Adversarial defense with differential geometry

1 Introduction

The study of machine learning robustness was motivated by 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].

In this paper, we shed an information geometric perspective to adversarial robustness in machine learning models. First, we formalize robustness against l_2 adversary with a condition relating the Fisher information metric (FIM) with the Euclidean metric. Then, we derive several approximations of this condition that can be efficiently implemented as regularization defense methods. We focus on 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 various defense methods are evaluated on Fashion-MNIST and CIFAR-10 against FGSM attack, PGD l_∞ attack, and AutoAttack [6].

The remaining of this paper is divided into six sections. Section 2 introduces the notations and definitions. Then, a sufficient condition for adversarial robustness at a sample point is derived, as well as two results used in the following sections. Section 3 presents our methods to approximate the robustness condition. The first method relies on encouraging the model to be isometric in the orthogonal complement of the kernel of the pullback of the FIM. The other methods are variants of the former obtained by randomization or by relaxing some assumptions. Section 4 presents several experiments to evaluate the proposed method. Section 5 discusses the results in the lights of related work on adversarial defense. 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 Preliminary results

2.1 Notations

Let $d, c \in \mathbb{N}^*$ such that $d \geq c > 1$. Let $m = c - 1$.
The range of a matrix M is denoted $\mathrm{rg}(M)$, its rank is denoted $\mathrm{rk}(M)$, and its spectrum is denoted $\mathrm{sp}(M)$. The Euclidean norm is denoted $\|\cdot\|$. We use the notation $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.
If M_1 and M_2 are two symmetric matrices, then $M_1 \preceq M_2$ means that $M_2 - M_1$ is positive semidefinite. If M is any matrix, its spectral norm (i.e., largest singular value) is denoted $\|M\|_2$. Moreover, we write $\|M\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{m} |M_{ij}|$ and $\|M\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |M_{ij}|$. We write $(M)_+$ the maximum between 0 and the largest eigenvalue of M. If r is any real number, we write $(r)_+ = \max\{0, r\}$.
For any real-valued function h_1, its gradient at a point x is denoted $\partial_x h_1$.
For any vector-valued function h_2, its Jacobian matrix at a point x is denoted $\nabla_x h_2$.
Following the convention in differential geometry, we denote the components of a vector v by v^i with a superscript.

2.2 Definitions

Definition 2.1 (Probability simplex)**.** Define the *probability simplex* of dimension m by:

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

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: $\theta = (\theta^1, \dots, \theta^m)$. Then, we define $\theta^c = 1 - \sum_{i=1}^{m} \theta^i$.

Definition 2.2 (Fisher information metric)**.** We endow Δ^m with the *Fisher information metric* (FIM) g. For each $\theta \in \Delta^m$, the FIM 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 [7]). The FIM induces a distance on Δ^m called the *Fisher-Rao distance* denoted $d(\theta_1, \theta_2)$ for any $\theta_1, \theta_2 \in \Delta^m$.

Definition 2.3 (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:

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

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: $\|x_1 - x_2\|_2$ for any $x_1, x_2 \in \mathbb{R}^d$.

From now on, we fix:

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

Definition 2.4 (Euclidean ball)**.** Define the Euclidean open ball centered at x with radius ϵ by:

$$\overline{b}(x, \epsilon) = \left\{ z \in \mathbb{R}^d : \|z - x\|_2 < \epsilon \right\}.$$

Definition 2.5. Define the set $\mathcal{A} = \left\{ \theta \in \Delta^m : \arg\max_i \theta^i = \arg\max_i f^i(x) \right\}$ (Figure 1). 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 2.6 (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 2.5, we propose a efficient approximation of δ.

Definition 2.7 (Pullback metric)**.** On $\mathbb{R}^d$, 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 $v, w \in T_x \mathbb{R}^d$ 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 $\tilde{g}_x$ is the standard coordinates of $\mathbb{R}^d$ by:

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

Definition 2.8 (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:

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

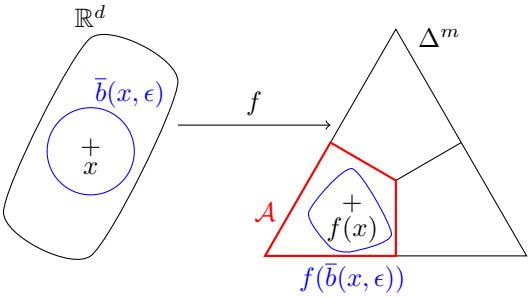

Figure 1: ϵ-robustness at x is enforced if and only if $f(\overline{b}(x,\epsilon)) \subseteq \mathcal{A}$.

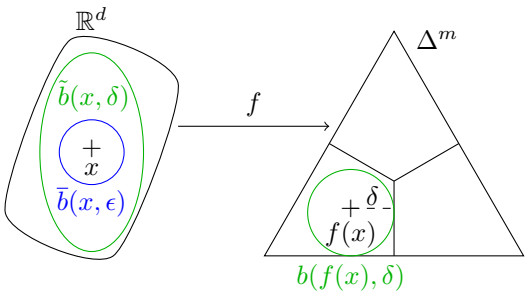

Figure 2: ϵ-robustness at x is enforced if $\overline{b}(x,\epsilon) \subseteq \tilde{b}(x,\delta)$.

2.3 Robustness condition

Definition 2.9 (Robustness)**.** We say that f is ϵ-robust at x if:

$$\forall z \in \mathbb{R}^d, \|z - x\|_2 < \epsilon \Rightarrow f(z) \in \mathcal{A}. \tag{3}$$

Equivalently, we can write (Figure 1):

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

Proposition 2.10 (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 2.10 and make several assumptions in order to derive a condition that can be efficiently implemented.

Working with geodesic balls $\overline{b}(x,\eta)$ and $\tilde{b}(x,\delta)$ is intractable, so our first assumption consists in using an "infinitesimal" condition by restating Proposition 2.10 in the tangent space $T_x\mathbb{R}^d$ instead of working directly on $\mathbb{R}^d$.

Definition 2.11. In $T_x\mathbb{R}^d$, define the Euclidean ball of radius ϵ by:

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

Definition 2.12. In $T_x\mathbb{R}^d$, define the $\tilde{g}_x$-ball of radius δ by:

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

Assumption 1. We replace Proposition 2.10 by:

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

For small enough δ, Equation (5) implies ϵ-robustness at x. However, contrary to Proposition 2.10, Equation (5) does not offer any guarantee on the ϵ-robustness at x for arbitrary δ.

Proposition 2.13. *Equation (5) is equivalent to:*

$$\forall v \in T_x\mathbb{R}^d, \quad \tilde{g}_x(v,v) \leq \frac{\delta^2}{\epsilon^2} \overline{g}_x(v,v). \tag{6}$$

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.

2.4 Coordinate change

In this section, we show how to compute the matrix P introduced for Corollary 3.6. 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 $\mu\left(\Delta^m\right)$ is the subset of $\mathcal{S}^m(2)$ where all coordinates are strictly positive (using the standard coordinates of $\mathbb{R}^c$).

Proposition 2.14. *Let g be the Fisher information metric on Δ^m (Definition 2.2), 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 2.15. *In the coordinates τ, the FIM is:*

$$G_{\tau,ij} = \frac{4}{\left(1 + \|\tau/2\|^2\right)^2} \delta_{ij}. \tag{7}$$

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

$$G = \widetilde{J}^T G_\tau \widetilde{J} = \frac{4}{\left(1 + \|\tau/2\|^2\right)^2} \widetilde{J}^T \widetilde{J}. \tag{8}$$

Thus, we can choose:

$$P = \frac{2}{1 + \|\tau/2\|^2} \widetilde{J}. \tag{9}$$

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

$$P_{ij} = \frac{\delta_{ij}}{\sqrt{\theta^i}} - \frac{\tau^i(\theta)}{2\sqrt{\theta^c}}. \tag{10}$$

2.5 The Fisher-Rao distance

As stated in Proposition 2.14, 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 terms, we replace δ by the following upper bound:

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

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

$$\delta \leq 2 \arccos \sum_{i=1}^{c} \sqrt{\frac{f(x)^i}{c}}. \tag{11}$$

3 Derivations of robustness conditions

3.1 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 3.1.

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

Fact 3.2. $J^T G J$ 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 $T_x\mathbb{R}^d$, denoted $\mathcal{B} = (e_1, \ldots, e_m, e_{m+1}, \ldots, e_d)$, 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 a m-dimensional subspace of $T_x\mathbb{R}^d$. $\tilde{g}_x$ does not define an inner product[1] on $T_x\mathbb{R}^d$ 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 3.3. 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 3.4. 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 Equation (5) with:

$$\overline{\mathcal{B}}_D(0, \epsilon) = \widetilde{\mathcal{B}}_D(0, \delta). \tag{12}$$

Equation 12 is the limit case of Equation 5, in the sense that if Equation 12 holds, then $\widetilde{\mathcal{B}}_x(0, \delta)$ is the smallest possible $\tilde{g}_x$-ball (for the inclusion) such that Equation 5 holds.

Proposition 3.5. *Equation (12) is equivalent to:*

$$\forall v \in D, \quad \tilde{g}_x(v, v) = \frac{\delta^2}{\epsilon^2} \overline{g}_x(v, v). \tag{13}$$

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

We can rewrite Equation (13) in a matrix form:

$$\forall v \in D, \quad v^T \widetilde{G} v = \frac{\delta^2}{\epsilon^2} v^T v. \tag{14}$$

In section 2.4, 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 that is equivalent with Equation 14.

Proposition 3.6. *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 $m \times m$.

Finally, we can define a regularization term:

$$\alpha_1(x, \epsilon, f) = \frac{1}{m^2} \left\| P J J^T P^T - \frac{\delta^2}{\epsilon^2} I_m \right\|, \tag{15}$$

where $\|\!| \cdot |\!\|$ is any matrix norm, such as the Frobenius norm or the spectral norm. The loss function is:

$$L(y, x, \epsilon, f) = l(y, f(x)) + \lambda\, \alpha(x, \epsilon, f), \tag{16}$$

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

3.2 Randomized isometry condition

For any function $h : \mathbb{R}^d \to \mathbb{R}^m$ and any vector $v \in \mathbb{R}^m$, let $\langle h, v \rangle : \mathbb{R}^d \to \mathbb{R}$ be the function defined by $\langle h, v \rangle(x) = h(x)^T v$.

The backpropagation algorithm is applied by computing the gradient $\partial_x \langle f, v \rangle$ where v is a vector of $\mathbb{R}^m$. Thus, in order to compute the Jacobian matrix J, we need to call the backpropagation algorithm m times, by computing each row $\partial_x \langle f, e_i \rangle$, where $e_i = (\delta_{ij})_{j=1,\dots,m}$ is the canonical basis of $\mathbb{R}^m$. To reduce the computational cost, we propose a randomized version of the isometry condition. Let u and v be two vectors sampled uniformly on the sphere of radius 1 in $\mathbb{R}^m$. Write $f(x) = \left(f^1(x), \dots, f^m(x), f^c(x)\right)$. Consider the bilinear form[2]:

$$(u, v) \mapsto F(u, v) = \left(1 - \sqrt{f^c(x)}\right)^2 \left(\partial_x \langle \tau \circ f, u \rangle\right)^T \partial_x \langle \tau \circ f, v \rangle - \frac{\delta^2}{\epsilon^2} u^T v. \tag{17}$$

Then, we can use the following regularization term:

$$\alpha_2(x, \epsilon, f) = \frac{1}{4} \left| F(u, u) \right| + \left| F(v, v) \right| + 2 \left| F(u, v) \right|. \tag{18}$$

With this regularization term, the backpropagation algorithm is only called two times: to compute $\partial_x \langle \tau \circ f, u \rangle$ and $\partial_x \langle \tau \circ f, v \rangle$.

3.3 Bound condition

In this subsection, our goal is to derive a looser condition by relaxing Assumption 3 and working directly from Assumption 1. Let us use the same notation's simplifications introduced at the beginning of section 3.1. We have that Proposition 2.13 is equivalent to:

$$\widetilde{G} \preceq \frac{\delta^2}{\epsilon^2} I_d, \tag{19}$$

[2]The factor $\left(1 - \sqrt{f^c(x)}\right)^2$ corresponds to the term $\frac{4}{(1 + \|\tau/2\|^2)^2}$ in Equation 8. See the proof of Proposition 2.16.

which means that $\widetilde{G} - \frac{\delta^2}{\epsilon^2} I_d$ must have nonpositive eigenvalues. Let $P \in \mathrm{GL}_m(\mathbb{R})$ be as in Proposition 2.16. According to Equation 31, $\widetilde{G} = J^T P^T P J = U P J J^T P^T U$ where U is some matrix that is orthogonal on D. Thus, $\widetilde{G}$ and PJJ^TP^T have the same nonzero eigenvalues. Thus, Equation 19 is equivalent to:

$$\left(PJJ^TP^T - \frac{\delta^2}{\epsilon^2} I_m \right)_+ = 0. \tag{20}$$

Using the bilinear form defined in Equation 17, we have:

$$\left(PJJ^TP^T - \frac{\delta^2}{\epsilon^2} I_m \right)_+ = \left(\max_{u:\|u\|=1} F(u,u) \right)_+, \tag{21}$$

In order to avoid computing $\max_{u:\|u\|=1} F(u,u)$, we use an approximate condition similar to Equation 18:

$$\alpha_3(x, \epsilon, f) = \frac{1}{4} \left\{ (F(u,u))_+ + (F(v,v))_+ + 2\,(F(u,v))_+ \right\}. \tag{22}$$

3.4 Adaptive softmax temperature

The matrix PJJ^TP^T is positive semi-definite. Thus, Equation 20 is equivalent to:

$$\begin{aligned} \left(PJJ^TP^T - \frac{\delta^2}{\epsilon^2} I_m \right)_+ = 0 &\Leftrightarrow \lambda_{max}\left(PJJ^TP^T\right) \leq \frac{\delta^2}{\epsilon^2}, \\ &\Leftrightarrow \|PJ\|_2 \leq \frac{\delta}{\epsilon}, \end{aligned}$$

where $\lambda_{max}()$ gives the largest eigenvalue, and $\|\cdot\|_2$ is the spectral norm. Define $k(x)$ by:

$$k(x) = \frac{\delta}{\epsilon \|PJ\|_2}. \tag{23}$$

Note that the factor $k(x)$ is adapted to the coordinates τ. However, we need to express the output in the coordinates θ in order to train the model (using the cross-entropy). Thus, we define the following function $\tilde{f}$:

$$x \mapsto \tilde{f}(x) = \tau^{-1}\left(k(x)\tau\left(f(x)\right)\right), \tag{24}$$

where the output is converted back to the coordinates θ using τ^{-1} (given explicitly in Equations 27 and 28).

4 Experiments

4.1 CIFAR-10

4.1.1 Set-up

We evaluate several defense methods on CIFAR-10 with a ResNet18 initialized with parameters pre-trained on ImageNet. The optimization algorithm is Stochastic Gradient Descent with learning rate of 10^{-3}, batch size of 32, and gradient clipping. The images are normalized in $[0, 1]$ and we use data augmentation with random crop and random flip.

A first model is trained without any defense. We use this model to adversarially perturb the test set with FGSM, using an attack budget of 8/255. This adversarially perturb test set is then used to evaluate all defense methods in the subsequent experiments. We compare our proposed methods with defensive distillation [8], Jacobian regularization [9], adversarial training with FGSM [3], and Gaussian augmentation [10]. All results are averaged over five runs with different random seeds.

4.1.2 Robustness against FGSM attack

4.1.3 Robustness against PGD attack

4.1.4 Robustness against AutoAttack

4.2 Fashion-MNIST

4.3 ImageNette

5 Discussion and related work

In 2019, Zhao *et al.* [11] 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.* [12] 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 [13]. In our framework, their method consists in expanding the geodesic ball $\widetilde{\mathcal{B}}(x, \delta)$ 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.

Cisse *et al.* [14] 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 isometry condition, albeit with completely different rationales. This suggests that geometric reasoning could successfully supplement the line of work on Lipschitz constants of neural networks.

Following another line of work, Hoffman *et al.* [9] advanced a Jacobian regularization to improve adversarial robustness. Their regularization consists in 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.* [12] 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 [15].

Our proposed methods 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 [16].

6 Conclusion

7 Proofs

Proof of Proposition 2.13. (6) $\Rightarrow$ (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) $\Rightarrow$ (6]). Assume (5). Let $v \in T_x\mathbb{R}^d$. 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 2.14. We need to show that $\mu^*\overline{g} = g$. Using the coordinates θ on Δ^m (Definition 2.1) and the standard coordinates on $\mathbb{R}^c$, and writing $f(x) = \theta_0 = (\theta_0^1, \ldots, \theta_0^m)$ we have:

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

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 2.2. □

Proof of Proposition 2.15. For $i = 1, \ldots, m$, the inverse transformation of $\tau(\mu)$ is (proof below):

$$\mu^i(\tau) = \frac{2\tau^i}{1 + \|\tau/2\|^2}, \tag{25}$$

and:

$$\mu^c(\tau) = 2\frac{\|\tau/2\|^2 - 1}{\|\tau/2\|^2 + 1}. \tag{26}$$

Moreover, according to Proposition 2.14, 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:

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

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

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

and for $\alpha = c$:

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

Thus:

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

□

Proof of Equations 25 and 26. 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:

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

Isolating μ^c, we get:

$$\mu^c(\tau) = \frac{2\|\tau\|^2 - 8}{\|\tau\|^2 + 4} = 2\frac{\|\tau/2\|^2 - 1}{\|\tau/2\|^2 + 1}. \tag{27}$$

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

$$\mu^i(\tau) = \frac{\tau^i}{\lambda} = \frac{2\tau^i}{1 + \|\tau/2\|^2} \tag{28}$$

□

Proof of Proposition 2.16. We have $\tau^i(\theta) = 2\sqrt{\theta^i}/\left(1 - \sqrt{\theta^c}\right)$. 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 $i = 1, \ldots, m$:

$$\frac{2}{1 + \|\tau(\theta)/2\|^2} = 1 - \sqrt{\theta^c} = \frac{2\sqrt{\theta^i}}{\tau^i(\theta)}. \tag{29}$$

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

$$\frac{\partial \tau^i(\theta)}{\partial \theta^j} = \frac{\delta_{ij}}{\sqrt{\theta^i}\left(1 - \sqrt{\theta^c}\right)} - \frac{\sqrt{\theta^i}}{\sqrt{\theta^c}\left(1 - \sqrt{\theta^c}\right)^2} = \frac{\tau^i(\theta)}{2}\left(\frac{\delta_{ij}}{\theta^i} - \frac{\tau^i(\theta)}{2\sqrt{\theta^i\theta^c}}\right) \tag{30}$$

Replacing Equations 29 and 30 into Equation 9 yields the result. □

Proof of Fact 3.1. 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 3.5. The implication (13) ⇒ (12]) is immediate (by double inclusion).
Now, assume (12]) holds. Let $v \in D$. Define $w_1 = \epsilon\, v/\sqrt{\overline{g}_x(v,v)}$ and $w_2 = \epsilon\, v/\sqrt{\tilde{g}_x(v,v)}$. Then, with a similar argument as in the proof of Proposition 2.13, we can obtain Equation (13). Note that w_2 is well defined because $v \notin \ker(J)$. □

Proof of Proposition 3.6. **Let us first introduce the polar decomposition**.
Let A be a $m \times d$ matrix.
Define the absolute value[3] of A by $|A| = (A^T A)^{\frac{1}{2}}$.
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 $\|Ax\|^2 = x^T A^T A x = (A^T A x)^T x = (|A|^2 x)^T x = x^T |A|^T |A| x = \||A|x\|^2$, thus u is an isometry[4].
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

$$PJ = U|PJ|,$$

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:

$$PJJ^T P^T = U|PJ|^2 U^T = UJ^T P^T PJ U^T \tag{31}$$

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

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

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 $J^T G J = \Pi_D$ 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:

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

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). □

References

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

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

[3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations*, 2018.

[3] The square root of $A^T A$ is well defined because it is a positive semidefinite matrix.
[4] We can arbitrarily extend u on the entire $\mathbb{R}^d$, e.g., by setting $\ker(u) = \ker(|A|)$.

[4] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," in *IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.

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

[6] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *37th International Conference on Machine Learning (ICML)*, 2020.

[7] O. Calin and C. Udrişte, *Geometric Modeling in Probability and Statistics*. Springer International Publishing, 2014.

[8] N. Papernot, P. D. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks.," *CoRR*, 2015.

[9] J. Hoffman, D. A. Roberts, and S. Yaida, "Robust Learning with Jacobian Regularization," 2019.

[10] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified Adversarial Robustness via Randomized Smoothing," in *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[11] C. Zhao, P. T. Fletcher, M. Yu, Y. Peng, G. Zhang, and C. Shen, "The Adversarial Attack and Detection under the Fisher Information Metric," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[12] C. Shen, Y. Peng, G. Zhang, and J. Fan, "Defending against adversarial attacks by suppressing the largest eigenvalue of fisher information matrix," *ArXiv*, 2019.

[13] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[14] M. Cissé, P. Bojanowski, E. Grave, Y. N. Dauphin, and N. Usunier, "Parseval Networks: Improving Robustness to Adversarial Examples," in *Proceedings of the 34th International Conference on Machine Learning*, pp. 854–863, PMLR, 2017.

[15] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," in *International Conference on Learning Representations*, 2018.

[16] L. T. Skovgaard, "A Riemannian Geometry of the Multivariate Normal Model," *Scandinavian Journal of Statistics*, vol. 11, no. 4, pp. 211–223, 1984.