Pullback bundles and the geometry of learning

Stéphane Puechmorel

MDPI

Article

Pullback bundles and the geometry of learning

Stéphane Puechmorel [1,*]

[1] ENAC, Université de Toulouse, 7, Avenue Edouard Belin, Toulouse, France; stephane.puechmorel@enac.fr
[*] Correspondence: stephane.puechmorel@enac.fr

Abstract: Explainable Artificial Intelligence (XAI) and acceptable artificial intelligence are active topics of research in machine learning. For critical applications, being able to prove or at least to ensure with a high probability the correctness of algorithms is of utmost importance. In practice, however, few theoretical tools are known that can be used for this purpose. Using the Fisher Information Metric (FIM) on the output space yields interesting indicators in both the input and parameter spaces, but the underlying geometry is not yet fully understood. In this work, an approach based on the pullback bundle, a well-known trick for describing bundle morphisms, is introduced and applied to the encoder-decoder block. With constant rank hypothesis on the derivative of the network with respect to its inputs, a description of its behavior is obtained. Further generalization is gained through the introduction of the pullback generalized bundle that takes into account the sensitivity with respect to weights.

Keywords: Pullback bundle; information geometry; machine learning

1. Introduction

Explainable Artificial Intelligence (XAI) is generally described as a collection of methods allowing humans to understand how an algorithm is able to learn from a database, reproduce and generalize. It is currently an active, multidisciplinary area of research [1,2] that relies on several theoretical or heuristic tools to identify salient features and indicators explaining the surprisingly performances of machine learning algorithms, especially deep neural networks. From a statistical point of view, a neural network is nothing but a parameterized regression or classification model, that can be described as a random variable whose probability distribution is known conditionally to external inputs and internal parameters [3] . Unfortunately, even if this approach seems the most natural one, it is not adapted to XAI as no insight is gained on the learning and inference process. Furthermore, it seems that there is a contradiction between the statistical procedure that appeals for models with the smallest possible number of free parameters and the performance of deep learning relying on thousands to millions weights. On the other hand, attempts have been made to design numerical [4] or visual [5] indicators aiming at producing a summary of salient features.

XAI is also related to acceptable AI, that is proving or at least ensuring with a high probability that the model will produce the intended result and is robust to perturbations, either inherent to the data acquisition process or intentional. In both cases, it is mandatory to be able to perform a sensitivity analysis on a trained network. In [6], an approach based on geometry was taken and the need of a metric on the set of admissible perturbations enforced. The problem of the so-called adversarial attacks is treated in several papers [7–9] where mitigating procedures are proposed. Adversarial attacks are a major concern for acceptable AI, especially in critical application like autonomous vehicles or air traffic control. From now, most of the research effort was dedicated to the design of such attacks with the idea of incorporating the fooling inputs in the learning database in order to increase robustness. The reader can refer for example to Fast Gradient Sign methods [10], robust optimization methods [11] or DeepFool [12,13]. Unfortunately, while these approaches are

Citation: Puechmorel, S. Title. *Entropy* **2023**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

relevant to acceptable AI, they do not provide XAI with usable tools. Furthermore, they rely on inputs in $\mathbb{R}^n$, or generally in a finite dimensional euclidean space, which is not always a valid hypothesis.

There is also a question on why learning from a high dimension data space is possible, and a possible answer is because data effectively lies on a low dimensional manifold [14,15]. As a consequence, most of the directions in the input space will have a very small impact on the output, while only a few number of them, namely those who are tangent to the data manifold, are going to be of great influence [16]. The manifold hypothesis also justifies the introduction of the encoder-decoder architecture [17,18] that is of wide use in the field of natural language processing [19] or time-series prediction [20]. The true underlying data manifold, if it exists, is most of the time not accessible, although some of its characteristics may be known and incorporated in the model. In particular, it may be subject to some action by a Lie group or possess extra geometric properties, like the existence of a symplectic structure. Specific networks have be designed to cope with such situations [21,22].

In a general setting, little is known about the data manifold and its geometric features, like metric, Levi-Civita connection and curvature. However, Riemannian properties are the most important ones as they dictate the behavior of the network under moves in the input space. Recalling the statistical approach invoked before, it makes sense to model the output of the network as a density probability parameterized by inputs and weights. Within this frame, there exists a well-defined Riemannian metric on the output space known as the Fisher Information Metric (FIM) originating from a second order expansion of the Kullback-Liebler divergence. The importance of this metric has already been pointed out in several past works [23,24]. The FIM can be pulled back to the input space, yielding in most cases a degenerate metric that can nevertheless be exploited to better understand the effect of perturbations [25], or to parameter space to improve gradient based learning algorithms [26]. In this last case, however, things tend to be less natural than for the input space.

In this work, a unifying framework for studying the geometry of deep networks is introduced, allowing a description of encoder-decoder blocks from the FIM perspective. The pullback bundle is a key ingredient in our approach.

In the sequel, features and outputs are random variables, thus characterized by their distribution functions, or their densities in the absolutely continuous case. Within this frame, a neural network is a random variable:

$$\begin{cases} Y = \mathcal{N}(X, W) \\ X \colon (\Omega, \mathcal{T}, \mathbb{P}) \to (E, \mathcal{E}) \\ W \colon (\Omega, \mathcal{T}, \mathbb{P}) \to (\Theta, \mathcal{F}) \end{cases} \tag{1}$$

Where $(\Omega, \mathcal{T}, \mathbb{P})$ is an underlying probability space and $(E, \mathcal{E})$, $(\Theta, \mathcal{F})$ are respectively the input and weight measure spaces Finally, Y is assumed to take its values in the output measure space $(O, \mathcal{O})$. Most of the time, the network has a layered structure so that the expression of $\mathcal{N}$ can be factored out as:

$$Y = \mathcal{N}(\mathcal{N}(\dots, W_2), W_1) \tag{2}$$

In many practical implementations, the weights W are deterministic, that is equivalent to saying that their probability distribution is a Dirac distribution. In this case, a neural network can be described as a parameterized family of random variables $\mathcal{N}_W \colon \omega \mapsto \mathcal{N}(X(\omega), W)$. A special case occur when a single decoder is considered [27], that is a measurable function:

$$f = \mathcal{N}(\cdot, W) \colon \mathbb{R}^d \to \mathbb{R}^m, d \leq m \tag{3}$$

with f a smooth mapping, assumed in [27] to be an immersion, that is, for any x, Df_x has maximal rank d. Conversely, one may consider an encoder

$$g = \mathcal{N}(\cdot, \tilde{W}) \colon \mathbb{R}^n \to \mathbb{R}^d, d \leq n \tag{4}$$

and assume f to be a submersion. In this paper, the geometry of the complete encoder-decoder network:

$$g \circ f = \mathcal{N}(\mathcal{N}(\cdot, W), \tilde{W}) \tag{5}$$

will be considered, as well as the case $d \geq m, d \leq n$.

The article is structured as follows: In section 2, the Fisher information metric is introduced and some formulas, valid when the parameter space is a smooth manifold, are given. In section 3, the pullback bundle is defined and applied to the encoder-decoder case. Finally, a conclusion is drawn in section 5.

2. The Fisher information metric

In this section, we recall some basic definitions and properties in information geometry. The fundational ideas can be traced back to [28], but the main developments occur quite recently. The reader is referred to [29] for a comprehensive introduction. The exposition below assumes a quite high degree of regularity for the parameterized density families, which is nevertheless a common situation in practice, especially in the field of machine learning we are interested in.

2.1. Definitions and properties

Definition 2.1. A statistical model is a pair $(\mathcal{M}, p)$ where $\mathcal{M}$ is an oriented n dimensional smooth manifold and $(p_\theta)_{\theta \in \mathcal{M}}$ is a parameterized family of probability densities on a measure space $(\Omega, \mathcal{T}, \mu)$ such that, putting $p(\theta, \omega) = p_\theta(\omega)$:

- For μ-almost all $\omega \in \Omega$, the mapping $\theta \mapsto p(\theta, \omega)$ is smooth.
- For any $\theta \in \mathcal{M}$, there exists an open neighborhood U_θ of θ and an integrable mapping $h \colon \Omega \to \mathbb{R}^+$ such that, for any $\xi \in U_\theta$, $|\partial_\theta p(\xi, \omega)| \leq h$.
- The mapping $\theta \to p_\theta \in L^1(\Omega, \mu)$ is one-to-one.
- The support of p_θ does not depend on θ.

Assuming p never vanishes, one can define the score $l \colon \mathcal{M} \times \Omega \to \mathbb{R}$ as:

$$l(\theta, \omega) = \log p(\theta, \omega) \tag{6}$$

For any $\theta \in \mathcal{M}$:

$$\int_\Omega p_\theta(\omega) d\mu(\omega) = 1 \tag{7}$$

Thus, using the fact that the assumptions made on family p_θ allow swapping derivatives and integrals, it comes:

$$\int_\Omega \partial_i p(\theta, \omega) d\mu(\omega) = 0, \, i = 1 \dots n \tag{8}$$

where ∂_i denotes the derivative with respect to the i-th component of θ in local coordinates. So the score $l_\theta = \log p_\theta$ satisfies by 8:

$$E[\partial_i l_\theta]_{p_\theta} = 0, \, i = 1 \dots n. \tag{9}$$

A simple computation shows that:

$$E\left[\partial_i l_\theta \partial_j l_\theta\right] = \int_\Omega \frac{\partial_i p_\theta}{\sqrt{p_\theta}} \frac{\partial_j p_\theta}{\sqrt{p_\theta}} d\mu(\omega) = 4 \int_\Omega \partial_i(\sqrt{p_\theta}) \partial_j(\sqrt{p_\theta}) d\mu(\omega), \, i, j = 1 \dots n \tag{10}$$

proving that:

$$g_{ij} = E\left[\partial_i l_\theta \partial_j l_\theta\right] = \langle \partial_i(\sqrt{p_\theta}), \partial_j(\sqrt{p_\theta}) \rangle_{L^2(\Omega, \mu)} \tag{11}$$

Let g be the section of $T\mathcal{M}^* \otimes T\mathcal{M}^*$ defined by:

$$g = g_{ij} d\theta^i \otimes d\theta^j \tag{12}$$

[1] Now, given any tangent vector $X = X^i\partial_i \in T_\theta\mathcal{M}$:

$$\begin{aligned} g(\theta; X, X) = g_{ij}X^iX^j &= \langle \partial_i(\sqrt{p_\theta}), \partial_j(\sqrt{p_\theta})\rangle_{L^2(\Omega,\mu)} \\ &= \langle X^i\partial_i(\sqrt{p_\theta}), X^j\partial_j(\sqrt{p_\theta})\rangle_{L^2(\Omega,\mu)} \\ &= \langle Z, Z\rangle_{L^2(\Omega,\mu)} \end{aligned} \tag{13}$$

with $Z = X^i\partial_i(\sqrt{p_\theta})$. Given the assumptions made on the family p_θ, g is a thus a positive definite symmetric section of $T\mathcal{M} \otimes T\mathcal{M}$, hence a Riemannian metric on $\mathcal{M}$ called the Fisher Information Metric (FIM).

Remark. The mapping $\mathcal{I}\colon \theta \mapsto \sqrt{p_\theta}$ embeds $\mathcal{M}$ as a submanifold of the unit sphere in $L^2_{\Omega,\mu}$ and the Fisher information metric is just the pullback of the ambient metric in $L^2_{\Omega,\mu}$ with respect to $\mathcal{I}$. However, in machine learning applications, it is common to consider parameter spaces for which the one-to-one assumption for $\mathcal{I}$ is non-valid so that g is only positive semidefinite. The study of the rank of the metric in this case is an important research topic.

It is quite fruitful to consider differential forms on $\mathcal{M}$ parameterized by Ω. The starting point is the definition of parameterized degree 0 forms.

Definition 2.2. A parameterized 0-form is a mapping $f\colon \mathcal{M} \times \Omega \to \mathbb{R}$ satisfying:

- For almost all $\omega \in \Omega$, the mapping $\theta \in \mathcal{M} \to f(\theta, \omega)$ is smooth.
- For all $\theta_0 \in \Omega$, and all integers n, there exists a neighborhood U_{n,θ_0} and an integrable positive mapping h_{n,θ_0} such that for all $\theta \in U_{n,\theta_0}$ and almost all $\omega \in \Omega$: $|\partial^n_\theta f(\theta, \omega)| \leq h_{n,\theta_0}(\omega)$.

Proposition 2.1. *Let X be a vector field on $T\mathcal{M}$ and f a parameterized 0-form in the previous sense. Then:*

$$X(E[f]) = E[X(f)] + E[fX(l)] \tag{14}$$

with $l(\theta, \omega) = \log p(\theta, \omega)n$.

Proof. $E[f]$ is a degree 0 form on $T\mathcal{M}$. If ψ is the flow of X, then:

$$\psi^\star E[f] = \int_\Omega f(\psi(t, \theta), \omega)p(\psi(t, \theta), \omega)d\mu(\omega) \tag{15}$$

The assumptions made on f allow swapping derivatives and integrals, so:

$$\frac{\partial}{\partial t}_{t=0} E[f] = \int_\Omega \partial_\theta f(\theta, \omega)X(\theta)p(\theta, \omega)d\mu(\omega) + \int_\Omega f(\theta, \omega)\frac{\partial_\theta p(\theta, \omega)}{p(\theta, \omega)}p(\theta, \omega)d\mu(\omega) \tag{16}$$

□

Remark. Applying proposition 2.1 to the constant function $f = 1$ yields $E[X(l)] = 0$, a result already known by equation 9

A parameterized degree k differential form on $T\mathcal{M}$ can be defined readily by requiring that the coefficients of the elementary forms $d\theta_{i_1} \wedge \cdots \wedge d\theta_{i_k}$ be parameterized differential forms of degree 0.

Proposition 2.2. *Let α be a degree k parameterized differential form on $T\mathcal{M}$. Then:*

$$dE[\alpha] = E[d\alpha] + E[dl \wedge \alpha] \tag{17}$$

Proof. It is enough to consider a form $\alpha(\theta, \omega) = f(\theta, \omega)d\theta_{i_1} \wedge \cdots \wedge d\theta_{i_k}$. Then:

$$dE[\alpha](\theta) = \sum_{j=1}^n E\Big[\partial_{\theta_j} f\Big]d\theta_j \wedge d\theta_{i_1} \wedge \cdots \wedge d\theta_{i_k} + \sum_{j=1}^n E\Big[f\partial_{\theta_j} l\Big]d\theta_j \wedge d\theta_{i_1} \wedge \cdots \wedge d\theta_{i_k} \tag{18}$$

[1] The summing convention on repeated indices is used through the document.

Since:

$$d\alpha = \sum_{j=1}^{n} \Big(\partial_{\theta_j} f\Big) d\theta_j \wedge d\theta_{i_1} \wedge \cdots \wedge d\theta_{i_k} \tag{19}$$

$$dl \wedge \alpha = \sum_{j=1}^{n} f\Big(\partial_{\theta_j} l\Big) d\theta_j \wedge d\theta_{i_1} \wedge \cdots \wedge d\theta_{i_k} \tag{20}$$

the claim follows. □

Proceeding the same way as in proposition 2.1, and using Cartan's homotopy formula, we obtain:

Proposition 2.3. *Let X be a vector field on $T\mathcal{M}$ and α a degree k parameterized differential form. Then*

$$\mathcal{L}_X(E[\alpha]) = E[i_X d\alpha] + E[d i_X \alpha] + E[(i_X dl) \wedge \alpha] \tag{21}$$

When $\alpha = dl$, equation 21 reads as:

$$\mathcal{L}_X E[dl] = E\Big[i_X d^2 l\Big] + E[d(i_X dl)] + E[(i_X dl) \wedge dl] \tag{22}$$

Since $E[dl] = 0$, it comes:

$$E[d(i_X dl)] = -E[(i_X dl) \wedge dl] \tag{23}$$

Given two vector fields X, Y:

$$i_Y E[(i_X dl) \wedge dl] = E[(i_X dl)(i_Y dl)] = g(X, Y) \tag{24}$$

with g the Fisher metric. Thus:

Proposition 2.4.

$$g(X, Y) = -E[i_Y d(i_X dl)] \tag{25}$$

Remark. In coordinates, $i_Y d(l_X dl) = \partial_{ij} X^j Y^i + \partial_j l \partial_i X^j Y^i$, and after taking the expectation:

$$g(X, Y) = -E\big[\partial_{ij} l\big] X^j Y^i \tag{26}$$

This is a well-known result in the $\mathbb{R}^n$ case.

Let ∇ be an affine connection on $T\mathcal{M}$. The same computation as above yields:

Proposition 2.5. *Let X be a vector field on $T\mathcal{M}$ and α a degree k parameterized differential form. Then:*

$$\nabla_X E[\alpha] = E[\nabla_X \alpha] + E[(i_X dl) \wedge \alpha] \tag{27}$$

When $\alpha = dl$, we recover $E[\nabla_X dl](Y) = -g(X, Y)$, showing that while the parameterized Hessian ∇dl depends on the connection ∇, it is not the case of its expectation. When $\Omega = \mathcal{M} = \mathbb{R}^n, \mu = dx_1 dx_2 \dots dx_n$, the Fisher metric is known to be twice the second order term in the Taylor expansion of the Kullback-Leibler divergence, which can be proved easily by iterating derivatives. More generally, let ∇ be a connection and let $\theta :]-\epsilon, \epsilon[\to \mathcal{M}$, $\epsilon > 0$ be a smooth curve with $\theta_0 = \theta(0), X = \theta'(0)$. e recall that the Kullback-Liebler divergence between two probability densities p, q is defined as:

$$\mathrm{KL}(p, q) = E_p[\log(p/q)] = \int \log\left(\frac{p(x)}{q(x)}\right) p(x) dx \tag{28}$$

The mapping:

$$t \in]-\epsilon, \epsilon[\mapsto \xi(t) = \mathrm{KL}\Big(p_{\theta_0}, p_{\theta(t)}\Big) = E_{p_{\theta_0}}\big[l_{\theta_0}(t) - l_\theta(t)\big] \tag{29}$$

is smooth, so Taylor formula applies for t close enough to 0:

$$\xi(t) = \sum_{i=1}^{n} \frac{\xi^{(i)}(0)}{i!} t^i + o(t^n) \tag{30}$$

With:

$$\xi^{(i)}(0) = E_{p_{\theta_0}}\left[\underbrace{X(X(\dots X(l)))}_{i \text{ times}}\right] = E_{p_{\theta_0}}\left[\underbrace{X(X(\dots dl(X)))}_{i-1 \text{ times}}\right] \tag{31}$$

If the curve $t \to \theta(t)$ is a geodesic for ∇, then:

$$X(dl(X)) = (\nabla_X dl)(X) + dl(\nabla_X X) = (\nabla_X dl)(X) \tag{32}$$

And by recurrence:

$$\xi^{(i)}(0) = E_{p_{\theta_0}}\left[\left(\nabla_X^{(i-1)} dl\right)\right](X). \tag{33}$$

The first derivative $\xi^{(1)}(0)$ is readily computed as:

$$-E\left[dl_{\theta_0}\right](X) = 0. \tag{34}$$

The second derivative $\xi^{(2)}(0)$ can be obtained using ∇ as :

$$-E\left[\nabla_X dl_{\theta_0}\right](X) = g_{\theta_0}(X, X). \tag{35}$$

Since g is symmetric, $g(X, Y) = (g(X + Y, X + Y) - g(X - Y, X - Y))/4$, thus 35 characterize g at θ_0. Higher order terms can be computed by repeatedly applying proposition 2.5 and are expressed thanks to the quantities:

$$E\left[(i_X dl) \wedge \nabla_X^{(i)} dl\right](X). \tag{36}$$

An interesting case occurs when the Fisher metric is non-degenerate and ∇^{lc} is its associated Levi-Civita connection. Normal coordinates at θ_0, denoted by x^i, $i = 1 \dots N$, are given by taking an orthonormal basis, with respect to the Fisher metric, $(v_1, \dots, v_N)$ and letting [30](p. 72):

$$x^i\left(\exp_{\theta_0} t^j v_j\right) = t^i \tag{37}$$

Using the $x^i, i = 1 \dots N$ system of coordinates in place of θ, and noting that θ_0 corresponds to the origin in normal coordinates, the KL divergence can be approximated at order 2 by:

$$\text{KL}(p_0, p_x) = \frac{1}{2} x^i x^j \tag{38}$$

where $x = \left(x^1, \dots, x^N\right)$.

2.2. The Fisher information in machine learning.

In machine learning applications, when the output is a probability distribution, then the Kullback-Leibler divergence is a natural measure for goodness-of-fit. Assuming that the database is given in the form of an iid sample of couples $(X_i, Y_i)_{i=1\dots N}$, then one can introduce the error function:

$$E(W) = \sum_{i=1}^{N} \text{KL}(Y_i, \mathcal{N}(X_i, W)) \tag{39}$$

That may be approximated by:

$$\tilde{E}(W) = -\sum_{i=1}^{N} \frac{1}{2} g\left(Y_i; \overrightarrow{Y_i \mathcal{N}(X_i, W)}, \overrightarrow{Y_i \mathcal{N}(X_i, W)}\right) \tag{40}$$

Where the notation $\overrightarrow{PQ}$ stands for the tangent vector at P such that a geodesic (for ∇^{lc}) θ with $\theta(0) = P, \theta'(0) = \overrightarrow{PQ}$ is such that $\theta(1) = Q$. Taking the derivative with respect to W yields:

$$\frac{\partial \tilde{E}}{\partial W} = -\sum_{i=1}^{N} g\left(Y_i; \frac{\partial \mathcal{N}(X_i, W)}{\partial W}, \overrightarrow{Y_i \mathcal{N}(X_i, W)}\right) \tag{41}$$

$\frac{\partial \mathcal{N}(X_i, W)}{\partial W}$ being a tangent vector at Y_i.

We recall the musical isomorphism $\flat \colon TM \to TM^\star$ defined by:

$$X^\flat(Y) = g(X, Y) \tag{42}$$

and use it to rewrite 41 as:

$$\frac{\partial \tilde{E}}{\partial W} = -\sum_{i=1}^{N} \left(\frac{\partial \mathcal{N}(X_i, W)}{\partial W}\right)^\flat \left(\overrightarrow{Y_i \mathcal{N}(X_i, W)}\right) \tag{43}$$

In this form, having a critical point of the energy $\tilde{E}$ with respect to W is equivalent to the vanishing of a totally symmetric multilinear form on $T\mathcal{M} \oplus T\mathcal{M}^\star$, the generalized tangent bundle of $\mathcal{M}$.

Finally, if $\psi \colon \mathcal{N} \to \mathcal{M}$ is a smooth mapping, one can take the pullback the Fisher metric on $\mathcal{M}$ to obtain a semi-definite symmetric bilinear form on $\mathcal{N}$:

$$\psi^\star g(\eta; X, Y) = g(\psi(\eta); \psi'(\eta)X, \psi'(\eta)Y) \tag{44}$$

When ψ is an embedding, $\psi^\star g$ is a Fisher metric on $\mathcal{N}$ with $p_{\psi(\eta)}, \eta \in \mathcal{N}$ as underlying densities. This is the case considered in [27].

As an example of a pullback metric, we are going to investigate the case of the Von-Mises Fisher distribution (VMF) on $\mathbb{S}^{n-1}$ with density:

$$p_{\kappa,\mu}(x) = \frac{\kappa^{n/2-1}}{(2\pi)^{n/2} I_{n/2-1}(\kappa)} \exp(\kappa \langle x, \mu \rangle) \tag{45}$$

where $\kappa \geq 0$ is the concentration parameter, $\mu \in \mathbb{S}^{n-1}$ is the location parameter and I_k is the modified Bessel function of the first kind of order k. The Fisher metric in the embedding space $\mathbb{R}^n$ can be deduced from the second moment $E\left[xx^t\right]$ since $l_{\kappa,\mu} = \log\left(p_{\kappa,\mu}\right) = f(\kappa) + \langle x, \mu \rangle$. If κ is assumed to be constant, then:

$$E\left[\partial_\mu l_{\kappa,\mu} l^t_{\kappa,\mu}\right] = E\left[xx^t\right] \tag{46}$$

Although the expression for $E\left[xx^t\right]$ has been given in [31], we present here an alternative proof based on the fact that for any integer n, $\mathbb{S}^{n-1}$ is a suspension of $\mathbb{S}^{n-2}$. If $x = (x_1, \ldots, x_n)$, then xx^t is a matrix whose (i, j) entry is $x_i x_j$. By the rotation invariance of the VMF, μ can be selected as the first vector of an orthonormal basis, with respect to which x is expressed in components as $x = (x_1, \ldots, x_n)$. If we specialize the first component, then, if $i \neq 1, j \neq 1$:

$$\int_{\mathbb{S}^{n-1}} x_i x_j p_{\kappa,\mu}(x) dx = c_\kappa \int_0^\pi \exp(\cos\theta) \sin^{n-2}(\theta) \int_{\mathbb{S}^{n-2}} \xi_i \xi_j d\sigma_{n-2}(\xi) \tag{47}$$

with $x_i = \sin\theta\xi_i, i = 1 \dots n-1$ and σ_{n-2} the Lebesgue measure on $\mathbb{S}^{n-2}$. If $i \neq j$, then the integral vanishes by symmetry, otherwise:

$$\begin{aligned}\int_{\mathbb{S}^{n-2}} \xi_i\xi_j d\sigma_{n-2}(\xi) &= \int_0^\pi \cos^2(\psi)\sin^{n-3}(\psi)\int_{\mathbb{S}^{n-3}} d\sigma_{n-3}d\psi \\ &= \int_0^\pi \cos^2(\psi)\sin^{n-3}(\psi)d\psi\mathcal{A}\left(\mathbb{S}^{n-3}\right)\end{aligned} \tag{48}$$

with $\mathcal{A}\left(\mathbb{S}^{n-3}\right)$ the area of the $n-3$-sphere, which is given by the general relation:

$$\mathcal{A}(\mathbb{S}^n) = \frac{2\pi^{\frac{n+1}{2}}}{\Gamma\left(\frac{n+1}{2}\right)} \tag{49}$$

Now, observing that [32]:

$$\int_0^\pi \cos^2(\psi)\sin^{n-3}(\psi)d\psi = B\left(\frac{3}{2}, \frac{n}{2}-1\right) \tag{50}$$

with B the beta function, the overall expression becomes, after using 49:

$$\begin{aligned}&\frac{(2\pi)^{n/2}\Gamma\left(\frac{n}{2}-1\right)I_{n/2}(\kappa)\kappa^{n/2-1}}{\kappa^{n/2}\Gamma(n/2-1)(2\pi)^{n/2}I_{n/2-1}(\kappa)} \\ &= \frac{1}{\kappa}\frac{I_{n/2}(\kappa)}{I_{n/2-1}(\kappa)}\end{aligned} \tag{51}$$

When $i = j = 1$, then the expression for the second moment becomes:

$$\begin{aligned}&\int_0^\pi \exp(\kappa\cos\theta)\cos^2(\theta)\sin^{n-2}(\theta)d\theta\mathcal{A}\left(\mathbb{S}^{n-2}\right) = \\ &\int_0^\pi \exp(\kappa\cos\theta)\left(1-\sin^2(\theta)\right)(\theta)\sin^{n-2}(\theta)d\theta\mathcal{A}\left(\mathbb{S}^{n-2}\right)\end{aligned} \tag{52}$$

The integral is a difference of two terms, each of which can be simplified as before to yield:

$$\left(1-\frac{n}{\kappa}\right)\frac{I_{n/2}(\kappa)}{I_{n/2-1}(\kappa)} \tag{53}$$

This procedure can easily be applied to an arbitrary moment, each of the integral involved being expressible using I_n and the Beta function.

Remark. Since μ is not a parameterization of the unit sphere, the Fisher metric defined that way is related to an ambient metric in $\mathbb{R}^n$, defined only on the unit sphere.

An obvious embedded dimension $n-2$ submanifold of $\mathbb{S}^{n-1}$ is obtained by taking a unit vector ν and computing the intersection of $\mathbb{S}^{n-1}$ with an hyperplane $\mathcal{H}$ defined by:

$$x \in \mathcal{H} \Leftrightarrow \langle x, \nu\rangle = \alpha \; \alpha \in]0,1[\tag{54}$$

An elementary computation proves that the intersection locus is a $n-2$ sphere contained in $\mathcal{H}$:

$$|x - \alpha\nu|^2 = 1 - \alpha^2 \tag{55}$$

Without loss of generality, ν can be taken as $\begin{pmatrix}1 & 0 & \dots & 0\end{pmatrix}$ and the embedding can be written easily as:

$$\begin{pmatrix}x_1 & \dots & x_{n-1}\end{pmatrix} \mapsto \begin{pmatrix}\alpha & \lambda x_1 & \dots & \lambda x_{n-1}\end{pmatrix}, \lambda = \sqrt{1-\alpha^2} \tag{56}$$

The pullback metric is just the original one scaled by $1 - \alpha^2$. Loss functions related to the VMF distribution are discussed in [33].

3. Pullback bundles

In this section, a neural network with weights W is a mapping $\mathcal{N}(\cdot, W)\colon \mathcal{I} \to \mathcal{O}$, where $\mathcal{I}$ (resp. $\mathcal{O}$) is the input (resp. output) manifold of dimension n (resp. m). Both manifolds are assumed to be smooth, and also the mapping $\mathcal{N}_W$. This last assumption is valid when the activation functions are smooth, which is the case for sigmoid functions, but not for the commonly used ReLu function. However, smooth approximations to the ReLu are easy to construct with an arbitrary degree of accuracy, so the framework introduced below can be still applied.

As mentioned in the introduction, $\mathcal{O}$ is further assumed to be a statistical model 2.1 with Fisher metric g. This setting is the one of a neural network whose output is a random variable with conditional density in a family $p_\theta, \theta \in \mathcal{O}$.

When the weights are kept fixed, the only free parameters are the inputs and the network is fully described by the mapping:

$$\begin{cases} \mathcal{N}(\cdot, W)\colon & \mathcal{I} \to \mathcal{O} \\ & x \mapsto \mathcal{N}(\cdot, W) = p_{\theta(x)} \end{cases} \tag{57}$$

For the ease of notation, the mapping $\mathcal{N}(\cdot, W)$ will be abbreviated by $\mathcal{N}_W(\cdot)$. When the activation functions in the network are smooth, $\mathcal{N}_W(\cdot)$ is a smooth mapping and its derivative will be denoted by $d\mathcal{N}_W(\cdots)$. With this convention, the pullback metric of g by $\mathcal{N}_W(\cdot)$, denoted $\tilde{g}$, is defined by:

$$\tilde{g}(X, Y) = g(d\mathcal{N}_W(X), d\mathcal{N}_W(Y)) \tag{58}$$

Unless the network $\mathcal{N}$ is a decoder, $\tilde{g}$ is generally degenerated and does not provide $\mathcal{I}$ with a Riemannian structure, so an ambient metric h on $\mathcal{I}$ is assumed to exist. The triple $(\mathcal{I}, h, \tilde{g})$ is called the data manifold of the network. The kernel of $\tilde{g}$, denoted $\ker \tilde{g}$, is the distribution in $T\mathcal{I}$ consisting of vectors X such that $\tilde{g}(X, \cdot)$ is the zero mapping. At a point $x \in \mathcal{I}$, the vectors in $T_x\mathcal{I}$ belonging to $\ker \tilde{g}$ give directions in which the output of the network will not change up to order 1. Figure 1 represents the case of a one dimensional output space and a 2-sphere input space. Since the dimension of the output is less than the one of the input, some moves in the data manifold will not induce any change at the output. Unless the

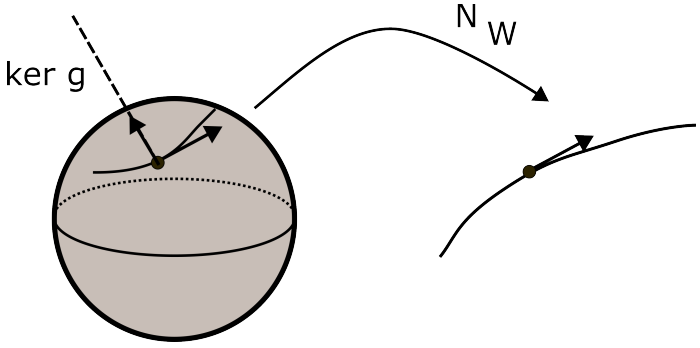

Figure 1. Kernel of the pullback metric.

dimension of $\ker \tilde{g}$ is constant, this distribution does not define a foliation. However, this is true locally in the neighborhood of points in $\mathcal{I}$ such that $d\mathcal{N}_W(\cdot)$ has maximal rank. Finally, if $E \xrightarrow{\pi} \mathcal{O}$ is an r-vector bundle on $\mathcal{O}$, then its pullback by $\mathcal{N}_W(\cdot)$ will be denoted in short by $E^{\mathcal{N}_W}$. We recall that if E has local charts:

$$(V_i, \xi_i),\ \xi_i\colon V_i \times \mathbb{R}^r \to \pi^{-1}(V_i),\ i \in I$$

and $\mathcal{I}$ has local charts (U_j, ϕ_i), $j \in J$, then $E^{\mathcal{N}_W}$ has local charts:

$$\begin{aligned} &\left(W_{ji} = U_j \cap \mathcal{N}_W^{-1}(V_i), \psi_{ji}\right),\ \psi_{ji} \colon W_{ji} \to W_{ji} \times \mathbb{R}^r \\ &\psi_{ji}(x) = \xi_i \circ f \circ \phi_j \end{aligned} \tag{59}$$

The pullback bundle enjoys a universal property that is in fact the main reason for introducing it in our context.

Proposition 3.1. *Let $(\tilde{E}, \tilde{\pi}, \mathcal{I})$ (resp. $(E, \pi, \mathcal{O})$) be a vector bundle on $\mathcal{I}$ (resp. $\mathcal{O}$)). For any bundle morphism (η_1, η_0), there exists a unique bundle morphism $(\tilde{\eta}_1, Id)$ such that the following diagram commutes:*

$$\begin{array}{ccccc} & & \overset{\eta_1}{\frown} & & \\ \tilde{E} & \xrightarrow[\tilde{\eta}_1]{} & E^{\eta_0} & \xrightarrow{\tilde{\eta}_0} & E \\ \downarrow{\scriptstyle \tilde{\pi}} & & \downarrow{\scriptstyle \pi^{\eta_0}} & & \downarrow{\scriptstyle \pi} \\ \mathcal{I} & = & \mathcal{I} & \xrightarrow{\eta_0} & \mathcal{O} \end{array} \tag{60}$$

where $\pi^{\eta_0} \colon (x, v) \mapsto x$ and $\tilde{\eta}_0 \colon (x, v) \mapsto (\eta_0(x), v)$.

This proposition is a classical one and its proof can be found in many textbooks. The one we give below is very simple, using only local charts.

The above construction is constructive and thus gives a practical mean of computation. For a network with fixed weights, e.g. a trained one, the derivative $d\mathcal{N}_W$ can be efficiently computed by back propagation, so the bundle morphism:

$$\begin{array}{ccc} T\mathcal{I} & \xrightarrow{d\mathcal{N}_W} & T\mathcal{O} \\ \downarrow{\scriptstyle \pi_{\mathcal{I}}} & & \downarrow{\scriptstyle \pi_{\mathcal{O}}} \\ \mathcal{I} & \xrightarrow{\mathcal{N}_W} & \mathcal{O} \end{array} \tag{61}$$

has a practical meaning.

Introducing the pullback bundle gives the diagram:

$$\begin{array}{ccccc} T\mathcal{I} & \xrightarrow{d\mathcal{N}_W} & T^{\mathcal{N}_W}\mathcal{O} & \xrightarrow{\widetilde{\mathcal{N}_W}} & T\mathcal{O} \\ \downarrow{\scriptstyle \pi_{\mathcal{I}}} & & \downarrow{\scriptstyle \pi^{\mathcal{N}_W}} & & \downarrow{\scriptstyle \pi_{\mathcal{O}}} \\ \mathcal{I} & = & \mathcal{I} & \xrightarrow{\mathcal{N}_W} & \mathcal{O} \end{array} \tag{62}$$

The bundle mapping $d\mathcal{N}_W$ to $T^{\mathcal{N}_W}\mathcal{O}$ is then the association:

$$(x, v) \in \mathbb{R}^m \times \mathbb{R}^n \mapsto (x, d\mathcal{N}_W \cdot v) \in \mathbb{R}^n \times \mathbb{R}^n \tag{63}$$

The pullback bundle is thus a mean of representing the action of the network on tangent vectors to the data manifold. As an example, the construction of adversarial attacks given in [34,35] can be revisited in this context, extending it to the general setting of network with manifold inputs.

The general problem of building an adversarial attack is, informally, to find, for an input point in the data manifold, a direction in which a perturbation will have the most important effect on the output, hopefully fooling the network. Following [35], we define:

Definition 3.1. Let h be a Riemannian metric on the input space. An optimal adversarial attack at $x \in \mathcal{I}$ with budget $\epsilon > 0$ is a solution to:

$$\max_{v \in T_x\mathcal{I}, h(v,v) \leq \epsilon} \tilde{g}(v, v) \tag{64}$$

Using 38, this optimization program can be viewed as a local approximation to the one based on the Kullback-Liebler divergence:

Definition 3.2. A Kullback-Liebler optimal adversarial attack at $x \in \mathcal{I}$ with budget $\epsilon > 0$ is a solution to:

$$\max_{y \in \mathcal{I}, h(x,y) \leq \epsilon} KL(\mathcal{N}(x, W), \mathcal{N}(y, W)) \tag{65}$$

The metric g on $T\mathcal{O}$ can be pulled back to $T^{\mathcal{N}_W}$ by letting:

$$g^{\mathcal{N}_W}(x; v, v) = g(f(x); v, v) \tag{66}$$

Due to the special form of the criterion, the optimal point is on the boundary, so that finally, the optimal adversarial attack problem may be formulated as:

Definition 3.3. An optimal adversarial attack at $x \in \mathcal{I}$ with budget $\epsilon > 0$ is a solution to:

$$\max_{v \in UT_x\mathcal{I}} \epsilon^2 g^{\mathcal{N}_W}(d\mathcal{N}_W v, d\mathcal{N}_W v) \tag{67}$$

Where $UT\mathcal{I}$ stands for the unit sphere bundle with respect to the metric h. Please note that due to bilinearity, the problem can be solved for $\epsilon = 1$, then let the optimal vector be scaled by the original ϵ. From standard linear algebra, if G_x is the matrix of the bilinear form $g^{\mathcal{N}_W}$ at x and H_x the one of h, then one can find unitary matrices A, B and diagonal matrices Λ, Σ such that:

$$H_x = A^t \Lambda A,\ G_x = B^t \Sigma B \tag{68}$$

Any vector v in $UT_x\mathcal{I}$ can be written as:

$$V = A^t \Lambda^{-1/2} w, w^t w = 1 \tag{69}$$

So that finally the original problem can be rewritten as:

$$\max_{w, w^t w = 1} w^t M^t M w,\ M = \Sigma^{1/2} B d_x \mathcal{N}_W A^t \Lambda^{-1/2} \tag{70}$$

Which is solved readily by taking w to be the unit eigenvector of M associated with the largest eigenvalue. This is the solution found in [35] when $H_x = \mathrm{Id}$.

In many cases, as the above example indicates, it is more convenient to work uniquely in the input space, thus justifying the introduction of the pullback bundle $T^{\mathcal{N}_W}\mathcal{O}$. From now, we are going to adopt this point of view.

Remark. Please note that a section in $T^{\mathcal{N}_w}\mathcal{I}$ is generally not related to a section of the form 63 in either $T\mathcal{O}$ or $T\mathcal{I}$ due to the fact that $d_x\mathcal{N}_W$ may not be a monomorphism or an epimorphism. The next proposition gives condition for the existence of global sections in $T\mathcal{O}$ associated to global sections in $T^{\mathcal{N}_W}\mathcal{O}$.

Proposition 3.2. *In the case of a decoding network, when $\mathcal{N}_W$ is an embedding, there is a natural embedding of bundles $T\mathcal{I} \xhookrightarrow{i} T^{\mathcal{N}_W}\mathcal{O}$ such that the image of (x, v) is $(x, d\mathcal{N}_W v)$. The pullback bundle then splits as:*

$$T^{\mathcal{N}_W}\mathcal{O} = i(T\mathcal{I}) \oplus F \tag{71}$$

where F has rank $n - m$.

Be careful that in this case, a section of the pullback bundle will not define a global section in $T\mathcal{O}$ since some points of the output space may have no preimage by $\mathcal{N}_W$. However, by the extension lemma [36](Lemma 5.34, p. 115), it exists local (global if $\mathcal{N}_W\mathcal{I}$ is closed) smooth vector fields on $T\mathcal{O}$ extending it.

Proof. If $\mathcal{N}_W$ is an embedding, $\mathcal{N}_W(\mathcal{I})$ is a submanifold of $\mathcal{O}$ and in an adapted chart, a vector field in $T\mathcal{N}_W(\mathcal{I})$ can be written as $v = \sum_{i=1}^{n} v^i \partial_i$, where the $\partial_i, i = 1 \dots n$ are the first n coordinate vector fields. It thus pulls back to a section $\tilde{v}$ of the same form in $T^{\mathcal{N}_W}\mathcal{O}$. Now, since $d\mathcal{N}_W$ is injective, $\tilde{v}$ is the image of a unique section in $T\mathcal{I}$, hence the claim. □

Proposition 3.3. *If* $\ker d\mathcal{N}_W$ *has constant rank* r*, then there exists a splitting* $T\mathcal{I} = \ker d\mathcal{N}_W \oplus F$*,* $T^{\mathcal{N}_W}\mathcal{O} = im\ d\mathcal{N}_W \oplus G$ *and bundle isomorphism* $F \to im\ d\mathcal{N}_W$ *that coincides with* $d\mathcal{N}_W$ *on the fibers.*

Proof. By Theorem 10.34, [36](p. 266), $\ker d\mathcal{N}_W$ is a subbundle of $T\mathcal{I}$ and $\mathsf{im}\ d\mathcal{N}_W$ a subbundle of $T^{\mathcal{N}_W}\mathcal{O}$. In local charts, the morphism $d\mathcal{N}_W$ gives rise to the decomposition:

$$\ker d\mathcal{N}_W \oplus \mathbb{R}^r \xrightarrow{d\mathcal{N}_W} \mathsf{im}\ d\mathcal{N}_W \oplus \mathbb{R}^{m-r} \tag{72}$$

with $d\mathcal{N}_W$ an isomorphism where restricted to $\mathbb{R}^r$. Passing to local sections yields the result. □

An important case is the one of submersions, corresponding to encoders in machine learning. In this case, $r = m$ and $d\mathcal{N}_W$ establishes a bundle isomorphism between F and $T^{\mathcal{N}_W}\mathcal{O}$. The pullback of Fisher-Rao metric g on $T\mathcal{O}$ gives rise to a metric $g^{\mathcal{N}_W}$ on $T^{\mathcal{N}_W}\mathcal{O}$, but only to a degenerate metric on $T\mathcal{I}$ that can, nevertheless, be quite well understood, as indicated below.

Definition 3.4. On the input bundle $T\mathcal{I}$, the symmetric tensor $\tilde{g}$ is defined using the splitting $T\mathcal{I} = \ker d\mathcal{N}_W \oplus F$, by:

$$\begin{aligned} g(X,Y) &= 0,\ X \in \ker d\mathcal{N}_W,\ Y \in T\mathcal{I} \\ g(X,Y) &= g^{\mathcal{N}_W}(d\mathcal{N}_W X, d\mathcal{N}_W Y),\ X, Y \in F \end{aligned} \tag{73}$$

Proposition 3.4. *There exists a symmetric* $(1,1)$*-tensor on* $\mathcal{I}$*, denoted by* Θ*, such that, for any tangent vectors* $(X,Y) \in T\mathcal{I}$*:*

$$h(\Theta X, Y) = \tilde{g}(X,Y) \tag{74}$$

Proof. From standard linear algebra, there exists an adjoint ${}^t d\mathcal{N}_W$ to $d\mathcal{N}_W$, defined by:

$$g^{\mathcal{N}_W}(d\mathcal{N}_W v, d\mathcal{N}_W v) = h\left({}^t d\mathcal{N}_W v, d\mathcal{N}_W v\right) \tag{75}$$

with, in local coordinates:

$${}^t N_i^j = h^{il} N_l^k g_{ij}^{\mathcal{N}_W} \tag{76}$$

where N (resp. ${}^t N$) is the matrix associated to $d\mathcal{N}_W$ (resp. ${}^t d\mathcal{N}_W$) and, as usual, $h^{il} = \left(h^{-1}\right)_{il}$. The $(1,1)$-tensor Θ is then the product ${}^t d\mathcal{N}_W d\mathcal{N}_W$. □

Remark. Θ is defined even if $d\mathcal{N}_W$ is not full rank.

Remark. All the relevant information concerning $d\mathcal{N}_W$ is encoded in Θ. As a consequence, the geometry of an encoder is described by this tensor, hence also the one of an encoder-decoder block.

Remark. The tensor Θ has expression $g_{pj} N_i^j N_k^p$ in a local orthonormal frame, hence is symmetric.

Definition 3.5. Let ∇ be a connection on $T\mathcal{I}$. Its dual connection $\nabla^\star$ is defined by the next equation:

$$\nabla_Z h(X,Y) = h(\nabla_Z X, Y) + h(X, \nabla_Z^\star Y) \tag{77}$$

where Z is any tangent vector in $T\mathcal{I}$ and X, Y are vector fields.

Definition 3.6. A $(1,1)$-tensor Θ is said to satisfy the gauge equation [37] if, for all tangent vectors Z:

$$\nabla_Z^\star \Theta = \Theta \nabla_Z \tag{78}$$

Proposition 3.5. *If Θ satisfies the gauge equation 78, then the $(0,2)$-tensor defined by:*

$$(X, Y) \mapsto h(\Theta X, Y) \tag{79}$$

is ∇ parallel.

Proof. For any vector fields X, Y and any tangent vector Z:

$$\begin{aligned} \nabla^\star_Z h(\Theta X, Y) &= h(\nabla^\star_Z \Theta X, Y) + h(\Theta X, \nabla_Z Y) \\ &= h(\Theta \nabla_Z X, Y) + h(\Theta X, \nabla_Z Y) \end{aligned} \tag{80}$$

hence the claim. □

Θ, being symmetric, admits a diagonal expression in a local orthonormal local frame $(X_1, \dots, X_n)$. When there exists a connection ∇ such that $\nabla^\star_Z \Theta X = \Theta \nabla_Z X$ for any vector fields X, Z, parallel transport of the $X_i, i = 1 \dots n$ shows that the eigenvalues are constant and the eigenspaces preserved. The existence of a solution to the gauge equation thus greatly simplifies the study of an encoder, as a local splitting of the input manifold exists. The reader is referred to [37] for more details. In fact, the tensor Θ is defined even if for general networks and the splitting may exist in this setting. This is the case when the rank of $d\mathcal{N}_W$ is locally constant, hence when it is maximal. A practical computation of Θ can be obtained through the singular value decomposition, as prop. 74 indicates. A numerical integration of the distribution given by the first singular vectors gives rise to a local system of coordinates, defining in turn a connection satisfying the gauge equation (the existence of a global solution has a cohomological obstruction that is outside the scope of this paper).

Finally, we introduce below a construction that takes into account the weight influence. As mentioned in section 2, the derivative of the network with respect to its weights is adequately described as a 1-form, thus a section of $T^\star\mathcal{O}$. In fact, when the inner layers of the network are manifolds, the parameters are no longer real values and a suitable extension has to be introduced. One possible approach is to take a connection ∇ on the layer manifold $\mathcal{L}$. Considering a point $p \in \mathcal{L}$, the exponential $\exp^\nabla$ defines a local chart centered at p. Given a point q in the injectivity domain of $\exp^\nabla$, one can obtain its coordinates as $\log^\nabla_p q = \vec{pq}$ and the activation of a neuron with input q as $\alpha(\vec{pq})$, with α a 1-from in $T^\star\mathcal{L}$. In this general setting, a manifold neuron will be defined by its input in an exponential chart, a 1-form corresponding to the weights in the euclidean setting and an activation function. Its free parameters are thus a couple $(q, \alpha) \in T\mathcal{L} \oplus T^\star\mathcal{L}$. This particular vector bundle is known as the generalized tangent bundle.

Recalling 43, it is worth to study the pullback of the generalized bundle $T\mathcal{O} \oplus T^\star\mathcal{O}$. The generalized pullback bundle is then $T^{\mathcal{N}_W}\mathcal{O} \oplus T^{\star\mathcal{N}_W}\mathcal{O}$ whose local sections are generated by the pullback local sections of the form:

$$(x, v(\mathcal{N}_W(x)), \alpha(\mathcal{N}_W(x))) \tag{81}$$

Please note that the pullback can be performed on any layer, internal or input. Most of the previous derivations can be carried out on the generalized bundle, which must be thus considered as a general, yet tractable framework for XAI.

4. A numerical example.

In this example, the input data are the handwritten digits from the MNIST database. A neural network with the next architecture was coded in torch 2 and trained on the dataset:

- First layer: convolutional, kernel size of 3, nonlinearity sigmoid.
- Second layer: convolutional, kernel size of 3, nonlinearity sigmoid.
- Pooling layer.
- Two linear layers.
- Softmax layer.

The input metric is euclidean, the output one is the Fisher metric of the multinomial distribution with ten classes, that is given by the matrix:

$$\begin{pmatrix} p_1^{-1} & 0 & \dots & 0 \\ 0 & p_2^{-1} & \ddots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & p_9^{-1} \end{pmatrix} + \frac{1}{p_{10}} \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \vdots & \vdots \end{pmatrix} \tag{82}$$

Since the output space has dimension 9, the pullback bundle also has dimension 9. At an input point x, a point in the pullback bundle is a couple (x, v) with v a vector from $\mathbb{R}^9$ at output point $\mathcal{N}_W(x)$. On the other hand, the image of the input tangent bundle (simply a vector space in our case) has points $(x, d\mathcal{N}_W u)$ with u an input vector. We are thus considering a bundle mapping $(x, d\mathcal{N}_W u) \mapsto (x, d\mathcal{N}_W u)$ where the right-hand term has values in the pullback bundle, equipped with the output Fisher metric. Tensor Θ is computed via singular value decomposition, already implemented in torch. We selected the rotation rate of the singular vector associated to the largest singular value as an indicator of the complexity of the decision process in the neighborhood of an input point. The code was adapted from https://github.com/eliot-tron/CurvNetAttack. A detection of outliers from a sample of 1000 points was performed. A visual analysis reveals that they correspond to poorly drawn digits, as indicated in figure 2 where the two digits with the highest curvature indicator are plotted: The first one is labeled "9", which is quite obvious for a

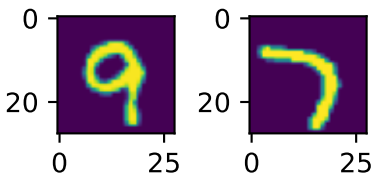

Figure 2. Samples with the highest rotation rate.

human operator, although the final stroke is vertical, while the second is labeled "7", easily confused with a "1".

5. Conclusion and future work

In this paper, several important constructions originating from information geometry were surveyed and some new ones introduced. The pullback bundle on a layer allows to describe the behavior of a network with respect to the Fisher information metric, and a simple description can be obtained when a gauge equation is satisfied. One important feature of this construction is its ability to fit in a general framework where layers take their inputs on a manifold.

Future work involves a companion paper describing computational procedures and examples from real case studies. An study of the properties of the pullback generalized bundle is also in progress. Finally, the case of networks with non constant rank $d\mathcal{N}_W$ must be considered. It is believed that they give rise to singular foliations.

Author Contributions: The author of this article is the only contributor.

Funding: "This research received no external funding"

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

References

1. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*. https://doi.org/10.3390/e23010018.
2. Chamola, V.; Hassija, V.; Sulthana, A.R.; Ghosh, D.; Dhingra, D.; Sikdar, B. A Review of Trustworthy and Explainable Artificial Intelligence (XAI). *IEEE Access* **2023**, *11*, 78994–79015. https://doi.org/10.1109/ACCESS.2023.3294569.
3. Chang, D.T. Probabilistic Deep Learning with Probabilistic Neural Networks and Deep Probabilistic Models, 2021, [arXiv:cs.LG/2106.00120].
4. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* **2013**.
5. Alicioglu, G.; Sun, B. A survey of visual analytics for Explainable Artificial Intelligence methods. *Computers & Graphics* **2022**, *102*, 502–520. https://doi.org/https://doi.org/10.1016/j.cag.2021.09.002.
6. Fawzi, A.; Moosavi-Dezfooli, S.M.; Frossard, P. The Robustness of Deep Networks: A Geometrical Perspective. *IEEE Signal Processing Magazine* **2017**, *34*, 50–62. https://doi.org/10.1109/MSP.2017.2740965.
7. Fawzi, A.; Fawzi, O.; Frossard, P. Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning* **2015**, *107*, 481–508.
8. Wong, E.; Kolter, Z. Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope. In Proceedings of the Proceedings of the 35th International Conference on Machine Learning; Dy, J.; Krause, A., Eds. PMLR, 10–15 Jul 2018, Vol. 80, *Proceedings of Machine Learning Research*, pp. 5286–5295.
9. Raghunathan, A.; Steinhardt, J.; Liang, P. Certified Defenses against Adversarial Examples. *ArXiv* **2018**, *abs/1801.09344*.
10. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. *CoRR* **2014**, *abs/1412.6572*.
11. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. *ArXiv* **2017**, *abs/1706.06083*.
12. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* **2015**, pp. 2574–2582.
13. Abdollahpourrostam, A.; Abroshan, M.; Moosavi-Dezfooli, S.M. Revisiting DeepFool: generalization and improvement. *ArXiv* **2023**, *abs/2303.12481*.
14. Fefferman, C.; Mitter, S.K.; Narayanan, H. Testing the Manifold Hypothesis. *arXiv: Statistics Theory* **2013**.
15. Narayanan, H.; Mitter, S.K. Sample Complexity of Testing the Manifold Hypothesis. In Proceedings of the NIPS, 2010.
16. Grementieri, L.; Fioresi, R. Model-centric Data Manifold: the Data Through the Eyes of the Model, 2021, [arXiv:cs.LG/2104.13289].
17. Ye, J.C.; Sung, W.K. Material for " Understanding Geometry of Encoder-Decoder CNNs " (Proof included). 2019.
18. Ye, J.C.; Sung, W.K. Understanding Geometry of Encoder-Decoder CNNs. *ArXiv* **2019**, *abs/1901.07647*.
19. Zhang, Z.; Yu, W.; Zhu, C.; Jiang, M. A Unified Encoder-Decoder Framework with Entity Memory. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2022.
20. Zhang, H.; Li, S.; Chen, Y.; Dai, J.; Yi, Y. A Novel Encoder-Decoder Model for Multivariate Time Series Forecasting. *Computational Intelligence and Neuroscience* **2022**, *2022*.
21. Ju, C.; Guan, C. Deep Optimal Transport for Domain Adaptation on SPD Manifolds. 2022.
22. Santos, S.; Ekal, M.; Ventura, R. Symplectic Momentum Neural Networks - Using Discrete Variational Mechanics as a prior in Deep Learning. In Proceedings of the Conference on Learning for Dynamics & Control, 2022.
23. Karakida, R.; Okada, M.; Amari, S. Adaptive Natural Gradient Learning Based on Riemannian Metric of Score Matching. 2016.
24. Amari, S.; Karakida, R.; Oizumi, M. Fisher Information and Natural Gradient Learning of Random Deep Networks. In Proceedings of the International Conference on Artificial Intelligence and Statistics, 2018.
25. Grementieri, L.; Fioresi, R. Model-centric Data Manifold: the Data Through the Eyes of the Model. *ArXiv* **2021**, *abs/2104.13289*.
26. Karakida, R.; Akaho, S.; Amari, S. Universal statistics of Fisher information in deep neural networks: mean field approach. *Journal of Statistical Mechanics: Theory and Experiment* **2018**, *2020*.
27. Arvanitidis, G.; González-Duque, M.; Pouplin, A.; Kalatzis, D.; Hauberg, S. Pulling back information geometry. In Proceedings of the 25th International Conference on Artificial Intelligence and Statistics; Camps-Valls, G.; Ruiz, F.J.R.; Valera, I., Eds. PMLR, 28–30 Mar 2022, Vol. 151, *Proceedings of Machine Learning Research*, pp. 4872–4894.
28. Rao, C.R., Information and the Accuracy Attainable in the Estimation of Statistical Parameters. In *Breakthroughs in Statistics: Foundations and Basic Theory*; Springer New York: New York, NY, 1992; pp. 235–247. https://doi.org/10.1007/978-1-4612-0919-5_16.
29. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Fields Institute Communications, American Mathematical Society, 2000.
30. Willmore, T. *Riemannian Geometry*; Oxford science publications, Oxford University Press, 1996.
31. Kitagawa, T.; Rowley, J. von Mises-Fisher distributions and their statistical divergence, 2022, [arXiv:econ.EM/2202.05192].
32. *NIST Digital Library of Mathematical Functions*. https://dlmf.nist.gov/, Release 1.1.10 of 2023-06-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.
33. Scott, T.R.; Gallagher, A.C.; Mozer, M.C. von Mises-Fisher Loss: An Exploration of Embedding Geometries for Supervised Learning, 2021, [arXiv:cs.LG/2103.15718].
34. Martin, J.; Elster, C. Inspecting adversarial examples using the fisher information. *Neurocomputing* **2020**, *382*, 80–86. https://doi.org/https://doi.org/10.1016/j.neucom.2019.11.052.

35. 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 Proceedings of the AAAI Conference on Artificial Intelligence, 2019, Vol. 33, pp. 5869–5876.
36. Lee, J. *Introduction to Smooth Manifolds*; Graduate Texts in Mathematics, Springer New York, 2012.
37. Boyom, M.N. Foliations-Webs-Hessian Geometry-Information Geometry-Entropy and Cohomology. *Entropy* **2016**, *18*, 433.