Adversarial robustness with partial isometry

1 Introduction

For the last few years, the machine learning community have started to study the robustness problems of machine learning models, and of neural networks in particular. This study was motivated by the high sensitivity of neural networks to adversarial attacks, i.e., small perturbations that are able to fool a network. Adversarial attacks has been shown to be both ubiquitous and transferable. Beyond the obvious security threat, adversarial attacks are the symptom of the dramatic lack of robustness of neural networks. We believe that understanding and mitigating adversarial attacks will solve a large portion of the robustness issue of neural networks. Nonetheless, it must be pointed out that adversarial attacks, which belongs the larger category of evasion attacks, are not the only robustness vulnerability of neural networks. Let us mention sensitivity to noise, ability to detect out-of-distribution data, quantification of the uncertainty in order to avoid model overconfidence, data poisoning, byzantine attacks, and model extraction, to name just a few.

In this document, a regularization method for adversarial robustness is presented. This method consists in encouraging the model to be isometric in each leaf of the data foliation (defined in section 2). The focus of this method is l_2 white-box attacks against multiclass classification tasks but it may be extended to more general settings (unrestricted attacks, black-box attacks etc.). In section 5, we discuss how this method can be extended to other supervised learning tasks. The method is evaluated on several image datasets, and using several state-of-the-art attacks. We pay a special attention to the computational efficiency of the method since we hope that it could be used in real-world applications.

The goal of this approach is not only to develop a new adversarial defense, but also to better understand the adversarial vulnerability phenomenon in deep learning models. Two questions are of particular interest: is there really a trade-off between accuracy and robustness? And is it possible to provide a certified defense in terms of both accuracy and robustness? This is a required step for the deployment of deep learning models in critical domains where certification is mandatory.

The remaining of this document is divided into four parts. Section 2 introduces the notations and describes the information geometric framework for the study of adversarial robustness. Section 3 presents the partial isometry regularization method in each leaf of the data foliation. Section 4 presents several experiments to evaluate the proposed method. Finally, section 5 suggests some directions to answer the questions asked in introduction and discusses potential extensions of this work.

2 Information geometric framework

2.1 Multiclass classification and family of categorical distributions

Let $\mathcal{Y} = \{1, \dots, m+1\} \subset \mathbb{N}$ be the set of labels for the classification task (hence there are $m+1$ different classes). Let $\mathcal{X} \subseteq \mathbb{R}^n$ be the **input domain**. We assume than $n > m$. For example, in MNIST we have $\mathcal{X} = [0,1]^n$ (with $n = 784$) since an admissible image must have its pixel's values between 0 and 1, and $m = 9$. In the following, "smooth" means C^∞. We assume that $\mathcal{X}$ is an n-dimensional embedded smooth connected submanifold of $\mathbb{R}^n$. For simplicity, we can think that $\mathcal{X} = \mathbb{R}^n$. 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 document, we see a model as assigning the *parameters* of a random variable Y to a given input $x \in \mathcal{X}$. Let us formalize this. We ask that the random variable Y verifies the following assumptions:

1. Y takes its values in $\mathbb{R}$ and its support is $\mathcal{Y}$.

2. Y is absolutely continuous with respect to the counting measure on $\mathbb{R}$. In particular, Y has a probability density function (pdf) with respect to the counting measure.

3. The pdf of Y belongs to a **parameterized family of pdf** $\mathcal{F} = \{p_\theta : \theta \in \Delta^m\}$ (more details below).

We will denote the components of a vector v as v^i with a superscript. The set Δ^m is defined as $\Delta^m = \{\theta \in \mathbb{R}^{m+1} : \sum_{i=1}^{m+1} \theta^i = 1, 0 < \theta^i < 1\}$. It is called the **probability m-simplex**. The inequality $0 < \theta^i < 1$ is strict in order to discard the boundary of the simplex, because the Fisher information metric introduced in the next subsection is not defined on the boundary of Δ^m.

Let $\theta \in \Delta^m$. The pdf $p_\theta : \mathbb{R} \to \mathbb{R}$ is defined as:

$$p_\theta(y) = \prod_{i=1}^{m+1} (\theta^i)^{\delta_i(y)} = \exp\left(\sum_{i=1}^{m+1} \delta_i(y) \ln \theta^i\right),$$

where $\delta_i(y) = 1$ if $y = i$ and 0 otherwise. Thus, the family $\mathcal{F}$ is the **family of $(m+1)$-dimensional categorical distributions**. $\mathcal{F}$ can be endowed with a differentiable structure by using $p_\theta \in \mathcal{F} \mapsto (\theta^1, \dots, \theta^m) \in \mathbb{R}^m$ as a global coordinate system. Hence, $\mathcal{F}$ becomes a smooth manifold of dimension m (more details on this construction can be found in Amari 1985 [1] Chapter 2). We can identify p_θ with $(\theta^1, \dots, \theta^m)$.

In the following, we will note $p_\theta(y) = p(y, \theta)$ where $p : \mathbb{R} \times \Delta^m \to \mathbb{R}$ is seen as a function of both y and θ. Note that for any $y \in \mathbb{R}$, the function $p(y, \cdot) : \Delta^m \to \mathbb{R}$ is smooth. We will often use the abuse of notation that consists in writing "the distribution $p(y, \theta)$" instead of "the distribution of Y characterized by the pdf $y \mapsto p(y, \theta)$".

Finally, we can define what we mean by "model". We call "**model**" any smooth map $F : \mathcal{X} \to \Delta^m$, that assigns to an input $x \in X$ the parameters $\theta = F(x) \in \Delta^m$ of a $(m+1)$-dimensional categorical distribution $p(y, \theta) \in \mathcal{F}$. 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: $\theta = \text{softmax}(s(x))$.

2.2 Riemannian metrics

Let F be a model. 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. Formally, a Riemannian metric is a covariant tensor field of rank 2 which is positive-definite. Intuitively, a Riemannian metric defines an inner product of on each tangent space, and this inner product changes smoothly when we move smoothly from one tangent space to another. With this inner product, we can compute angles and norms of tangent vectors, which can be seen as "infinitesimal" vectors. By integrating these "infinitesimal" lengths along a curve in the manifold, we can compute the length of the curve. The distance between two points can thus be defined as the infimum of the length of all the curves joining the two points.

Let us begin with $\mathcal{X}$. 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**. Using the standard coordinate of $\mathbb{R}^n$ as a global coordinate system for $\mathcal{X}$, the Euclidean metric is defined as

$$\overline{g}_x = \sum_{i,j=1}^{n} \delta_{ij} dx^i dx^j,$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. The Euclidean metric corresponds to the l_2 norm. Please note that there are other "dissimilarity" 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 on 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 a l_∞ budget ϵ_∞, we can choose a 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. It may be possible to find a coordinate change on $\mathcal{X}$ such that these measures correspond to the l_2 norm in the new coordinate system,

or to use another metric. Anyway, in the rest of this document, we assume that $(\mathcal{X}, \overline{g})$ is a Riemannian manifold equipped with the Euclidean metric.

Now, we consider Δ^m. The manifold Δ^m can also be equipped with the Euclidean metric induced from $\mathbb{R}^{m+1}$. However, as described above, we see Δ^m as the family of categorical distributions. In order to find a more natural Riemannian metric for Δ^m, i.e., a metric that reflects the statistical properties of Δ^m, let us consider the loss function used to train a multiclass classification model: the **relative entropy**, also known as Kullback-Leibler divergence. Let θ and ϕ be two points of Δ^m. Then, the relative entropy is:

$$\begin{aligned} H\left(\phi||\theta\right) &= -\mathbb{E}_{\phi}\left[\ln\frac{\theta}{\phi}\right], \\ &= \sum_{i=1}^{m+1}\phi^i \ln\frac{\phi^i}{\theta^i}. \end{aligned}$$

When training a model, $\theta = F(x)$ is the distribution predicted by the model, while ϕ is the true distribution. The relative entropy can be rewritten $H(\phi||\theta) = -\sum_{i=1}^{m+1}\phi^i \ln\theta^i + \sum_{i=1}^{m+1}\phi^i \ln\phi^i = H(\phi,\theta) - H(\phi)$ where $H(\phi,\theta)$ is the cross entropy and $H(\phi)$ is the entropy of ϕ. In practice, only the cross entropy term is used as a loss function since $H(\phi)$ does not depend on θ.

More generally, we can define a large family of divergences by:

$$H_{r,f}\left(\phi||\theta\right) = \frac{1}{r-1} f\left(\sum_{i=1}^{m+1}(\phi^i)^r(\theta^i)^{1-r}\right),$$

where $r \geq 0$, $r \neq 1$, and f is a C^1 function such that $f(1) = 0$ and $f'(1) = 1$. If we choose $f = \ln$, we obtain the Rényi divergences. Then, it can be shown that:

$$H_{r,\ln}\left(\phi||\theta\right) \xrightarrow[r\to 1]{} H\left(\phi||\theta\right).$$

The second-order Taylor approximation of $H_{r,f}$ gives:

$$H_{r,f}\left(\theta||\theta+\delta\theta\right) = \frac{r}{2}\sum_{i=1}^{m+1}\frac{1}{\theta^i}\left(\delta\theta^i\right)^2 + o\left(\|\delta\theta\|^2\right). \tag{1}$$

Surprisingly, equation 1 does not depend on f. When $\delta\theta$ is small, $H_{r,f}\left(\theta||\theta+\delta\theta\right)$ looks almost like a *squared* Euclidean distance but with an additional coefficient $\frac{r}{2\phi^i}$. Let $T_\theta\Delta^m$ be the tangent space of Δ^m at θ. Let $X, Y \in T_\theta\Delta^m$ be two tangent vectors at θ. Equation 1 suggests that $H_{r,f}$ is an infinitesimal squared distance. If we discard the multiplicative constant $\frac{r}{2}$, the corresponding norm is:

$$\|X\|^2 = \sum_{i=1}^{m+1}\frac{1}{\theta^i}X^i,$$

and the inner product is:

$$\langle X, Y\rangle = \sum_{i=1}^{m+1}\frac{1}{\theta^i}X^iY^i.$$

We have thus defined a Riemannian metric on Δ^m:

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

The metric g is called the **Fisher information metric** (FIM). Indeed, if we consider the Fisher information defined from the Cramér-Rao bound:

$$\mathcal{I}(\theta) = \mathbb{E}_{y\sim p_\theta}[\nabla_\theta \ln p_\theta(y) \nabla_\theta \ln p_\theta(y)^T],$$

then an easy calculation shows that $\mathcal{I}(\theta) = g_\theta^{FIM}$.

We have already seen that the FIM is the infinitesimal metric associated to a large family of divergences. Another remarkable property of the FIM is Chentsov's theorem that claims that the FIM is the *unique* Riemannian metric on Δ^m that is invariant under sufficient statistics (up to a multiplicative constant). To summarize, the FIM is invariant when we deform the relative entropy, and it is invariant when we deform $\mathcal{Y}$ with a sufficient statistic. Moreover, the FIM tells us how our loss function changes when we move in Δ^m. For all those reasons, we endow Δ^m with the FIM g^{FIM}, making it a Riemannian manifold (Δ^m, g^{FIM}).

The reader may have noticed that we have defined the FIM using the parameter $\theta \in \mathbb{R}^{m+1}$. However, we also stated that Δ^m is a m-dimensional smooth manifold. This means that $\theta \in \mathbb{R}^{m+1}$ is not a coordinate system on Δ^m, while $(\theta^1, \dots, \theta^m) \in \mathbb{R}^m$ is such a coordinate system. Hence, we have defined the FIM on $\mathbb{R}^{m+1}$ by seeing Δ^m as an embedded submanifold of $\mathbb{R}^{m+1}$. In order to see Δ^m as a manifold in its own right, we must defined the FIM with respect to a coordinate system. Let g be the FIM expressed in the coordinate system $(\theta^1, \dots, \theta^m)$. In these coordinates, define the inclusion function $i : \Delta^m \to \mathbb{R}^{m+1}$ by $i(\theta^1, \dots, \theta^m) = (\theta^1, \dots, \theta^m, 1 - \sum_{k=1}^m \theta^k)$. Then, g is the pullback of g^{FIM} by the inclusion function i, i.e., $g = i^* g^{FIM}$. It can easily be shown that:

$$g_\theta = \sum_{i,j=1}^m \left(\frac{1}{\theta^i} \delta_{ij} + \frac{1}{1 - \sum_{k=1}^m \theta^k} \right) d\theta^i d\theta^j.$$

Let $\theta^{m+1} = 1 - \sum_{k=1}^m \theta^k$. Then:

$$g_\theta = \sum_{i,j=1}^m \left(\frac{1}{\theta^i} \delta_{ij} + \frac{1}{\theta^{m+1}} \right) d\theta^i d\theta^j. \tag{2}$$

In the rest of this document, we use the FIM on Δ^m as defined in equation 2.

2.3 Data and kernel foliations

Consider the smooth manifold $\mathcal{X}$. Let $x \in \mathcal{X}$ and let $T_x\mathcal{X}$ be the tangent space of $\mathcal{X}$ at x. Since $\mathcal{X}$ is a n-dimensional embedded submanifold of $\mathbb{R}^n$, $T_x\mathcal{X}$ can be identified with the vector space $\mathbb{R}^n$. A **distribution**[1] D on $\mathcal{X}$ is the assignment to every $x \in \mathcal{X}$ of a vector subspace D_x of $T_x\mathcal{X} \approx \mathbb{R}^n$. We say that the distribution D has constant rank if $\dim D_x = \dim D_{x'}$ for every $x, x' \in \mathcal{X}$.

A subset L of $\mathcal{X}$ is said to be a **regular submanifold** of dimension d if, for any point $x \in L$, there exists a local coordinate chart $(U, r^1, \dots, r^n)$ with $x \in U$ such that L is locally defined by $r^{d+1} = c^{d+1}, \dots, r^n = c^n$ for some constants $c^{d+1}, \dots, c^n$.

A constant rank distribution D is said to be **completely integrable** if, for any point $x \in \mathcal{X}$, there exists a regular submanifold L such that $T_x L = D_x$, where the tangent space $T_x L$ is seen as a vector subspace of $T_x\mathcal{X}$. We say that L is an integral submanifold of D. The set $\mathfrak{F}$ of all maximal integral submanifolds of D is called a **foliation** of $\mathcal{X}$. The submanifolds of the foliation $\mathfrak{F}$ are called the **leaves** of the foliation. The leaves of $\mathfrak{F}$ form a partition of $\mathcal{X}$. The Frobenius theorem gives a necessary and sufficient condition for a constant rank distribution D to be completely integrable: D is completely integrable if and only if it is involutive, i.e., if and only if $[X, Y] = XY - YX \in D$ for any smooth vector fields X, Y in D.

For example, assume that $\mathcal{X} = \mathbb{R}^2 \setminus \{(0,0)\}$ and let (x^1, x^2) be the standard coordinates. Consider the distribution $x \mapsto D_x = \text{Span}\{-x^2 \partial x^1 + x^1 \partial x^2\}$. Let $x_0 \in \mathcal{X}$ and $\|x_0\|^2 = (x_0^1)^2 + (x_0^2)^2$. Consider the regular submanifold L defined in polar coordinates (r, θ) by $r = \|x_0\|$. L is simply the circle centered at the origin with radius $\|x_0\|$. We can see that $T_{x_0} L = D_{x_0}$, thus the distribution D is completely integrable. The set of all circles centered at the origin is a foliation of $\mathcal{X}$.

Now, we go back to the general case where $\mathcal{X}$ is a submanifold of $\mathbb{R}^n$. Let $F : \mathcal{X} \to \Delta^m$ be a model. Consider the pullback metric $\tilde{g}$ of the FIM by F, i.e. $\tilde{g} = F^* g$. Using the standard coordinates on $\mathcal{X}$ and the coordinates $(\theta^1, \dots, \theta^m)$ on Δ^m, we have for any $x \in \mathcal{X}$:

$$\tilde{g}_x = J_x^T g_{F(x)} J_x, \tag{3}$$

[1] Be careful to not confuse a "distribution" in this geometrical sense with a "probability distribution" as used in susbsection 2.1.

where J_x is the Jacobian matrix of F at x. Since $n > m$, $\tilde{g}$ cannot be positive definite. It is only positive semi-definite, i.e., its kernel is non-trivial. Hence we can define a distribution K on $\mathcal{X}$ by

$$x \mapsto K_x = \ker \tilde{g}_x.$$

We also consider the distribution D that is orthogonal to K in $T\mathcal{X}$ (using the Euclidean metric $\overline{g}$):

$$x \mapsto D_x = (\ker \tilde{g}_x)^{\perp}.$$

Grementieri & Fioresi [2] showed that if F is a neural network using piecewise linear activations (such as ReLU), and if D and K have constant rank, then both D and K are completely integrable. The foliation $\mathfrak{D}$ associated to D is called the **data foliation**, and the foliation $\mathfrak{K}$ associated to K is called the **kernel foliation**. Moreover, Grementieri & Fioresi [2] showed that the rank of D is less or equal to m and that

$$D_x = \mathrm{Span}\{\nabla_x F^1(x), \dots, \nabla_x F^{m+1}(x)\}, \tag{4}$$

where the F^is are the components of F.

3 Partial isometry regularization

3.1 A necessary condition for adversarial robustness at a point

Let $F : \mathcal{X} \to \Delta^m$ be a model. Let $x \in \mathcal{X}$ and consider the Euclidean open ball $B(x, \epsilon_2)$ in $\mathcal{X}$ centered at x with radius $\epsilon_2 > 0$. The radius ϵ_2 is chosen such that all points in $B(x, \epsilon_2)$ are considered to be indistinguishable from x (according to some relevant perspective), while every point outside $B(x, \epsilon_2)$ can be distinguished from x.

In order to ensure adversarial robustness at x, it is sufficient that every point in $B(x, \epsilon_2)$ have the same class, i.e., $B(x, \epsilon_2)$ does not intersect any decision boundary. This condition must be strongly enforced when the model is very confident in classifying x, i.e., when $F(x)$ is close to a vertex of Δ^m. However, when F is unsure how to classify x (e.g., when x is an edge case, or an out-of-distribution example), then this condition may be relaxed since the model could indicate that it is very uncertain about its prediction.

The image by F in Δ^m of the decision boundaries is the set $\mathcal{B} = \{\theta \in \Delta^m : \theta^i = \theta^j \text{ for some } i \neq j\}$. For $\theta \in \Delta^m$, let $d(\theta, \mathcal{B})$ be the distance between θ and $\mathcal{B}$ (here, we are using the distance induced by the FIM g). It can be shown that:

$$\sup_{\theta \in \Delta^m} d(\theta, \mathcal{B}) = 2 \arccos \frac{1}{\sqrt{m+1}}. \tag{5}$$

We write $\delta_m = 2 \arccos \frac{1}{\sqrt{m+1}}$. Let $\delta(F, x, \epsilon_2)$ be the diameter[2] of $F(B(x, \epsilon_2))$ in Δ^m. A necessary condition to ensure adversarial robustness at x is:

$$\delta(F, x, \epsilon_2) \leq \delta_m. \tag{6}$$

3.2 Restricting to a leaf of the data foliation

Let L be the leaf of the data foliation $\mathfrak{D}$ that contains x. The pullback of the FIM $\tilde{g}$ is not a Riemannian metric on $\mathcal{X}$ because it is not positive definite. However, the restriction $\tilde{g}|_L$ of $\tilde{g}$ to L is a Riemannian metric on L. In this section, we restrict our study to $B_L(x, \epsilon_2)$, the open ball in L centered at x of radius ϵ_2, defined using the metric $\overline{g}|_L$.

Let $\exp : U \subset T_x L \to L$ be the exponential map induced by the pullback of the FIM $\tilde{g}|_L$, restricted to a subset U. The subset U is chosen such that $\exp(U) = B_L(x, \epsilon_2)$. Let $X \in U$ be a tangent vector. The point $\exp(X) \in L$ is defined as the point reached by following the geodesic starting at x with initial velocity X during a time interval of 1. The distance between $F(x)$ and $F(\exp(X))$ in Δ^m is $\sqrt{(\tilde{g}|_L)_x(X, X)}$. This distance is maximized when X is an eigenvector of $(\tilde{g}|_L)_x$ associated with the highest eigenvalue. Hence, it is reasonable to regularize the model F in order to reduce the highest eigenvalue of $(\tilde{g}|_L)_x$. This approach is developed by Shen et al. [3].

[2]The diameter of a set is the supremum of the distances between two points of this set.

In this section, we aim to enforce the following stronger condition:

$$(\tilde{g}|_L)_x = \frac{\delta_m}{\epsilon}(\overline{g}|_L)_x. \tag{7}$$

The value of ϵ should be chosen such that $\epsilon \geq \epsilon_2$. If this condition is enforced and if ϵ is small enough, then we have $\delta_L(F, x, \epsilon_2) \leq \delta_m$ where $\delta_L(F, x, \epsilon_2)$ is the diameter of $F(B_L(x, \epsilon_2))$ in Δ^m. This condition consists in choosing a model F such that $F|_L$: $(L, \frac{\delta_m}{\epsilon}\overline{g}|_L) \to (\Delta^m, g)$ is a local isometry at x.

3.3 Other conditions equivalent to the partial isometry condition

Let us assume that L has dimension m. Let $v_1, \ldots, v_m$ be a basis of D_x. Let $\tilde{G}_x$ be the matrix of $\tilde{g}_x$ in the standard coordinates of $\mathbb{R}^n$. Equation 7 translates into:

$$v_i^T \tilde{G}_x v_j = \frac{\delta_m}{\epsilon} v_i^T v_j, \text{ for all } 1 \leq i, j \leq m, \tag{8}$$

because the matrix of $\overline{g}$ in the standard coordinates is the identity. Our goal is to design a regularization term to enforce equation 8 in a computationally efficient way. Since L has dimension m, then $D_x = (\ker \tilde{G}_x)^\perp$ has also dimension m. Hence, $\tilde{G}_x$ has m strictly positive eigenvalues $\lambda_1, \ldots, \lambda_m$, and $n - m$ eigenvalues equal to zero. It is easy to see that equation 8 becomes:

$$\lambda_i = \frac{\delta_m}{\epsilon}, \text{ for all } 1 \leq i \leq m. \tag{9}$$

Since $\tilde{G}_x$ is symmetric, its eigenvectors associated to different eigenvalues are orthogonal. Hence, all eigenvectors associated to non-zero eigenvalues are in $D_x = (\ker \tilde{G}_x)^\perp$. Let $(e_1, \ldots, e_m)$ be an orthonormal basis of D_x. Let W be the $m \times n$ matrix whose rows are the e_i's. According to the Rayleigh-Ritz procedure, the eigenvalues of WG_xW^T are exactly $\lambda_1, \ldots, \lambda_m$. Using equation 3, we obtain $W\tilde{G}_xW^T = WJ_x^T G_{F(x)} J_x W^T$, where $G_{F(x)}$ is the matrix of $g_{F(x)}$ in the coordinates[3] $(\theta^1, \ldots, \theta^m)$. Thus, equation 9 translates into

$$WJ_x^T G_\theta J_x W^T = \frac{\delta_m}{\epsilon} I_m, \tag{10}$$

where I_m is the $m \times m$ identity matrix and $\theta = F(x)$. In order to enforce equation 10, we need to find a change of basis matrix P such that $P^T G_\theta P = \frac{\delta_m}{\epsilon} I_m$. Then equation 10 would become $J_x W^T = P$. However, we must compute the eigenvalues and eigenvectors of G_θ in order to compute P, which is not computationally efficient. Instead, we will do several coordinate changes such that the matrix of g in the new coordinates is diagonal for every $\theta \in \Delta^m$.

3.4 Coordinate changes to obtain a diagonal FIM

First, consider the diffeomorphism $T_1 : \Delta^m \to \Delta^m$ defined by:

$$\mu = T_1(\theta) = 2\sqrt{\theta},$$

where the square root is applied element-wise. Let $\mu^{m+1} = 2\sqrt{1 - \frac{1}{4}\sum_{k=1}^m (\mu^i)^2}$. Then, we can compute g is these new coordinates:

$$g = \sum_{i,j=1}^m \left(\delta_{ij} + \frac{\mu^i \mu^j}{(\mu^{m+1})^2}\right) d\mu^i d\mu^j. \tag{11}$$

The expression of g in equation 11 is the same as the expression of the Euclidean metric induced on the sphere of radius 2 in the standard coordinates. In other words, Δ^m is isometric with a portion of the m-sphere of radius 2. As a side note, we can use this result to prove equation 5.

[3]Here, we consider that $\theta = (\theta^1, \ldots, \theta^m) = F(x) \in \mathbb{R}^m$. In other words, we discard θ^{m+1}, the last component of $F(x)$. Moreover, notice that we see F as a map from the *coordinates* $(x^1, \ldots, x^n)$ to the *coordinates* $(\theta^1, \ldots, \theta^m)$ i.e., we do not see F as being defined in a coordinate-independent way.

Now, consider the stereographic projection $T_2 : \Delta^m \to \Delta^m$ defined by:

$$t = T_2(\mu) = \frac{2\mu}{2 - \mu^{m+1}},$$

where the expression is applied element-wise. Let $\|t\|^2 = \sum_{k=1}^{m} (t^k)^2$. Then, the expression of g is these new coordinates is:

$$g = \sum_{i,j=1}^{m} \frac{4\delta_{ij}}{\left(\left\|\frac{t}{2}\right\|^2 + 1\right)^2} dt^i dt^j, \tag{12}$$

which is diagonal. Now, we can easily compute P as:

$$P = \frac{\delta_m \left(\left\|\frac{t}{2}\right\|^2 + 1\right)}{2\epsilon} I_m.$$

Notice that P is a homothety matrix. Let $\tilde{J}_x$ be the Jacobian matrix of $T_2 \circ T_1 \circ F$ at x. Equation 10 becomes $\tilde{J}_x W^T = P$. According to equation 4, the rows of $\tilde{J}_x$ is a basis of D_x. Remember that the rows of W is an *orthonormal* basis of D_x. Since P is diagonal, the condition $\tilde{J}_x W^T = P$ is equivalent to the following conditions:

1. The rows of $\tilde{J}_x$ are orthogonal.
2. $\|\nabla_x T_2(T_1(F^i(x)))\| = \frac{\delta_m \left(\left\|\frac{t}{2}\right\|^2 + 1\right)}{2\epsilon}$ for all $1 \leq i \leq m$.

With these conditions, J is a **semi-orthogonal matrix** multiplied by a homothety matrix, and F is a local **partial isometry**. Equation 10 becomes $\tilde{J}_x \tilde{J}_x^T = \delta_m^2 \left(\|T_2(T_1(F(x)))/2\|^2 + 1\right)^2 / (4\epsilon^2) I_m$. This condition can be rewritten as:

$$\tilde{J}_x \tilde{J}_x^T = \frac{\delta_m^2 F(x)^{m+1}}{\epsilon^2 \left(2\sqrt{F(x)^{m+1}} - \|F(x)\|_1\right)^2} I_m, \tag{13}$$

where $F(x) = (\theta^1, \ldots, \theta^m)$, $\|F(x)\|_1 = \sum_{i=1}^{m} \theta^i$, and $F(x)^{m+1} = 1 - \|F(x)\|_1$.

3.5 A regularization term to enforce the partial isometry condition

Now, we can define a regularization term:

$$\alpha(x, F) = \epsilon^2 \left\| \tilde{J}_x \tilde{J}_x^T - \frac{\delta_m^2 F(x)^{m+1}}{\epsilon^2 \left(2\sqrt{F(x)^{m+1}} - \|F(x)\|_1\right)^2} I_m \right\|_F, \tag{14}$$

where $\|\cdot\|_F$ is the Frobenius norm. We multiplied the regularization term by ϵ^2 in order to compensate the $1/\epsilon^2$. This will facilitate the hyperparameter tuning when using different values of ϵ. To compute $\alpha(x, F)$, we only need to compute the Jacobian matrix $\tilde{J}_x$ which can be efficiently achieved with backpropagation. The regularized loss function is:

$$\mathcal{L}(\hat{\theta}, x, F) = (1 - \eta) H(\hat{\theta} || F(x)) + \eta \alpha(x, F), \tag{15}$$

where η is a hyperparameter to control the trade-off between the relative entropy and the regularization term. In practice, using the loss $\mathcal{L}$ to train a model will approximately enforce the robustness condition 6 at each training point in expectation.

Model \ Budget	0.1	0.2	0.3
Vanilla (9740)	4138 (42%)	428 (4%)	103 (1%)
Regularized (8829)	8040 (91%)	3708 (42%)	211 (2%)

Table 1: Adversarial robustness of the vanilla and regularized models for various attack budgets. The reported values are the number of perturbed images that were **correctly classified** (i.e., same class as the original image). Next to each model's name, we report the total number of correctly classified original images.

4 Experiments

4.1 Toy model

4.2 First results on MNIST

We present preliminary results obtained on MNIST for the regularization method introduced in section 3. We implement a simple LeNet model with two convolutional layers of 32 and 64 channels respectively, followed by one hidden layer with 128 neurons. We train two models, one with the regularization, and one without. Both models are trained with vanilla SGD on 30 epochs, with batch size 32 and learning rate 0.01. For the regularization, we choose ϵ such that $\frac{\delta_m}{\epsilon} = 10^{-3}$. Instead of choosing a fixed η, we increase η at each epoch of the training according to the following rule:

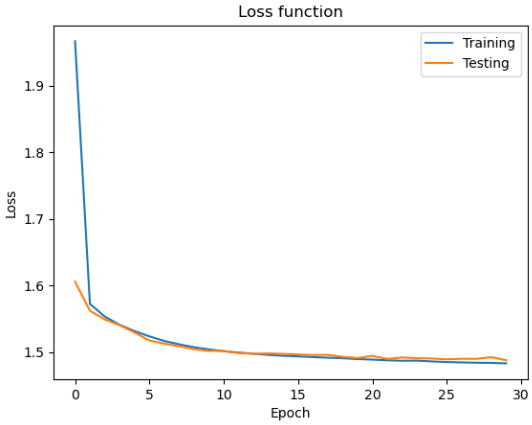
$$\eta_i = \eta_{min} \left(\frac{\eta_{max}}{\eta_{min}} \right)^{(i-1)/(N_{epoch}-1)},$$

where N_{epoch} is the total number of epochs and $1 \leq i \leq N_{epoch}$ is the current epoch. We chose $\eta_{min} = 0.002$ and $\eta_{max} = 0.006$.

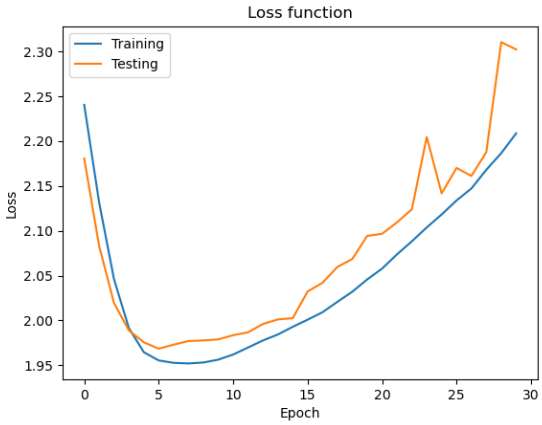
The models are trained on the 60000 images of MNIST's training set, then tested on the 10000 images of the test set. **The vanilla model achieves an accuracy of 97% (9740/10000) while the regularized model achieves an accuracy of 88% (8829/10000)**. It must be mentioned that, with the current implementation, the regularized model is still almost **12 times slower** to train than the vanilla model. However, it may be possible to accelerate the training using for example the technique proposed by Shafahi et al. [4].

Figure 1 presents the loss functions of both models as well as the cross entropy and the regularization term (equation 14) of the regularized model alone. We can see that the total loss function of the regularized model (figure 1b) decreases until epoch 7 and then increases. This is due to the increase of η. Indeed, we see in figure 1d that the regularization of the initial (random) model is very low and then increases dramatically when the model starts to learn. When η increases, the model is forced to reduce the regularization term but seems to struggle at achieve a large decrease. It may achieve a lower regularization with longer training. We also tried to train the model with a fixed $\eta = 0.005$. With this setting, the model immediately minimize the regularization but is unable to reduce the cross entropy and thus achieves low accuracy.

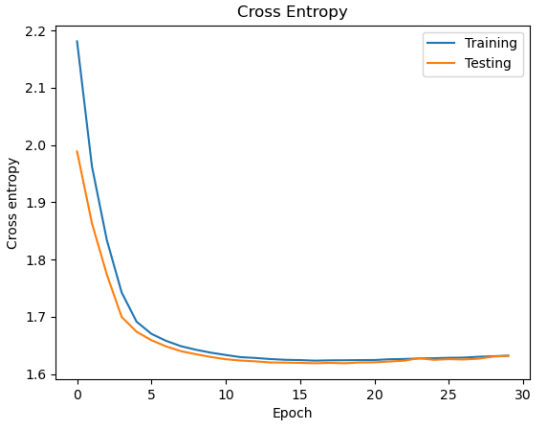
To measure the adversarial robustness of both models, we use the PGD attack with the l_∞ norm, 40 iterations, and a step size of 0.01. We chose to use the l_∞ norm instead of the l_2 in order to use the hardest possible attack for our method, and because the l_∞ norm corresponds more to the human notion of "indistinguishable images" than the l_2 norm. The attacks are performed on the test set, and only on images that were correctly classified by each model. The results are reported in table 1. These preliminary results seems to indicate that the regularization proposed in equation 14 improves the adversarial robustness, except for high attack budget. It must be mentioned that a l_∞ budget of 0.3 on MNIST can clearly be noticed with the naked eye.



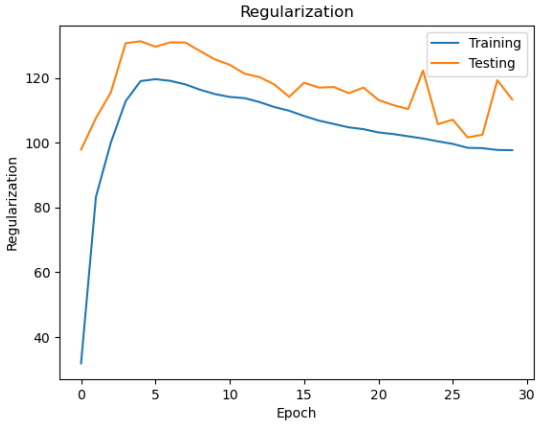
(a) Loss function (cross entropy) of the vanilla model.



(b) Loss function (cross entropy & regularization) of the regularized model (equation 15).



(c) Cross entropy of the regularized model alone.



(d) Regularization of the regularized model alone.

Figure 1: Loss functions of the vanilla model and the regularized model (equation 14).

5 Discussion and conclusion

References

[1] S.-i. Amari, *Differential-Geometrical Methods in Statistics*, vol. 28 of *Lecture Notes in Statistics*. Springer New York, 1985.

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

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

[4] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.