Adversarial Robustness with Partial Isometry

Loïc Shi-Garrier, Carla Nidhal, Daniel Delahaye

▶ **To cite this version:**

Loïc Shi-Garrier, Carla Nidhal, Daniel Delahaye. Adversarial Robustness with Partial Isometry. Entropy, 2024, 26 (103), 10.3390/e26020103 . hal-04414736

HAL Id: hal-04414736
https://enac.hal.science/hal-04414736

Submitted on 24 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

entropy

MDPI

Article

Adversarial Robustness with Partial Isometry

Loïc Shi-Garrier [1,*], Nidhal Carla Bouaynaya [2] and Daniel Delahaye [1]

[1] ENAC, Université de Toulouse, 31400 Toulouse, France; delahaye@recherche.enac.fr
[2] Department of Electrical and Computer Engineering, Rowan University, Glassboro, NJ 08028, USA; bouaynaya@rowan.edu
* Correspondence: loic.shi-garrier@aviation-civile.gouv.fr

Abstract: Despite their remarkable performance, deep learning models still lack robustness guarantees, particularly in the presence of adversarial examples. This significant vulnerability raises concerns about their trustworthiness and hinders their deployment in critical domains that require certified levels of robustness. In this paper, we introduce an information geometric framework to establish 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 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. We evaluate our approach using MNIST and CIFAR-10 datasets against adversarial attacks, revealing its substantial improvements over defensive distillation and Jacobian regularization for medium-sized perturbations and its superior robustness performance to adversarial training for large perturbations, all while maintaining the desired accuracy.

Keywords: adversarial robustness; information geometry; fisher information metric; multi-class classification

check for updates

Citation: Shi-Garrier, L.; Bouaynaya, N.C.; Delahaye, D. Adversarial Robustness with Partial Isometry. *Entropy* **2024**, *26*, 103. https://doi.org/10.3390/e26020103

Academic Editor: Boris Ryabko

Received: 7 November 2023
Revised: 9 January 2024
Accepted: 19 January 2024
Published: 24 January 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

One of the primary motivations for investigating machine learning robustness stems from the susceptibility of neural networks to adversarial attacks, wherein small perturbations in the input data can deceive the network into making the wrong decision. These adversarial attacks have been shown to be both ubiquitous and transferable [1–3]. Beyond posing a security threat, adversarial attacks underscore the glaring lack of robustness in machine learning models [4,5]. This deficiency in robustness is a critical challenge as it undermines trustworthiness in machine learning systems [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 information metric (FIM). We subsequently formulate a regularization defense method for adversarial robustness. While our focus is on l_2 white-box attacks within multi-class classification tasks, the method's applicability extends to more general settings, including unrestricted attacks and black-box attacks across various supervised learning tasks. The regularized model is evaluated on MNIST and CIFAR-10 datasets against projected gradient descent (PGD) l_∞ attacks and AutoAttack [7] with l_∞ and l_2 norms. Comparisons with the 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 against larger perturbations compared to adversarial training.

The remainder of this paper is organized as follows. Section 2 introduces notations, notions of adversarial machine learning, and definitions related to geometry. Then, we derive a sufficient condition for adversarial robustness at a given sample point. Section 3

presents our method for approximating the robustness condition, which involves promoting model isometry in the orthogonal complement of the kernel of the pullback of the FIM. In Section 4, several experiments are presented to evaluate the proposed method. Section 5 discusses the results in the context of related work on adversarial defense. Finally, Section 6 concludes the paper and outlines potential extensions of this research. Appendix A provides the proof of the results stated in the main text.

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 as $\mathrm{rg}(M)$. The rank of M is denoted as $\mathrm{rk}(M)$. The Euclidean norm (i.e., l_2 norm) is denoted as $\|\cdot\|$. 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. Adversarial Machine Learning

An adversarial attack is any strategy aiming at deliberately altering the expected behavior of a model or extracting information from a model. In this work, we focus on attacks performed at inference time (i.e., after training), sometimes referred to as evasion attacks. The most well-known evasion attacks are gradient-based. Such gradient-based attacks all follow the same idea that we explain thereafter.

To reach good accuracy and generalization, a machine learning model f (with input x and parameter w) is typically trained by minimizing a loss function $L(y, f(x, w))$ with respect to the parameters w of the model. In its simpler form, the loss function quantifies the error between the prediction of the model $f(x, w)$ and ground-truth y. Given a clean input x_0, an adversarial example x_* can be crafted by maximizing the loss function $L(y, f(x, w))$, starting from x_0 and using gradient ascent $x_{t+1} - x_t \propto \nabla_x L(y, f(x, w))$, where the gradient is computed with respect to the input x (and not the parameter w as during training). In order for x_* to be an adversarial example, x_0 and x_* must be close to each other according to some dissimilarity measure, typically a l_p norm. An adversarial example x_* is successful if the model f classifies x_* differently from x_0. Some well-known gradient-based attacks include the fast gradient sign method [2] and projected gradient descent [3].

Adversarial attacks can be classified according to their threat model. White-box attacks assume that the adversary has perfect knowledge of the targeted model, including access to the training data, model architecture, and model parameters. Such an adversary can directly compute the gradient $\nabla_x L(y, f(x, w))$ of the targeted model and craft adversarial examples. More realistic threat models are classified as gray-box or black-box attacks, where some or all of the information is unknown to the adversary. In this work, we use both white-box attacks as well as simple gray-box attacks where the adversary can access the training data and model architecture, but not the model parameters. To craft such gray-box adversarial examples, another model is trained with the same data and architecture. Then, white-box attacks are performed on this model. Finally, the adversarial examples can be transferred to the targeted model.

Adversarial robustness aims to build models that classify both x_* and x_0 with the same class while preserving sufficient accuracy for the clean examples x_0. Various defenses have been proposed to improve adversarial robustness. The most efficient defense is called adversarial training, which was first described in [2] and further developed in [3]. The idea behind adversarial training is to obtain the parameters w_* of the trained model as:

$$w_* = \arg\min_w \max_{\epsilon \in \Delta(x)} L(y, f(x + \epsilon, w)),$$

in place of the original parameters $\arg\min_w L(y, f(x, w))$. The set $\Delta(x)$ is a set of allowed adversarial attacks for x, e.g., a l_2 ball with a given radius (or budget). In practice, adversar-

ial training is performed by adding adversarial examples to the training set, thus providing a lower bound for $\max_{\epsilon \in \Delta(x)} L(y, f(x+\epsilon, w))$.

2.3. Geometrical 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 a d-dimensional embedded smooth connected submanifold of $\mathbb{R}^d$. Let $m = c - 1$.

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

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

Δ^m is a smooth submanifold of $\mathbb{R}^c$ of dimension m. We can see $\theta = (\theta^1, \ldots, \theta^m)$ as a coordinate system from Δ^m to $\mathbb{R}^m$. Then, let us 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: $\theta = \text{softmax}(s(x))$.

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 $\mathbb{R}^c$, for all $\theta \in \Delta^m$ and all tangent vectors $v, w \in T_\theta \Delta^m$, we have*

$$g_\theta(v, w) = v^T G_\theta w, \tag{2}$$

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

$$G_{\theta,ij} = \frac{\delta_{ij}}{\theta^i} + \frac{1}{\theta^c}. \tag{3}$$

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

The FIM has two remarkable properties. First, it is the "infinitesimal distance" of the relative entropy, which is the loss function used to train a multi-class classification model. More precisely, if D is the relative entropy (also known as the Kullback–Leibler divergence)

and if d is the Fisher–Rao distance, then given two distributions θ_1 and θ_2, we have (see Theorem 4.4.5 in [12]):

$$D(\theta_1||\theta_2) = \frac{1}{2}d^2(\theta_1,\theta_2) + o\Big(d^2(\theta_1,\theta_2)\Big).$$

The same result can be restated infinitesimally using the FIM g, as follows:

$$D(\theta||\theta+d\theta) = \frac{1}{2}g_\theta(d\theta,d\theta) + o(g_\theta(d\theta,d\theta)), \tag{4}$$

where $d\theta$ is seen as a tangent vector of $T_\theta\mathcal{S}$.

The other remarkable property of the FIM is Chentsov's theorem [13], claiming that the FIM is the unique Riemannian metric on Δ^m, which is invariant under sufficient statistics (up to a multiplicative constant). Informally, the FIM is the only Riemannian metric that is statistically meaningful. In [14], Amari and Nagaoka state a more general result. Along with the FIM, they introduce a family of affine connections parameterized by a real parameter α, called the α-connections. Theorem 2.6 in [14] states that an affine connection is invariant under sufficient statistics if and only if it is an α-connection for some $\alpha \in \mathbb{R}$. In other words, the α-connections are the only affine connections that have a statistical meaning. While Equation (4) gives the second-order approximation of the relative entropy, an α-connection can be seen as the third-order term in the Taylor approximation of some divergence [14]. More precisely, a given α-connection can be canonically associated with a unique divergence (while the second-order term is always given by the FIM). If $\alpha = \pm 1$, the canonical divergences are the relative entropy and its dual (obtained by switching the arguments in $D(\theta_2||\theta_1)$). More generally, for $\alpha \neq 0$, the canonical divergence is not symmetric. The only canonical divergence that is symmetric is obtained for $\alpha = 0$, and is precisely the square of the Fisher–Rao distance. Thus, the Fisher–Rao distance is the only statistically meaningful distance. This provides a motivation for using the Fisher–Rao distance to measure lengths in Δ^m.

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 $\overline{g}$. 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*

$$\overline{g}_x(v,w) = v^T w, \tag{5}$$

thus, its matrix is the identity matrix of dimension d, denoted as 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 $f : (\mathcal{X}, \overline{g}) \to (\Delta^m, g)$. 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 $\epsilon > 0$.

Define the Euclidean open ball centered at x with radius ϵ by

$$\overline{b}(x,\epsilon) = \Big\{z \in \mathbb{R}^d : \|z - x\| < \epsilon\Big\}. \tag{6}$$

Definition 4. *Define the set (Figure 1):*

$$\mathcal{A}_x = \Big\{\theta \in \Delta^m : \arg\max_i \theta^i = \arg\max_i f^i(x)\Big\}. \tag{7}$$

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

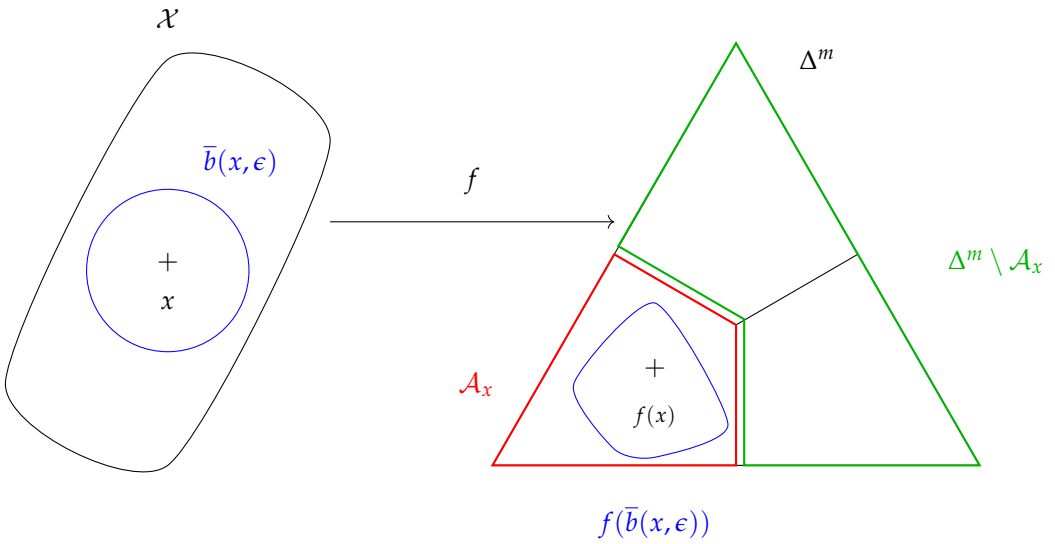

Figure 1. ϵ-robustness at x is enforced if and only if $f(\overline{b}(x,\epsilon)) \subseteq \mathcal{A}_x$.

The set $\mathcal{A}_x$ is the subset of distributions of Δ^m that have the same class as $f(x)$.

Definition 5 (Geodesic ball of the FIM). *Let $\delta > 0$ be the Fisher–Rao distance between $f(x)$ and $\Delta^m \setminus \mathcal{A}_x$ (Figure 2), i.e., the Fisher–Rao distance between $f(x)$ and the closest distribution of Δ^m with a different class.*

Define the geodesic ball centered at $f(x) \in \Delta^m$ with radius δ by

$$b(f(x),\delta) = \{\theta \in \Delta^m : d(f(x),\theta) \leq \delta\}. \tag{8}$$

In Section 3.3, we propose an efficient approximation of δ.

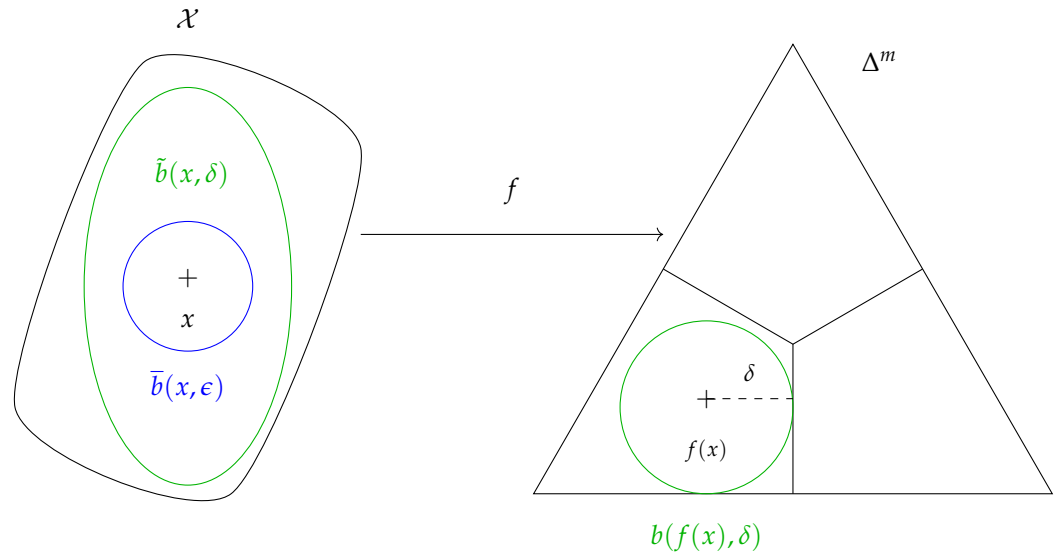

Figure 2. ϵ-robustness at x is enforced if $\overline{b}(x,\epsilon) \subseteq \tilde{b}(x,\delta)$.

Definition 6 (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*

$$\tilde{g}_x(v,w) = v^T J_x^T G_{f(x)} J_x w, \tag{9}$$

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

$$\widetilde{G}_x = J_x^T G_{f(x)} J_x. \tag{10}$$

Definition 7 (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*

$$\tilde{b}(x,\delta) = \left\{ z \in \mathbb{R}^d : \tilde{d}(x,z) \leq \delta \right\}. \tag{11}$$

Note that the radius δ is the Fisher–Rao distance between $f(x)$ and $\Delta^m \setminus \mathcal{A}_x$ as defined in Definition 5.

2.4. Robustness Condition

Definition 8 (Robustness)**.** *We say that f is ϵ-robust at x if*

$$\forall z \in \mathbb{R}^d, \|z - x\| < \epsilon \Rightarrow f(z) \in \mathcal{A}_x. \tag{12}$$

Equivalently, we can write (Figure 1):

$$f(\overline{b}(x,\epsilon)) \subseteq \mathcal{A}_x. \tag{13}$$

Proposition 1 (Sufficient condition for robustness)**.** *If $\overline{b}(x,\epsilon) \subseteq \tilde{b}(x,\delta)$, then f is ϵ-robust at x (Figure 2).*

Our goal is to start from Proposition 1 and make several assumptions in order to derive a condition that can be efficiently implemented.

Working with geodesic balls $\overline{b}(x,\epsilon)$ and $\tilde{b}(x,\delta)$ is intractable, so our first assumption consists of using an "infinitesimal" condition by restating Proposition 1 in the tangent space $T_x\mathcal{X}$ instead of working directly on $\mathcal{X}$. In $T_x\mathcal{X}$, define the Euclidean ball of radius ϵ by

$$\overline{\mathcal{B}}_x(0,\epsilon) = \left\{ v \in T_x\mathcal{X} : \overline{g}_x(v,v) = v^T v \leq \epsilon^2 \right\}. \tag{14}$$

Similarly, in $T_x\mathcal{X}$, define the $\tilde{g}_x$-ball of radius δ by

$$\widetilde{\mathcal{B}}_x(0,\delta) = \left\{ v \in T_x\mathcal{X} : \tilde{g}_x(v,v) = v^T \widetilde{G}_x v \leq \delta^2 \right\}. \tag{15}$$

Assumption 1. *We replace Proposition 1 by*

$$\overline{\mathcal{B}}_x(0,\epsilon) \subseteq \widetilde{\mathcal{B}}_x(0,\delta). \tag{16}$$

Proposition 2. *Equation (16) is equivalent to*

$$\forall v \in T_x\mathcal{X}, \quad \tilde{g}_x(v,v) \leq \frac{\delta^2}{\epsilon^2} \overline{g}_x(v,v). \tag{17}$$

Since $m < d$, the Jacobian matrix J_x has a 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 a rank of at most m (when J_x has a rank of m).

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

Using Assumptions 1 and 2, the constant rank theorem ensures that for small enough δ, f is ϵ-robust at x. However, contrary to Proposition 1, Assumption 1 does not offer any guarantee on the ϵ-robustness at x for arbitrary δ.

3. Derivation of the Regularization Method

In this section, we derive a condition for robustness (Proposition 4), which can be implemented as a regularization method. Then, we provide two useful results for the practical implementation of this method: an explicit formula for the decomposition of the FIM as $G = P^T P$ (Section 3.2), and an easy-to-compute upper-bound of δ, i.e., the Fisher–Rao distance between $f(x)$ and $\Delta^m \setminus \mathcal{A}_x$ (section 3.3).

3.1. The Partial Isometry Condition

In order to simplify the notations, we replace

- J_x with J, which is a full-rank $m \times d$ real matrix.
- $G_{f(x)}$ with G, which is an $m \times m$ symmetric positive definite real matrix.
- $\widetilde{G}_x$ with $\widetilde{G}$, 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 1.

$$D = \mathrm{rg}(J^T) = (\ker(J))^\perp = \Big(\ker(J^T G J)\Big)^\perp \tag{18}$$

Fact 2. *$J^T G J$ is symmetric positive semidefinite. Thus, by the spectral theorem, the eigenvectors associated with 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 as $\mathcal{B} = (e_1, \ldots, e_m, e_{m+1}, \ldots, e_d)$, such that each e_i is an eigenvector of $J^T G J$, and such that $(e_1, \ldots, e_m)$ is a basis of $D = \mathrm{rg}(J^T)$ and $(e_{m+1}, \ldots, e_d)$ is a basis of $\ker(J)$.

The set $D = \mathrm{rg}(J^T)$ is an m-dimensional subspace of $T_x\mathcal{X}$. $\tilde{g}_x$ does not define an inner product on $T_x\mathcal{X}$ because $\widetilde{G}$ has a nontrivial kernel of dimension $d - m$. In particular, the set $\widetilde{\mathcal{B}}_x(0, \delta)$ is not bounded, i.e., it is a cylinder rather than a ball. However, when restricted to D, $\tilde{g}_x|_D$ defines an inner product. We define the restriction of $\widetilde{\mathcal{B}}_x(0, \delta)$ to D:

$$\widetilde{\mathcal{B}}_D(0, \delta) = \Big\{ v \in D : v^T \widetilde{G} v \leq \delta \Big\}, \tag{19}$$

and similarly, we define the restriction of $\overline{\mathcal{B}}_x(0, \epsilon)$ to D:

$$\overline{\mathcal{B}}_D(0, \epsilon) = \Big\{ v \in D : v^T v \leq \epsilon^2 \Big\}. \tag{20}$$

Assume that f is such that Equation (16) holds (i.e., $\overline{\mathcal{B}}_x(0, \epsilon) \subseteq \widetilde{\mathcal{B}}_x(0, \delta)$). Moreover, assume that we are in the limit case defined as follows: for any perturbation size, we can find a smaller perturbation of f such that Equation (16) does not hold anymore. This limit case is equivalent to having $\overline{\mathcal{B}}_D(0, \epsilon) = \widetilde{\mathcal{B}}_D(0, \delta)$. In this case, $\widetilde{\mathcal{B}}_x(0, \delta)$ is the smallest possible $\tilde{g}_x$-ball (for the inclusion) such that Equation (16) holds. We noticed experimentally that enforcing this stronger criteria yields a larger robustifying effect. Thus, we make the following assumption:

Assumption 3. *We replace Equation (16) with*

$$\overline{\mathcal{B}}_D(0, \epsilon) = \widetilde{\mathcal{B}}_D(0, \delta). \tag{21}$$

Proposition 3. *Equation (21) is equivalent to*

$$\forall v \in D, \quad \tilde{g}_x(v, v) = \frac{\delta^2}{\epsilon^2} \overline{g}_x(v, v). \tag{22}$$

We can rewrite Equation (22) in matrix form:

$$\forall v \in D, \quad v^T \widetilde{G} v = \frac{\delta^2}{\epsilon^2} v^T v. \tag{23}$$

In Section 3.2, we 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, which is equivalent to Equation (23).

Proposition 4. *The following statements are equivalent:*

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

$$(ii) \qquad P J J^T P^T = \frac{\delta^2}{\epsilon^2} I_m,$$

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

Proposition 4 constrains the matrix PJ to be a semi-orthogonal matrix (multiplied by a homothety matrix). A smooth map f between Riemannian manifolds $(\mathcal{X}, \overline{g})$ and (Δ^m, g) is said to be (locally) isometric if the pullback metric (denoted $f^* g$) coincides with $\overline{g}$, i.e., $f^* g = \overline{g}$. Such a map f locally preserves distances. In our case, $f^* g = \tilde{g}$ is not a metric (since its kernel is non-trivial); thus, f cannot be an isometry. However, Equation (22) ensures that f locally preserves distances along directions spanned by D. Hence, f becomes a partial isometry, at least in the neighborhood of the training points.

Under the Assumptions 1–3, Equation (ii) in Proposition 4 implies robustness as defined in Definition 8. In other words, Equation (ii) is a sufficient condition for robustness. However, there is no reason for a neural network to satisfy Equation (ii). This is why we define the following regularization term:

$$\alpha(x, \epsilon, f) = \frac{1}{m^2} \left\| P J J^T P^T - \frac{\delta^2}{\epsilon^2} I_m \right\|, \tag{24}$$

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. Finally, the loss function is:

$$L(y, x, \epsilon, f) = l(y, f(x)) + \lambda \, \alpha(x, \epsilon, f), \tag{25}$$

where l is the cross-entropy loss, and $\lambda > 0$ is a hyperparameter controlling the strength of the regularization with respect to the cross-entropy loss. The regularization term $\alpha(x, \epsilon, f)$ is minimized during training, such that the model is pushed to satisfy the sufficient condition of robustness.

3.2. Coordinate Change

In this subsection, we show how to compute the matrix P that was introduced in Proposition 4. To this end, we isometrically embed Δ^m into the Euclidean space $\mathbb{R}^c$ using the following inclusion map:

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

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 5. *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 $\mathbb{R}^m$:

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

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

Proposition 6. *In the coordinates τ, the FIM is:*

$$G_{\tau,ij} = \frac{4}{\left(1 + \|\tau/2\|^2\right)^2}\delta_{ij}. \tag{26}$$

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

$$G = \widetilde{J}^T G_\tau \widetilde{J} = \frac{4}{\left(1 + \|\tau/2\|^2\right)^2}\widetilde{J}^T\widetilde{J}. \tag{27}$$

Thus, we can choose:

$$P = \frac{2}{1 + \|\tau/2\|^2}\widetilde{J}. \tag{28}$$

Write $f(x) = \theta = (\theta^1, \ldots, \theta^m)$ 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 $i = 1, \ldots, m$. More explicitly, we have:

Proposition 7. *For $i, j = 1, \ldots, m$:*

$$P_{ij} = \frac{\delta_{ij}}{\sqrt{\theta^i}} - \frac{\tau^i(\theta)}{2\sqrt{\theta^c}}. \tag{29}$$

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 $\Delta^m \setminus \mathcal{A}_x$). In Proposition 5, we show 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:

$$\cos(\beta) = \frac{1}{4}\sum_{i=1}^c \mu_1^i \mu_2^i = \sum_{i=1}^c \sqrt{\theta_1^i \theta_2^i}. \tag{30}$$

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

$$d(\theta_1, \theta_2) = 2\arccos \sum_{i=1}^c \sqrt{\theta_1^i \theta_2^i}. \tag{31}$$

In the regularization term defined in Equation (24), we replace δ with the following upper bound:

$$\delta = d(f(x), \Delta^m \setminus \mathcal{A}_x) \leq d(f(x), O), \tag{32}$$

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

$$\delta \leq 2\arccos \sum_{i=1}^c \sqrt{\frac{f^i(x)}{c}}. \tag{33}$$

4. Experiments

The regularization method introduced in Section 3 is evaluated on MNIST and CIFAR-10 datasets. Our method uses the loss function introduced in Equation (25).

4.1. Experiments on MNIST Dataset

4.1.1. Experimental Setup

For the MNIST dataset, we implement a LeNet model with two convolutional layers of 32 and 64 channels, respectively, followed by one hidden layer with 128 neurons. The code is available here: https://github.com/lshigarrier/geometric_robustness.git (accessed on 1 December 2022). We train three models: one regularized model, one baseline unregularized model, and one model trained with adversarial training. All three models are trained with the 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 trade-off; 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 and then tested on 10,000 images of the test set. The baseline model achieves an accuracy of 98.9% (9893/10,000), the regularized model achieves an accuracy of 94.0% (9403/10,000), and the adversarially trained model achieves an accuracy of 98.8% (9883/10,000). Although the current implementation of the regularized model is almost six 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. [15], 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 use 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 are correctly classified by each model. The results are reported in Figure 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 the 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 a fair comparison, and in addition to a baseline model (BASE), we compare the partial isometry defense (ISO) with several other computationally efficient defenses: 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. ISO is the best defense that does not rely on adversarial training. In future work, ISO may be combined with AT to further boost performance. Note that ISO and JAC are more robust against l_2 attacks since they were designed to defend the model against such attacks. On the other hand, AT is more robust against l_∞ attacks, because the adversarial training was conducted with the l_∞ norm.

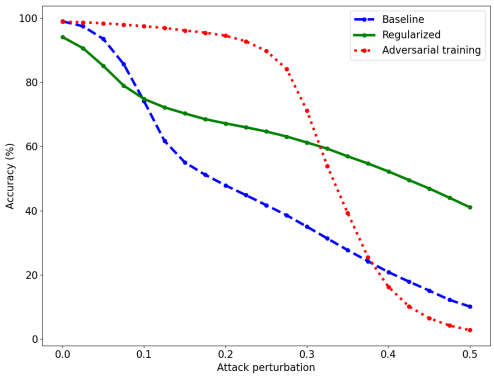

Figure 3. Accuracy of the baseline (dashed, blue), regularized (solid, green), and adversarially trained (dotted, red) models for various attack perturbations on the MNIST dataset. The perturbations are obtained with PGD using l_∞ norm.

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

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

4.2. Experiments on CIFAR-10 Dataset

We consider a DenseNet121 model fine-tuned on CIFAR-10 using pre-trained weights for ImageNet. The code is available here: https://github.com/lshigarrier/iso_defense.git (accessed on 26 January 2023). As for the MNIST experiments, we compare the partial isometry defense with distillation (DIST), Jacobian regularization (JAC), and Fisher information regularization (FIR). Here, adversarial training (AT) relies on the fast gradient sign method (FGSM) attack [16]. All defenses are compared against PGD for various attack strengths. The results are presented 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). Apart from high computational costs, AT suffers from several limitations: it only robustifies against the chosen attack at the chosen budget and it does not offer a robustness guarantee. For example, under Gaussian noise, AT accuracy decreases faster than baseline accuracy (i.e., no defense). Achieving high robustness accuracy against specific attacks on a 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.

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

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

5. Discussion and Related Work

In 2019, Zhao et al. [17] proposed to use the Fisher information metric in the setting of adversarial attacks. They used the eigenvector associated with 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 [18]. In our framework, their method consists of 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, 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. [19] introduced another adversarial defense called Parseval networks. To achieve adversarial robustness, the authors aim to control 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 another name for 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 [20].

Following another line of work, Hoffman et al. [9] advanced a Jacobian regularization to improve adversarial robustness. Their regularization consists of using the Frobenius 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 $\epsilon_2 = \sqrt{d}\epsilon_\infty$, 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, and color changes of the original image. Contrary to the l_2 or l_∞ norms, these measures do not rely on a pixel-based coordinate system. However, it is possible to define unrestricted attacks based on these spatial dissimilarities, for example, in [21].

In this work, we derive 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 [22].

Finally, the precise quantification of the robustness condition presented in Equation (12) and Proposition 4 paves the way for the development of a certified defense [23] in this framework. By strongly enforcing Proposition 4 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 [24,25]. Certifiable defenses are a necessary 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. Conclusions and Future Work

In this paper, we introduce 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 a 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 will consider other "distinguishability" measures.

Our work extends a recent, promising but understudied framework for adversarial robustness based on information geometric tools. The FIM has already been harnessed to develop attacks [17] and defenses [10,26] but a precise robustness analysis is yet to be proposed. Our work is a step toward the development of such an analysis, which might 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 an 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 [27] or randomized smoothing [23], where statistical models naturally appear.

Author Contributions: Conceptualization, L.S.-G., N.C.B. and D.D.; methodology, L.S.-G. and N.C.B.; software, L.S.-G.; validation, L.S.-G.; formal analysis, L.S.-G., N.C.B. and D.D.; investigation, L.S.-G.; writing—original draft preparation, L.S.-G.; writing—review and editing, L.S.-G., N.C.B. and D.D.; supervision, N.C.B. and D.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: http://yann.lecun.com/exdb/mnist/ (accessed on 1 December 2022) and https://www.cs.toronto.edu/~kriz/cifar.html (accessed on 26 January 2023).

Acknowledgments: We thank Roman Shterenberg for useful discussions.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Proofs

Proof of Proposition 2. (17) ⇒ (16). Assume (17). Let $v \in \overline{\mathcal{B}}_x(0,\epsilon)$. Thus, $\overline{g}_x(v,v) \leq \epsilon^2$. We have

$$\tilde{g}_x(v,v) \leq \frac{\delta^2}{\epsilon^2}\overline{g}_x(v,v) \leq \frac{\delta^2}{\epsilon^2}\epsilon^2 = \delta^2. \tag{A1}$$

Thus, $v \in \widetilde{\mathcal{B}}_x(0,\delta)$.

(16) ⇒ (17). Assume (16). Let $v \in T_x\mathcal{X}$, $v \neq 0$. 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

$$\tilde{g}_x(w,w) = \frac{\epsilon^2}{\overline{g}_x(v,v)}\tilde{g}_x(v,v) < \delta^2. \tag{A2}$$

We obtain Equation (17) by multiplying by $\overline{g}_x(v,v)/\epsilon^2$. □

Proof of Fact 1. We prove the third equality (the second equality is a well-known fact of linear algebra).

Let $u \in \ker J$. Then, $J^T G J u = 0$; thus, $u \in \ker(J^T G J)$. Hence, $\big(\ker(J^T G J)\big)^\perp \subseteq (\ker(J))^\perp$.

Let $v \in \ker J^T G J$. Since G is symmetric positive-definite, the function $w \mapsto N(w) = \sqrt{w^T G w}$ is a norm. We have $0 = v^T J^T G J v = N(Jv)^2$. The positive-definiteness of the norm N implies $Jv = 0$. Thus, $v \in \ker J$. Hence, $(\ker(J))^\perp \subseteq \big(\ker(J^T G J)\big)^\perp$. □

Proof of Proposition 3. The implication (22) ⇒ (21) is immediate (by double inclusion).

Now, assume (21) 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 2, we can obtain Equation (22). Note that w_2 is well-defined because $v \notin \ker(J)$. □

Proof of Proposition 4. Let us first introduce the polar decomposition.

Let A be a $m \times d$ matrix.

Define the absolute value of A by $|A| = (A^T A)^{\frac{1}{2}}$. Note that the square root of $A^T A$ is well-defined because it is a positive-semidefinite matrix

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 $\|Ax\|^2 = x^T A^T A x = (A^T A x)^T x = (|A|^2 x)^T x = x^T |A|^T |A| x = \||A|x\|^2$; thus, u is an isometry (we can arbitrarily extend u on the entire $\mathbb{R}^d$, e.g., by setting $\ker(u) = \ker(|A|)$.).

Let U be the matrix associated to u in the canonical basis.

We now prove the main result.

Let $A = PJ$. Using the polar decomposition, we have

$$PJ = U|PJ|, \tag{A3}$$

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

$$J^T P^T = |PJ| U^T. \tag{A4}$$

Hence, by multiplying both relations, we have

$$PJJ^T P^T = U|PJ|^2 U^T = UJ^T P^T PJ U^T \tag{A5}$$

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

$$J^T G J = J^T P^T PJ = U^T PJJ^T P^T U = U^T U. \tag{A6}$$

Since U is an isometry from D to $\mathbb{R}^m$, then $U^T U$ is the projection onto D, denoted as Π_D. Thus, we have $J^T G J = \Pi_D$, 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

$$PJJ^T P^T = UJ^T P^T PJ U^T = U \Pi_D U^T. \tag{A7}$$

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). □

Proof of Proposition 5. 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

$$\frac{\partial \mu^\alpha(\theta_0)}{\partial \theta^i} = \frac{\delta_{i\alpha}}{\sqrt{\theta_0^i}}, \tag{A8}$$

and for $\alpha = c$:

$$\frac{\partial \mu^c(\theta_0)}{\partial \theta^i} = -\frac{1}{\sqrt{\theta_0^c}}, \tag{A9}$$

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

$$G_{\theta, ij} = \frac{\delta_{ij}}{\theta_0^i} + \frac{1}{\theta_0^c}, \tag{A10}$$

which is the FIM, as defined in Definition 2. □

Proof of Proposition 6. For $i = 1, \ldots, m$, the inverse transformation of $\tau(\mu)$ is

$$\mu^i(\tau) = \frac{2\tau^i}{1 + \|\tau/2\|^2}, \tag{A11}$$

and

$$\mu^c(\tau) = 2\frac{\|\tau/2\|^2 - 1}{\|\tau/2\|^2 + 1}. \tag{A12}$$

The proofs of Equations (A11) and (A12) are provided below.

Moreover, according to Proposition 5, the FIM in the coordinates $(\mu^1, \ldots, \mu^m)$ 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 $i = 1, \ldots, m$ and $\alpha = 1, \ldots, m$, we have

$$\frac{\partial \mu^\alpha(\tau)}{\partial \tau^i} = \frac{2}{1 + \|\tau/2\|^2} \left(\delta_{i\alpha} - \frac{\tau^\alpha \tau^i}{2(1 + \|\tau/2\|^2)} \right), \tag{A13}$$

and for $\alpha = c$:

$$\frac{\partial \mu^c(\tau)}{\partial \tau^i} = \frac{2\tau^i}{(1 + \|\tau/2\|^2)^2}, \tag{A14}$$

Thus,

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

□

Proof of Equations (A11) and (A12). 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

$$\|\tau\|^2 = \sum_{i=1}^{m}(\tau^i)^2 = \lambda^2\|\mu\|^2. \tag{A15}$$

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

$$\|\tau\|^2 = \lambda^2\left(4 - (\mu^c)^2\right) = 4\frac{4-(\mu^c)^2}{(2-\mu^c)^2} = 4\frac{2+\mu^c}{2-\mu^c}. \tag{A16}$$

Isolating μ^c, we obtain

$$\mu^c(\tau) = \frac{2\|\tau\|^2 - 8}{\|\tau\|^2 + 4} = 2\frac{\|\tau/2\|^2 - 1}{\|\tau/2\|^2 + 1}. \tag{A17}$$

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

$$\mu^i(\tau) = \frac{\tau^i}{\lambda} = \frac{2\tau^i}{1+\|\tau/2\|^2} \tag{A18}$$

□

Proof of Proposition 7. We have

$$\tau^i(\theta) = 2\sqrt{\theta^i}\Big/\left(1 - \sqrt{\theta^c}\right). \tag{A19}$$

Thus,

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

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

$$\frac{2}{1+\|\tau(\theta)/2\|^2} = 1 - \sqrt{\theta^c} = \frac{2\sqrt{\theta^i}}{\tau^i(\theta)}. \tag{A20}$$

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

$$\frac{\partial \tau^i(\theta)}{\partial \theta^j} = \frac{\delta_{ij}}{\sqrt{\theta^i}\left(1-\sqrt{\theta^c}\right)} - \frac{\sqrt{\theta^i}}{\sqrt{\theta^c}\left(1-\sqrt{\theta^c}\right)^2}, \tag{A21}$$

$$= \frac{\tau^i(\theta)}{2}\left(\frac{\delta_{ij}}{\theta^i} - \frac{\tau^i(\theta)}{2\sqrt{\theta^i\theta^c}}\right). \tag{A22}$$

Replacing Equations (A20) and (A22) with Equation (28) yields the result. □

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, Banff, AB, Canada, 14–16 April 2014.
2. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
3. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
4. Carlini, N.; Wagner, D. Towards Evaluating the Robustness of Neural Networks. In Proceedings of the IEEE Symposium on Security and Privacy, San Jose, CA, USA, 22–24 May 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 International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 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 Comput. Surv.* **2022**, *55*, 1–46. [CrossRef]
7. Croce, F.; Hein, M. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-Free Attacks. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020.
8. Papernot, N.; McDaniel, P.D.; Wu, X.; Jha, S.; Swami, A. Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. In Proceedings of the IEEE Symposium on Security and Privacy, San Jose, CA, USA, 22–26 May 2016.
9. Hoffman, J.; Roberts, D.A.; Yaida, S. Robust Learning with Jacobian Regularization. *arXiv* **2018**, arXiv:1908.02729.
10. Shen, C.; Peng, Y.; Zhang, G.; Fan, J. Defending Against Adversarial Attacks by Suppressing the Largest Eigenvalue of Fisher Information Matrix. *arXiv* **2019**, arXiv:1909.06137.
11. Amari, S.i. *Differential-Geometrical Methods in Statistics*; Lecture Notes in Statistics; Springer: New York, NY, USA, 1985; Volume 28.
12. Calin, O.; Udrişte, C. *Geometric Modeling in Probability and Statistics*; Springer International Publishing: Berlin/Heidelberg, Germany, 2014.
13. Čencov, N. Algebraic foundation of mathematical statistics. *Ser. Stat.* **1978**, *9*, 267–276. [CrossRef]
14. Amari, S.I.; Nagaoka, H. *Methods of Information Geometry*; American Mathematical Society: Providence, RI, USA, 2000.
15. 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, Vancouver, BC, Canada, 8–14 December 2019.
16. Wong, E.; Rice, L.; Kolter, J.Z. Fast is better than free: Revisiting adversarial training. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
17. Zhao, C.; Fletcher, P.T.; Yu, M.; Peng, Y.; Zhang, G.; Shen, C. The Adversarial Attack and Detection under the Fisher Information Metric. In Proceedings of the AAAI Conference on Artificial Intelligence Honolulu, HI, USA, 27 January–1 February 2019.
18. Müller, R.; Kornblith, S.; Hinton, G.E. When does label smoothing help? In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.
19. Cissé, M.; Bojanowski, P.; Grave, E.; Dauphin, Y.N.; Usunier, N. Parseval Networks: Improving Robustness to Adversarial Examples. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 854–863.
20. 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. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022.
21. 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, Vancouver, BC, Canada, 30 April–3 May 2018.
22. Skovgaard, L.T. A Riemannian Geometry of the Multivariate Normal Model. *Scand. J. Stat.* **1984**, *11*, 211–223.
23. Cohen, J.; Rosenfeld, E.; Kolter, Z. Certified Adversarial Robustness via Randomized Smoothing. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 1310–1320.
24. 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 International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 7472–7482.
25. 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, New Orleans, LA, USA, 6–9 May 2019.

26. Picot, M.; Messina, F.; Boudiaf, M.; Labeau, F.; Ben Ayed, I.; Piantanida, P. Adversarial Robustness via Fisher-Rao Regularization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2698–2710. [CrossRef] [PubMed]
27. Leino, K.; Wang, Z.; Fredrikson, M. Globally-Robust Neural Networks. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 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.