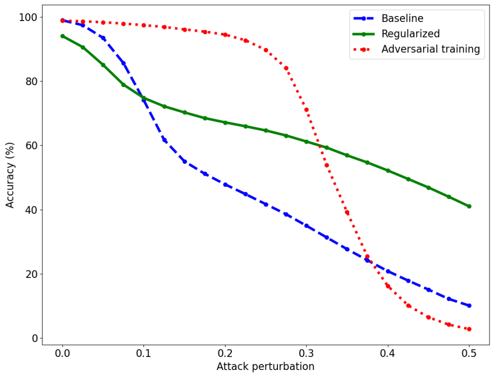

Figure 3. Accuracy of the baseline (dashed, blue), regularized (solid, green), and adversarially trained (dotted, red) models for various attack perturbations on the MNIST dataset. The perturbations are obtained with PGD using l_∞ norm.

Defense	BASE	ISO	DIST	JAC	FIR	AT
Clean	99.01	96.51	98.81	98.95	98.84	98.98
AA-L_2 (0.15)	35.70	43.38	35.35	38.74	1.68	73.34
AA-L_∞ (1.5)	10.38	22.15	9.63	13.30	0.03	95.43

Table 1. Clean and robust accuracy on MNIST against AA, averaged over 10 runs. The number in parentheses is the attack strength.

adversarially trained model achieves an accuracy of 98.8% (9883/10000). Although the current implementation of the regularized model is almost 6 times slower to train than the baseline model, it may be possible to accelerate the training using, for example, the technique proposed by Shafahi *et al.* [14], or using another method to approximate the spectral norm of $\widetilde{J}$. Even without relying on these acceleration techniques, the regularized model is still faster to train than the adversarially trained model.

4.1.2. Robustness to Adversarial Attacks

To measure the adversarial robustness of the models, we used the PGD attack with the l_∞ norm, 40 iterations, and a step size of 0.01. The l_∞ norm yields the hardest possible attack for our method, and corresponds more to the human notion of “indistinguishable images" than the l_2 norm. The attacks are performed on the test set, and only on images that were correctly classified by each model. The results are reported in Fig. 3. The regularized model has a slightly lower accuracy than the baseline model for small perturbations, but the baseline model suffers a drop in accuracy above attack level $\epsilon = 0.1$. Adversarial training achieves high accuracy for small to medium-sized perturbations but the accuracy decreases sharply above $\epsilon = 0.3$. The regularized model remains robust even for large perturbations. The baseline model reaches 50% accuracy at $\epsilon = 0.2$ and the adversarially trained model at $\epsilon = 0.325$, while the regularized model reaches 50% accuracy at $\epsilon = 0.4$.

Table 1 provides more results against AutoAttack (AA) [7], which was designed to offer a more reliable evaluation of adversarial robustness. For fair comparison, and in addition to a baseline model (BASE), we compare the partial isometry defense with several other computationally efficient: distillation (DIST) [8], Jacobian regularization (JAC) [9], which also relies on the Jacobian matrix of the network, and Fisher information regularization (FIR) [10], which also leverages information geometry. We also consider an adversarially trained (AT) model using PGD. We see that the DIST method, which is the best method outside AT against FGSM (not reported), drops below BASE against AA. This is also the case of FIR. ISO is the best defense not relying on adversarial training. In future work, ISO may be combined with AT to further boost performance.

Defense	BASE	ISO	DIST	JAC	FIR	AT
Clean	92.93	76.86	84.96	86.17	89.98	80.78
PGD (4/255)	2.49	40.17	7.54	8.56	9.74	68.82
PGD (8/255)	0.47	39.68	3.35	3.66	4.05	66.61

Table 2. Clean and robust accuracy on CIFAR-10 against PGD. The number in parentheses is the attack strength.

4.2. *Experiments on CIFAR-10 dataset*

We consider[3] a DenseNet121 model fine-tuned on CIFAR-10 using pre-trained weights for ImageNet. As for the MNIST experiments, we compare the partial isometry defense with distillation (DIST), Jacobian regularization (JAC), and Fisher information regularization (FIR). Here, the adversarial training (AT) relies on FGSM attack [15]. All defenses are compared against PGD for various attack strengths. The results are reported in Table 2. The defenses are evaluated in a "gray-box" setting where the adversary can access the architecture and the data but not the weights. More precisely, the adversarial examples are crafted from the test set of CIFAR-10 using another unregularized DenseNet121 model. AT is the more robust method, but ISO achieves a robust accuracy 30% higher than the next best analogous method (FIR).

One of our goals is to provide alternatives to adversarial training (AT). Besides high computational cost, AT suffers from several limitations: it only robustifies against the chosen attack at the chosen budget, and does not offer robustness guarantees. For example, under Gaussian noise, AT accuracy decreases *faster* than baseline accuracy (i.e., no defense). Achieving high robustness accuracy against specific attacks on specific benchmark is insufficient and misleading to measure the true robustness of the evaluated model. Our method offers a new point of view that can be extended to certified defense methods in future works.

5. Discussion and related work

In 2019, Zhao *et al.* [16] proposed to use the Fisher information metric in the setting of adversarial attacks. They used the eigenvector associated to the largest eigenvalue of the pullback of the FIM as an attack direction. Following their work, Shen *et al.* [10] suggested a defense mechanism by suppressing the largest eigenvalue of the FIM. They upper-bounded the largest eigenvalue by the trace of the FIM. As in our work, they added a regularization term to encourage the model to have smaller eigenvalues. Moreover, they showed that their approach is equivalent to label smoothing [17]. In our framework, their method consists in expanding the geodesic ball $\tilde{b}(x, \delta)$ as much as possible. However, their approach does not guarantee that the constraint imposed on the model will not harm the accuracy more than necessary. In our framework, the matrix PJ (compared with δ/ϵ) informs the model on the precise restriction that must be imposed to achieve adversarial robustness in the l_2 ball of radius ϵ.

Cisse *et al.* [18] introduced an other adversarial defense called *Parseval networks*. To achieve adversarial robustness, the authors aim at controlling the Lipschitz constant of each layer of the model to be close to unity. This is achieved by constraining the weight matrix of each layer to be a *Parseval tight frame*, which is synonymous with semi-orthogonal matrix. Since the Jacobian matrix of the entire model with respect to the input is almost the product of the weight matrices, the Parseval network defense is similar to our proposed defense, albeit with completely different rationales. This suggests that geometric reasoning could successfully supplement the line of work on Lipschitz constants of neural networks, such as in [19].

Following another line of work, Hoffman *et al.* [9] advanced a Jacobian regularization to improve adversarial robustness. Their regularization consists in using the Frobenius

[3] Code available here: https://github.com/lshigarrier/iso_defense.git

norm of the input-output Jacobian matrix. To avoid computing the true Frobenius norm, they relied on random projections, which are shown to be both efficient and accurate. This method is similar to the method of Shen *et al.* [10] in the sense that it will also increase the radius of the geodesic ball. However, the Jacobian regularization does not take into account the geometry of the output space (i.e., the Fisher information metric) and assumes that the probability simplex Δ^m is Euclidean.

Although this study focuses on l_2 norm robustness, it must be pointed out that there are other “distinguishability" measures that can be used to study adversarial robustness, including all other l_p norms. In particular, the l_∞ norm is often considered to be the most natural choice when working with images. However, the l_∞ norm is not induced by any inner product, and hence, there is no Riemannian metric that induces the l_∞ norm. However, given an l_∞ budget ϵ_∞, we can choose an l_2 budget $\epsilon_2 = \sqrt{n}\epsilon_\infty$ such that any attack in the ϵ_∞ budget will also respect the ϵ_2 budget. When working on images, other dissimilarity measures are: rotations, deformations, or color changes of the original image. Contrary to the l_2 or l_∞, these measures are not based on a pixel-based coordinate system. However, it is possible to define *unrestricted attacks* based on these spatial dissimilarities, for example in [20].

In this work, we derived the partial isometry regularization for a classification task. The method can be extended to regression tasks by considering the family of multivariate normal distributions as the output space. On the probability simplex Δ^m, the FIM is a metric with constant positive curvature, while it has constant negative curvature on the manifold of multivariate normal distributions [21].

Finally, the precise quantification of the robustness condition presented in Eq. (3) and Proposition 19 paves the way to the development of a certified defense [22] in this framework. By strongly enforcing Proposition 19 on a chosen proportion of the training set, it may be possible to maximize the accuracy under the constraint of a chosen robustness level, which offers another solution to the robustness-accuracy trade-off [23], [24]. Certifiable defenses are a require step for the deployment of deep learning models in critical domains and missions, such as civil aviation, security, defense and healthcare, where a certification may be required to ensure a sufficient level of trustworthiness.

6. Conclusion and future work

In this paper, we introduced an information geometric approach to the problem of adversarial robustness in machine learning models. The proposed defense consists of enforcing a partial isometry between the input space endowed with the Euclidean metric and the probability simplex endowed with the Fisher information metric. We subsequently derived a regularization term to achieve robustness during training. The proposed strategy is tested on the MNIST and CIFAR-10 datasets, and shows considerable increase in robustness without harming the accuracy. Future works will evaluate the method on other benchmarks and real-world datasets. Several attack methods will also be considered in addition to PGD and AutoAttack. Although this work focuses on l_2 norm robustness, future work would consider other “distinguishability" measures.

Our work extends a recent, promising but under-studied framework for adversarial robustness based on information geometric tools. The FIM has already been harnessed to develop attacks [16] and defenses [10,25] but a precise robustness analysis is yet to be proposed. Our work is a step towards the development of such analysis which may yield certified guarantees relying on these geometric tools. The study of adversarial robustness, which is non-local by definition and contrary to accuracy, should benefit greatly from a geometrical vision. However, the current literature on adversarial robustness is mainly concerned with the FIM and its spectrum (which are very local objects) without unfolding the full arsenal developed in information geometry. In our work, we demonstrate the usefulness of such approach by developing a preliminary robustification method. Model robustification is a hard, unsolved yet vital problem to ensure the trustworthiness of deep learning tools in safety-critical applications. Our framework could be extended and applied to existing certification strategies, such as Lipschitz-based [26] or randomized smoothing [22] where statistical models naturally appear.

7. Proofs

Proof of Proposition 13. (6) ⇒ (5]). Assume (6). Let $v \in \overline{\mathcal{B}}_x(0,\epsilon)$. Thus $\overline{g}_x(v,v) \leq \epsilon^2$. We have:

$$\tilde{g}_x(v,v) \leq \frac{\delta^2}{\epsilon^2}\overline{g}_x(v,v) \leq \frac{\delta^2}{\epsilon^2}\epsilon^2 = \delta^2.$$

Thus $v \in \widetilde{\mathcal{B}}_x(0,\delta)$.
(5) ⇒ (6]). Assume (5). Let $v \in T_x\mathcal{X}$. Define $w = \epsilon\, v / \sqrt{\overline{g}_x(v,v)}$. Then $\overline{g}_x(w,w) = \epsilon^2$. Thus, $w \in \overline{\mathcal{B}}_x(0,\epsilon)$. Hence, $w \in \widetilde{\mathcal{B}}_x(0,\delta)$. Thus, $\tilde{g}_x(w,w) < \delta^2$. Finally, we have:

$$\tilde{g}_x(w,w) = \frac{\epsilon^2}{\overline{g}_x(v,v)}\tilde{g}_x(v,v) < \delta^2.$$

We obtain Equation (6) by multiplying by $\overline{g}_x(v,v)/\epsilon^2$. □

Proof of Proposition 20. We need to show that $\mu^*\overline{g} = g$. Using the coordinates θ on Δ^m (Definition 1) and the standard coordinates on $\mathbb{R}^c$, and writing $f(x) = \theta_0 = (\theta_0^1, \ldots, \theta_0^m)$ we have:

$$\begin{aligned} G_{ij} &= G_{\theta_0,ij}, \\ &= \sum_{\alpha=1}^{c}\sum_{\beta=1}^{c} \frac{\partial\mu^\alpha(\theta_0)}{\partial\theta^i}\frac{\partial\mu^\beta(\theta_0)}{\partial\theta^j}\delta_{\alpha\beta}, \\ &= \sum_{\alpha=1}^{c} \frac{\partial\mu^\alpha(\theta_0)}{\partial\theta^i}\frac{\partial\mu^\alpha(\theta_0)}{\partial\theta^j}. \end{aligned}$$

For $i = 1, \ldots, m$ and $\alpha = 1, \ldots, m$ we have:

$$\frac{\partial\mu^\alpha(\theta_0)}{\partial\theta^i} = \frac{\delta_{i\alpha}}{\sqrt{\theta_0^i}},$$

and for $\alpha = c$:

$$\frac{\partial\mu^c(\theta_0)}{\partial\theta^i} = -\frac{1}{\sqrt{\theta_0^c}},$$

with $\theta_0^c = \sqrt{1 - \sum_{i=1}^m \theta_0^i}$. Thus:

$$G_{\theta,ij} = \frac{\delta_{ij}}{\theta_0^i} + \frac{1}{\theta_0^c},$$

which is the FIM as defined in Definition 2. □

Proof of Proposition 21. For $i = 1, \ldots, m$, the inverse transformation of $\tau(\mu)$ is:

$$\mu^i(\tau) = \frac{2\tau^i}{1 + \|\tau/2\|^2}, \tag{17}$$

and:

$$\mu^c(\tau) = 2\frac{\|\tau/2\|^2 - 1}{\|\tau/2\|^2 + 1}. \tag{18}$$

The proofs of Eqs. (17) and (18) are provided below.
Moreover, according to Proposition 20, the FIM in the coordinates $(\mu^1, \dots, \mu^m)$ is the metric induced on $\mu(\Delta^m)$ by the identity matrix (i.e., the Euclidean metric) of $\mathbb{R}^c$. Hence, we have:

$$\begin{aligned} G_{\tau,ij} &= \sum_{\alpha=1}^{c} \sum_{\beta=1}^{c} \frac{\partial \mu^\alpha(\tau)}{\partial \tau^i} \frac{\partial \mu^\beta(\tau)}{\partial \tau^j} \delta_{\alpha\beta}, \\ &= \sum_{\alpha=1}^{c} \frac{\partial \mu^\alpha(\tau)}{\partial \tau^i} \frac{\partial \mu^\alpha(\tau)}{\partial \tau^j}. \end{aligned}$$

For $i = 1, \dots, m$ and $\alpha = 1, \dots, m$ we have:

$$\frac{\partial \mu^\alpha(\tau)}{\partial \tau^i} = \frac{2}{1 + \|\tau/2\|^2} \left(\delta_{i\alpha} - \frac{\tau^\alpha \tau^i}{2(1 + \|\tau/2\|^2)} \right),$$

and for $\alpha = c$:

$$\frac{\partial \mu^c(\tau)}{\partial \tau^i} = \frac{2\tau^i}{(1 + \|\tau/2\|^2)^2},$$

Thus:

$$\begin{aligned} G_{\tau,ij} =& \frac{4}{(1+\|\tau/2\|^2)^2} \left(\sum_{\alpha=1}^{m} \left\{ \delta_{i\alpha}\delta_{j\alpha} - \frac{\delta_{i\alpha}\tau^j\tau^\alpha}{2(1+\|\tau/2\|^2)} \right. \right. \\ & \left. \left. - \frac{\delta_{j\alpha}\tau^i\tau^\alpha}{2(1+\|\tau/2\|^2)} + \frac{\tau^i\tau^j(\tau^\alpha)^2}{4(1+\|\tau/2\|^2)^2} \right\} + \frac{\tau^i\tau^j}{(1+\|\tau/2\|^2)^2} \right), \\ =& \frac{4}{(1+\|\tau/2\|^2)^2} \left(\delta_{ij} - \frac{\tau^i\tau^j}{1+\|\tau/2\|^2} + \frac{\tau^i\tau^j\|\tau/2\|^2}{(1+\|\tau/2\|^2)^2} \right. \\ & \left. + \frac{\tau^i\tau^j}{(1+\|\tau/2\|^2)^2} \right), \\ =& \frac{4}{(1+\|\tau/2\|^2)^2} \left(\delta_{ij} - \frac{\tau^i\tau^j}{1+\|\tau/2\|^2} + \frac{\tau^i\tau^j}{1+\|\tau/2\|^2} \right), \\ =& \frac{4}{(1+\|\tau/2\|^2)^2} \delta_{ij}. \end{aligned}$$

□

Proof of Eqs. (17) and (18). We have $\tau^i(\mu) = \lambda\mu^i$ with $\lambda = 2/(2 - \mu^c)$. Let us express μ^c as a function of τ. We have:

$$\|\tau\|^2 = \sum_{i=1}^{m} (\tau^i)^2 = \lambda^2 \|\mu\|^2.$$

Since μ belongs to the sphere of radius 2, we have $\|\mu\|^2 + (\mu^c)^2 = 4$. Thus:

$$\|\tau\|^2 = \lambda^2 \left(4 - (\mu^c)^2\right) = 4\frac{4 - (\mu^c)^2}{(2 - \mu^c)^2} = 4\frac{2 + \mu^c}{2 - \mu^c}.$$

Isolating μ^c, we get:

$$\mu^c(\tau) = \frac{2\|\tau\|^2 - 8}{\|\tau\|^2 + 4} = 2\frac{\|\tau/2\|^2 - 1}{\|\tau/2\|^2 + 1}. \tag{19}$$

Now, we can replace μ^c into the expression of λ. We obtain $\lambda = \left(1 + \|\tau/2\|^2\right)/2$, and thus:

$$\mu^i(\tau) = \frac{\tau^i}{\lambda} = \frac{2\tau^i}{1 + \|\tau/2\|^2} \tag{20}$$

□

Proof of Proposition 22. We have:

$$\tau^i(\theta) = 2\sqrt{\theta^i}/\left(1-\sqrt{\theta^c}\right).$$

Thus:

$$\begin{aligned}\left\|\frac{\tau(\theta)}{2}\right\|^2 &= \sum_{i=1}^{m}\frac{\tau^i(\theta)^2}{4},\\ &= \frac{\sum_{i=1}^{m}\theta^i}{\left(1-\sqrt{\theta^c}\right)^2},\\ &= \frac{1-\theta^c}{\left(1-\sqrt{\theta^c}\right)^2},\\ &= \frac{1+\sqrt{\theta^c}}{1-\sqrt{\theta^c}}.\end{aligned}$$

Hence, for any $i = 1,\ldots,m$:

$$\frac{2}{1+\|\tau(\theta)/2\|^2} = 1-\sqrt{\theta^c} = \frac{2\sqrt{\theta^i}}{\tau^i(\theta)}. \tag{21}$$

Now, we compute $\widetilde{J}$. Let i and j in $\{1,\ldots,m\}$:

$$\begin{aligned}\frac{\partial\tau^i(\theta)}{\partial\theta^j} &= \frac{\delta_{ij}}{\sqrt{\theta^i}\left(1-\sqrt{\theta^c}\right)} - \frac{\sqrt{\theta^i}}{\sqrt{\theta^c}\left(1-\sqrt{\theta^c}\right)^2}, &(22)\\ &= \frac{\tau^i(\theta)}{2}\left(\frac{\delta_{ij}}{\theta^i} - \frac{\tau^i(\theta)}{2\sqrt{\theta^i\theta^c}}\right). &(23)\end{aligned}$$

Replacing Eqs. (21) and (23) into Eq. (14) yields the result. □

Proof of Fact 14. We prove the third equality (the second equality is a well-known fact of linear algebra).
Let $u \in \ker J$. Then $J^TGJu = 0$, thus $u \in \ker(J^TGJ)$. Hence $\left(\ker(J^TGJ)\right)^{\perp} \subseteq (\ker(J))^{\perp}$.
Let $v \in \ker J^TGJ$. Since G is symmetric positive-definite, the function $w \mapsto N(w) = \sqrt{w^TGw}$ is a norm. We have $0 = v^TJ^TGJv = N(Jv)^2$. The positive-definiteness of the norm N implies $Jv = 0$. Thus, $v \in \ker J$. Hence $(\ker(J))^{\perp} \subseteq \left(\ker(J^TGJ)\right)^{\perp}$. □

Proof of Proposition 18. The implication (8) ⇒ (7]) is immediate (by double inclusion).
Now, assume (7]) holds. Let $v \in D$. Define $w_1 = \epsilon\, v/\sqrt{\bar{g}_x(v,v)}$ and $w_2 = \epsilon\, v/\sqrt{\tilde{g}_x(v,v)}$. Then, with a similar argument as in the proof of Proposition 13, we can obtain Eq. (8). Note that w_2 is well defined because $v \notin \ker(J)$. □

Proof of Proposition 19. Let us first introduce the polar decomposition.
Let A be a $m \times d$ matrix.
Define the absolute value[4] of A by $|A| = (A^TA)^{\frac{1}{2}}$.
Define the linear map $u : \mathrm{rg}(|A|) \to \mathrm{rg}(A)$ by $u(|A|x) = Ax$ for any $x \in \mathbb{R}^d$.
Using the fact that $|A|$ is symmetric, we have that $\|Ax\|^2 = x^TA^TAx = (A^TAx)^Tx = (|A|^2x)^Tx = x^T|A|^T|A|x = \||A|x\|^2$, thus u is an isometry[5].
Let U be the matrix associated to u in the canonical basis.

[4] The square root of A^TA is well defined because it is a positive semidefinite matrix.

[5] We can arbitrarily extend u on the entire $\mathbb{R}^d$, e.g., by setting $\ker(u) = \ker(|A|)$.

We now prove the main result.
Let $A = PJ$. Using the polar decomposition, we have

$$PJ = U|PJ|,$$

where U is an isometry from $\mathrm{rg}(|PJ|) = (\ker|PJ|)^{\perp} = (\ker(PJ))^{\perp} = (\ker(J))^{\perp} = D$ to $\mathrm{rg}(PJ) = \mathbb{R}^m$ (using our assumption that $\mathrm{rk}(J) = m$). Transposing this relation, we obtain:

$$J^T P^T = |PJ|U^T.$$

Hence, by multiplying both relations, we have:

$$PJJ^T P^T = U|PJ|^2 U^T = UJ^T P^T PJU^T \tag{24}$$

Assume that (ii) holds, i.e., $PJJ^T P = I_m$. Then:

$$J^T GJ = J^T P^T PJ = U^T PJJ^T P^T U = U^T U.$$

Since U is an isometry from D to $\mathbb{R}^m$, then $U^T U$ is the projection onto D, denoted Π_D. Thus, we have $J^T GJ = \Pi_D$ which is (i).
Now, assume that (i) holds, i.e., $J^T P^T PJ = \Pi_D$ where Π_D is the projection onto D. We have:

$$PJJ^T P^T = UJ^T P^T PJU^T = U\Pi_D U^T.$$

Since $\mathrm{rg}(U^T) = D$, then $\Pi_D U^T = U^T$. Since U is an isometry from D to $\mathbb{R}^m$, then $UU^T = I_m$. Thus, $PJJ^T P^T = I_m$ which is (ii). □

Author Contributions: The authors contribute equally in this work.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

1. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.J.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the International Conference on Learning Representations, 2014.
2. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the International Conference on Learning Representations, 2015.
3. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the 6th International Conference on Learning Representations, 2018.
4. Carlini, N.; Wagner, D. Towards Evaluating the Robustness of Neural Networks. In Proceedings of the IEEE Symposium on Security and Privacy, 2017, pp. 39–57.
5. Gilmer, J.; Metz, L.; Faghri, F.; Schoenholz, S.S.; Raghu, M.; Wattenberg, M.; Goodfellow, I.J. Adversarial Spheres. In Proceedings of the 6th International Conference on Learning Representations, 2018.
6. Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; Zhou, B. Trustworthy AI: From Principles to Practices. *ACM Computing Surveys* **2022**, *55*.
7. Croce, F.; Hein, M. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-Free Attacks. In Proceedings of the 37th International Conference on Machine Learning (ICML), 2020.
8. Papernot, N.; McDaniel, P.D.; Wu, X.; Jha, S.; Swami, A. Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. *CoRR* **2015**.
9. Hoffman, J.; Roberts, D.A.; Yaida, S. Robust Learning with Jacobian Regularization. *ArXiv* **2018**.
10. Shen, C.; Peng, Y.; Zhang, G.; Fan, J. Defending Against Adversarial Attacks by Suppressing the Largest Eigenvalue of Fisher Information Matrix. *ArXiv* **2019**.
11. Amari, S.i. *Differential-Geometrical Methods in Statistics*; Vol. 28, *Lecture Notes in Statistics*, Springer New York, 1985.
12. Calin, O.; Udrişte, C. *Geometric Modeling in Probability and Statistics*; Springer International Publishing, 2014.
13. Čencov, N. Algebraic foundation of mathematical statistics. *Series Statistics* **1978**, *9*, 267–276.
14. Shafahi, A.; Najibi, M.; Ghiasi, M.A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L.S.; Taylor, G.; Goldstein, T. Adversarial training for free! In Proceedings of the Advances in Neural Information Processing Systems, 2019, Vol. 32.
15. Wong, E.; Rice, L.; Kolter, J.Z. Fast is better than free: Revisiting adversarial training. In Proceedings of the 8th International Conference on Learning Representations (ICLR), 2020.

16. Zhao, C.; Fletcher, P.T.; Yu, M.; Peng, Y.; Zhang, G.; Shen, C. The Adversarial Attack and Detection under the Fisher Information Metric. *Proceedings of the AAAI Conference on Artificial Intelligence* **2019**.
17. Müller, R.; Kornblith, S.; Hinton, G.E. When does label smoothing help? In Proceedings of the Advances in Neural Information Processing Systems; Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; Garnett, R., Eds. Curran Associates, Inc., 2019, Vol. 32.
18. Cissé, M.; Bojanowski, P.; Grave, E.; Dauphin, Y.N.; Usunier, N. Parseval Networks: Improving Robustness to Adversarial Examples. In Proceedings of the Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 854–863.
19. Béthune, L.; Boissin, T.; Serrurier, M.; Mamalet, F.; Friedrich, C.; González-Sanz, A. Pay Attention to Your Loss: Understanding Misconceptions about 1-Lipschitz Neural Networks, 2022.
20. Xiao, C.; Zhu, J.Y.; Li, B.; He, W.; Liu, M.; Song, D. Spatially Transformed Adversarial Examples. In Proceedings of the International Conference on Learning Representations, 2018.
21. Skovgaard, L.T. A Riemannian Geometry of the Multivariate Normal Model. *Scandinavian Journal of Statistics* **1984**, *11*, 211–223.
22. Cohen, J.; Rosenfeld, E.; Kolter, Z. Certified Adversarial Robustness via Randomized Smoothing. In Proceedings of the Proceedings of the 36th International Conference on Machine Learning, 2019, pp. 1310–1320.
23. Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; Ghaoui, L.E.; Jordan, M. Theoretically Principled Trade-off between Robustness and Accuracy. In Proceedings of the Proceedings of the 36th International Conference on Machine Learning, 2019, pp. 7472–7482.
24. Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; Madry, A. Robustness May Be at Odds with Accuracy. In Proceedings of the International Conference on Learning Representations, 2019.
25. Picot, M.; Messina, F.; Boudiaf, M.; Labeau, F.; Ben Ayed, I.; Piantanida, P. Adversarial Robustness via Fisher-Rao Regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**. https://doi.org/10.1109/TPAMI.2022.3174724.
26. Leino, K.; Wang, Z.; Fredrikson, M. Globally-Robust Neural Networks. In Proceedings of the Proceedings of the 38th International Conference on Machine Learning, 2021.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.