PhD Defense

The Geometry of Neural Networks: a Riemannian Foliation Perspective on Robustness

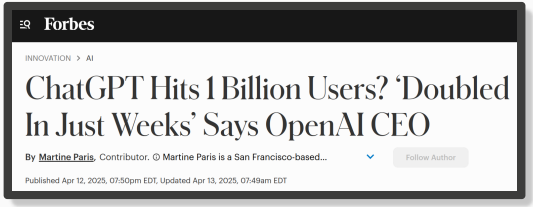
Eliot Tron

Supervision:
Nicolas Couellan, ENAC (Supervisor)
Rita Fioresi, Università di Bologna (Co-supervisor)
Stéphane Puechmorel, ENAC (Co-advisor)

20th October 2025
ENAC, Toulouse, France

Jury:
Jesús Angulo Lopez, Mines Paris (Reviewer)
Alice Barbara Tumpach, Université de Lille & Institut CNRS Pauli (Reviewer)
Stéphanie Allassonniere, Université Paris Cité (Examiner)
Mathieu Serrurier, Université de Toulouse Jean-Jaurès (Examiner)

General context



Forbes

INNOVATION > AI

ChatGPT Hits 1 Billion Users? 'Doubled In Just Weeks' Says OpenAI CEO

By Martine Paris, Contributor. ⓘ Martine Paris is a San Francisco-based...

Follow Author

Published Apr 12, 2025, 07:50pm EDT, Updated Apr 13, 2025, 07:49am EDT

General context

Forbes

INNOVATION > AI

ChatGPT Hits 1 Billion Users? 'Doubled In Just Weeks' Says OpenAI CEO

By Martine Paris, Contributor. ⓘ Martine Paris is a San Francisco-based...

Follow Author

Published Apr 12, 2025, 07:50pm EDT, Updated Apr 13, 2025, 07:49am EDT

The Washington Post

Tesla launches long-awaited Robotaxi in Austin

June 22, 2025

General context

Forbes

INNOVATION > AI

ChatGPT Hits 1 Billion Users? 'Doubled In Just Weeks' Says OpenAI CEO

By Martine Paris, Contributor. ⓘ Martine Paris is a San Francisco-based...

Follow Author

Published Apr 12, 2025, 07:50pm EDT, Updated Apr 13, 2025, 07:49am EDT

The Washington Post

Tesla launches long-awaited Robotaxi in Austin

June 22, 2025

Forbes

SMALL BUSINESS

The Future Of Aviation: How Could AI Reshape The Industry?

By Tanya Fileva, Forbes Councils Member.

for Forbes Business Council, COUNCIL POST | Membership (fee-based)

Published Jun 27, 2025, 07:00am EDT, Updated Jun 30, 2025, 03:03pm EDT

Good results

▷ accuracy, performances...

Controlled results

▷ fairness, explainability...

Consistent results

▷ robustness, controlled sensitivity...

Good results

▷ accuracy, performances...

Controlled results

▷ fairness, explainability...

Consistent results

▷ robustness, controlled sensitivity...

Good results

▷ accuracy, performances...

Controlled results

▷ fairness, explainability...

Consistent results

▷ robustness, controlled sensitivity...

Good results

▷ accuracy, performances...

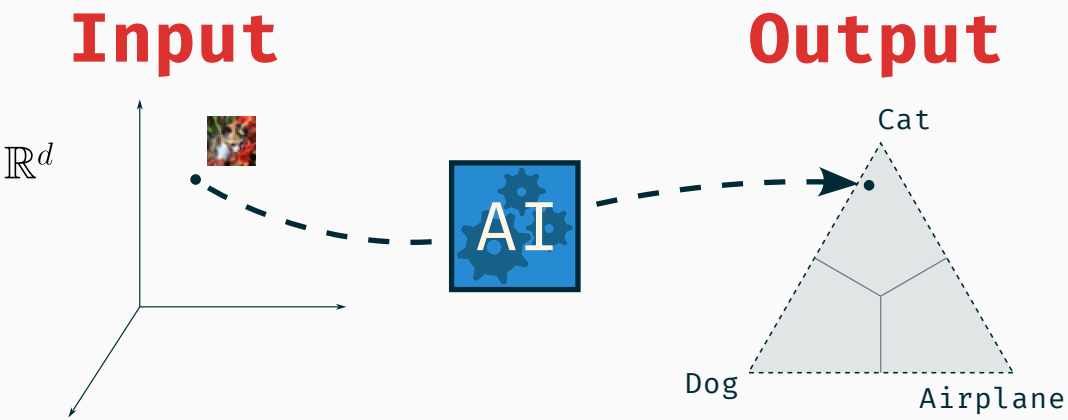
Controlled results

▷ fairness, explainability...

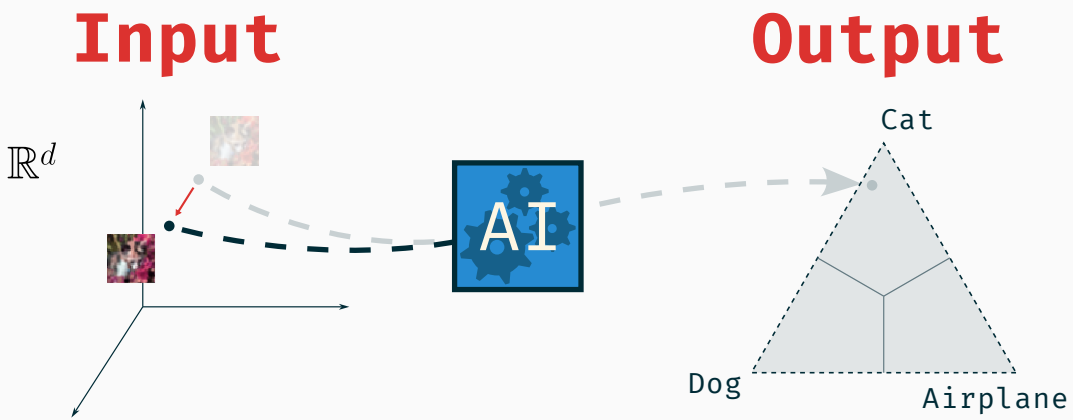
Consistent results

▷ robustness, controlled sensitivity...

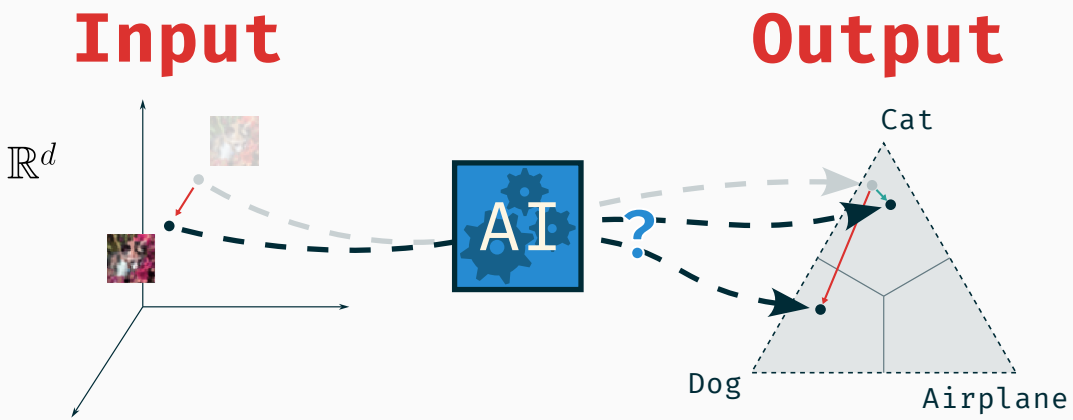
What do we mean by “consistent results”?



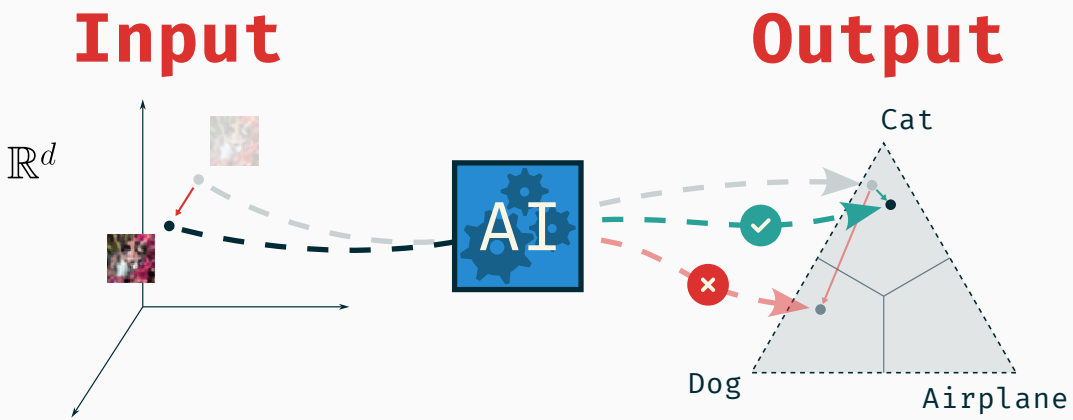
What do we mean by “consistent results”?



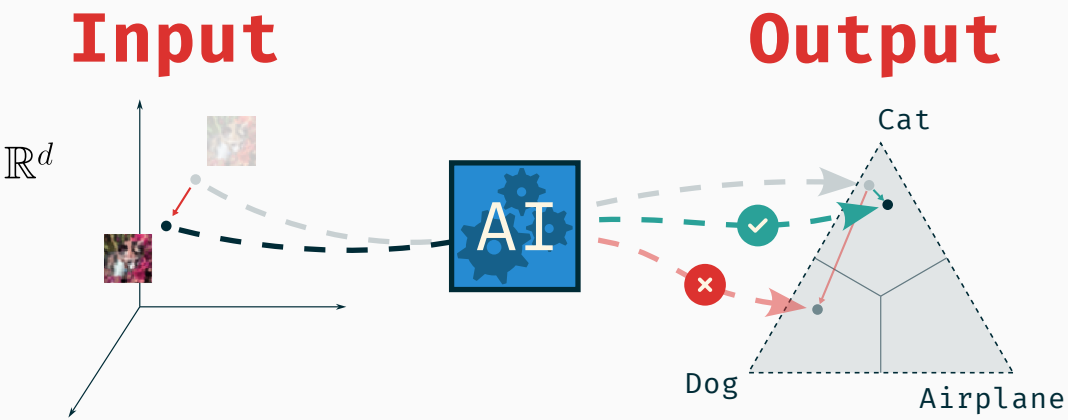
What do we mean by “consistent results”?



What do we mean by “consistent results”?



What do we mean by “consistent results”?



Changing the input slightly should not result in a significant change to the output.

In more mathematical terms?

Definition (Robustness)

Let d_{out} be a distance function on the output and d_{obs} one on the input. We say that N is *K-Lipschitz* if

$$\exists K > 0,\ \forall x_1, x_2, \quad d_{\text{out}}\left(N(x_1), N(x_2)\right) \leq K\, d_{\text{obs}}\left(x_1, x_2\right).$$

In more mathematical terms?

Definition (Robustness)

Let d_{out} be a distance function on the output and d_{obs} one on the input. We say that N is *K-Lipschitz* if

$$\exists K > 0,\ \forall x_1, x_2, \quad d_{\text{out}}\left(N(x_1), N(x_2)\right) \leq K\, d_{\text{obs}}\left(x_1, x_2\right).$$

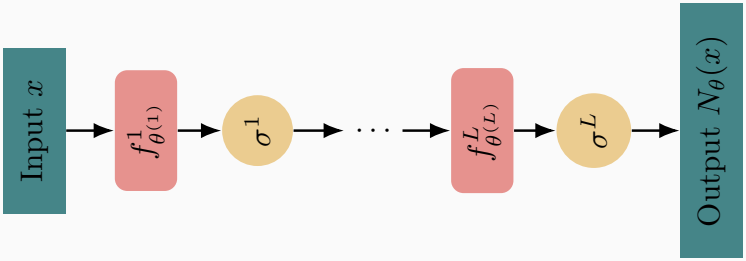
In more mathematical terms?

Definition (Robustness)

Let d_{out} be a distance function on the output and d_{obs} one on the input. We say that N is *K-Lipschitz* if

$$\exists K > 0,\ \forall x_1, x_2, \quad d_{\text{out}}\left(N(x_1), N(x_2)\right) \leq K\, d_{\text{obs}}\left(x_1, x_2\right).$$

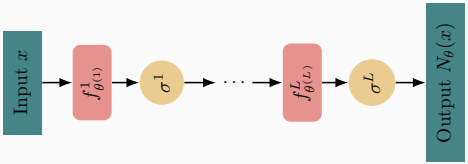
What is a neural network?



Definition (Neural network)

$$N_\theta : x \in \mathcal{X} \mapsto \sigma^L \circ f^L_{\theta^{(L)}} \circ \cdots \circ \sigma^1 \circ f^1_{\theta^{(1)}}(x) \in \mathcal{Y}.$$

What is a neural network?



Definition (Neural network)

$$N_\theta : x \in \mathcal{X} \mapsto \sigma^L \circ f^L_{\theta^{(L)}} \circ \cdots \circ \sigma^1 \circ f^1_{\theta^{(1)}}(x) \in \mathcal{Y}.$$

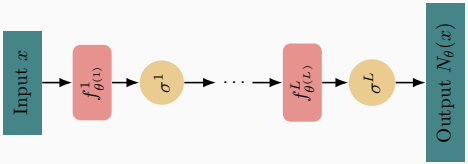
$\mathcal{X}$ = input space, $\mathcal{Y}$ = output space

$f^k_{\theta^{(k)}}$ = affine functions, convolution, ... (layers)

$\theta = \left(\theta^{(1)}, \ldots, \theta^{(L)}\right) \in \Theta \subseteq \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_L}$, *e.g.* weights and biases (parameters)

σ^k = GeLU, ReLU, or other non-linear maps (activation functions/non-linearities)

What is a neural network?



Definition (Neural network)

$$N_\theta : x \in \mathcal{X} \mapsto \sigma^L \circ f^L_{\theta^{(L)}} \circ \cdots \circ \sigma^1 \circ f^1_{\theta^{(1)}}(x) \in \mathcal{Y}.$$

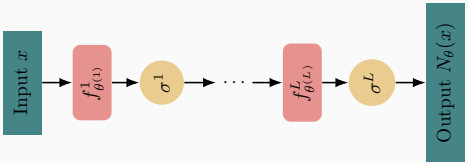
$\mathcal{X}$ = input space, $\mathcal{Y}$ = output space

$f^k_{\theta^{(k)}}$ = affine functions, convolution, ... (layers)

$\theta = \left(\theta^{(1)}, \ldots, \theta^{(L)}\right) \in \Theta \subseteq \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_L}$, *e.g.* weights and biases (parameters)

σ^k = GeLU, ReLU, or other non-linear maps (activation functions/non-linearities)

What is a neural network?



Definition (Neural network)

$$N_\theta : x \in \mathcal{X} \mapsto \sigma^L \circ f^L_{\theta^{(L)}} \circ \cdots \circ \sigma^1 \circ f^1_{\theta^{(1)}}(x) \in \mathcal{Y}.$$

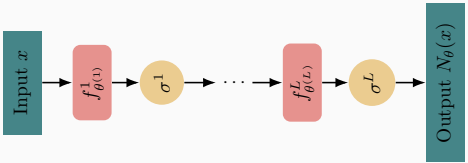
$\mathcal{X}$ = input space, $\mathcal{Y}$ = output space

$f^k_{\theta^{(k)}}$ = affine functions, convolution, ... (layers)

$\theta = \left(\theta^{(1)}, \ldots, \theta^{(L)}\right) \in \Theta \subseteq \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_L}$, *e.g.* weights and biases (parameters)

σ^k = GeLU, ReLU, or other non-linear maps (activation functions/non-linearities)

What is a neural network?



Definition (Neural network)

$$N_\theta : x \in \mathcal{X} \mapsto \sigma^L \circ f^L_{\theta^{(L)}} \circ \cdots \circ \sigma^1 \circ f^1_{\theta^{(1)}}(x) \in \mathcal{Y}.$$

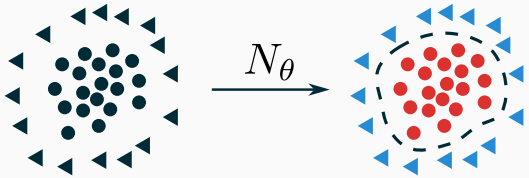
$\mathcal{X}$ = input space, $\mathcal{Y}$ = output space

$f^k_{\theta^{(k)}}$ = affine functions, convolution, ... (layers)

$\theta = \left(\theta^{(1)}, \ldots, \theta^{(L)}\right) \in \Theta \subseteq \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_L}$, *e.g.* weights and biases (parameters)

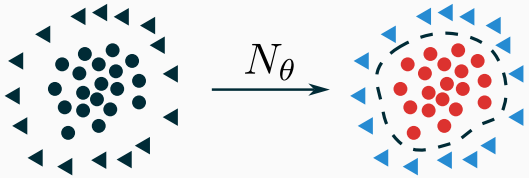
σ^k = GeLU, ReLU, or other non-linear maps (activation functions/non-linearities)

Popular problems in machine learning

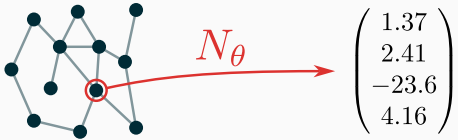


(a) Classification.

Popular problems in machine learning

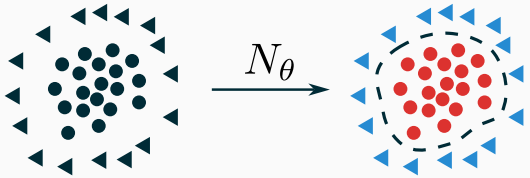


(a) Classification.

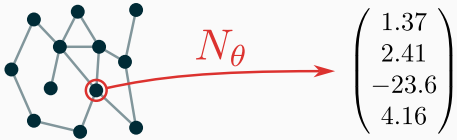


(b) Regression.

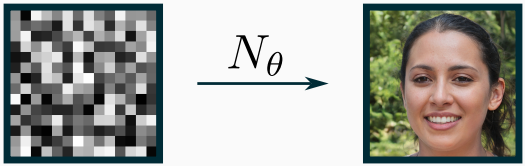
Popular problems in machine learning



(a) Classification.

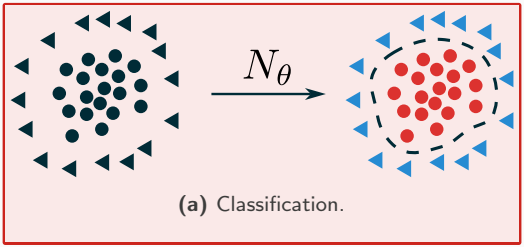


(b) Regression.

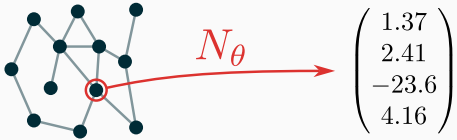


(c) Generation.

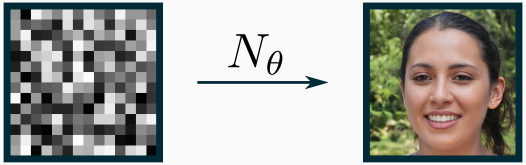
Popular problems in machine learning



(a) Classification.

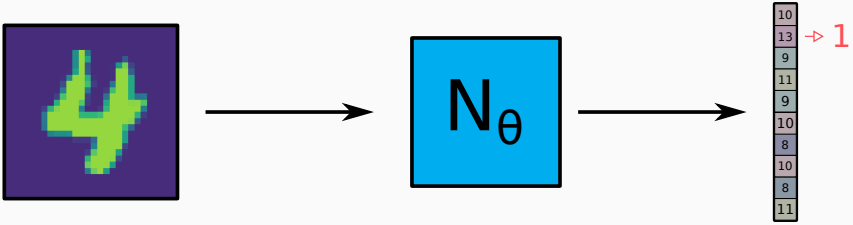


(b) Regression.



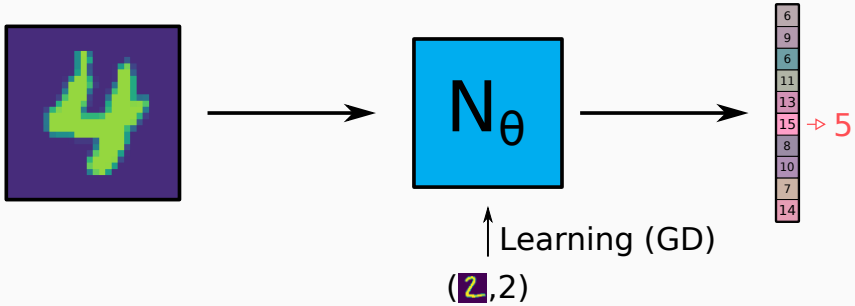
(c) Generation.

How does it work?

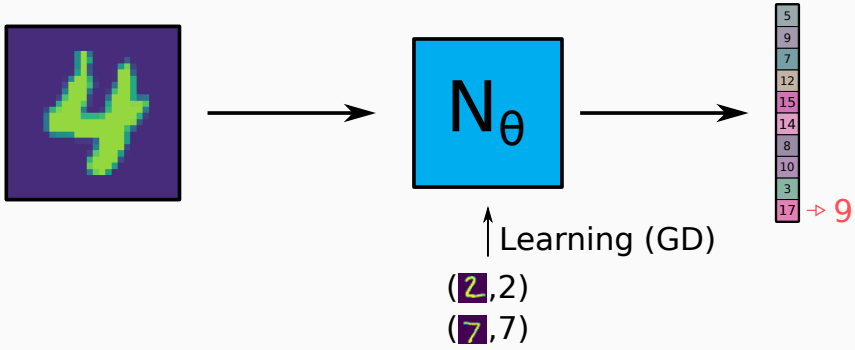


Step 0: Initialized at random.

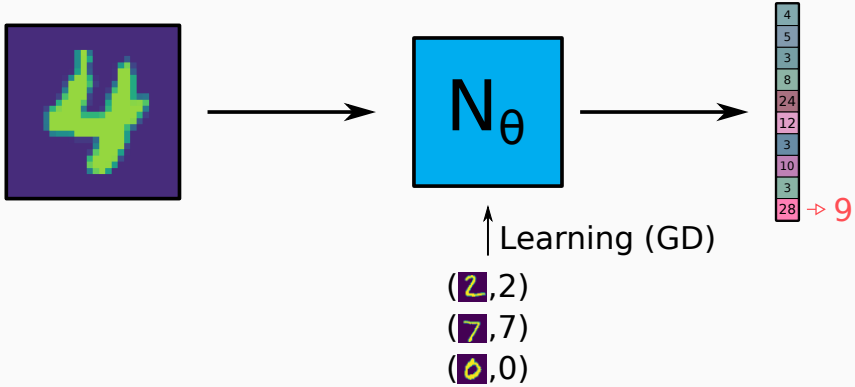
How does it work?



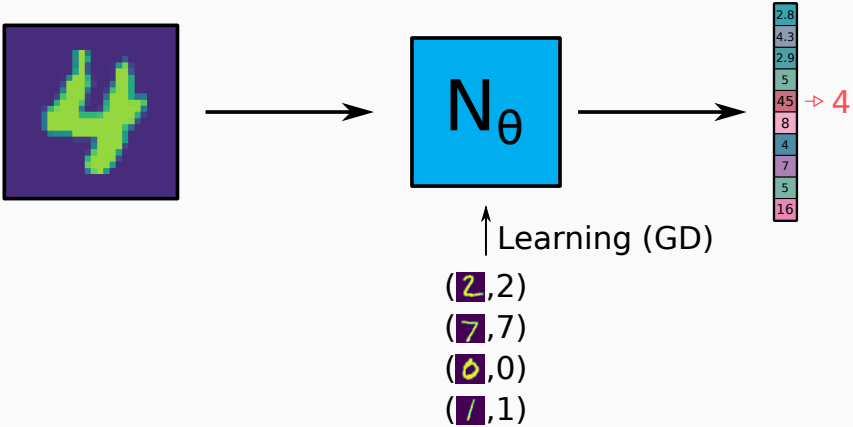
How does it work?



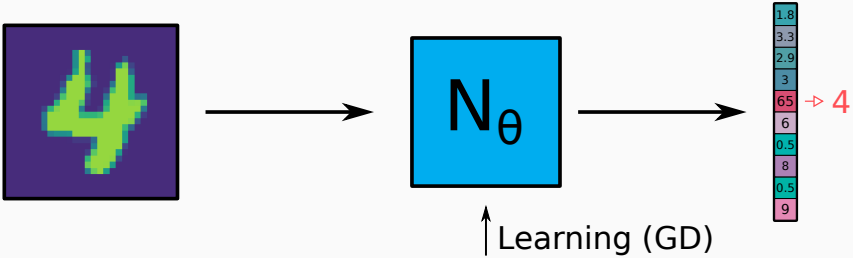
How does it work?



How does it work?



How does it work?



How does it work?

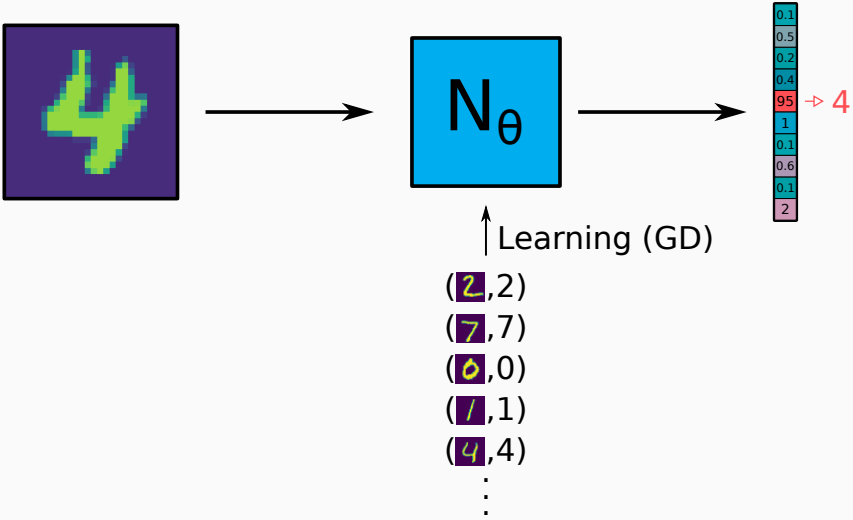


Table of content

Chapters

Table of content

Chapters

Table of content

Chapters

Table of content

Chapters

Table of content

Chapters

Table of content

Chapters

Table of content

Chapters

A Geometrical Framework for the Analysis of Neural Networks

(Chap. 2)

A Geometrical Framework for the Analysis of Neural Networks (Chap. 2)

A Simple Analogy

A Simple Analogy: Latitudes of the Sphere

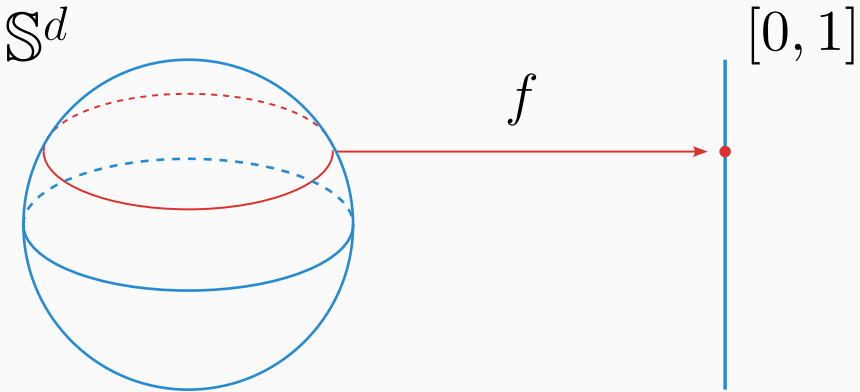
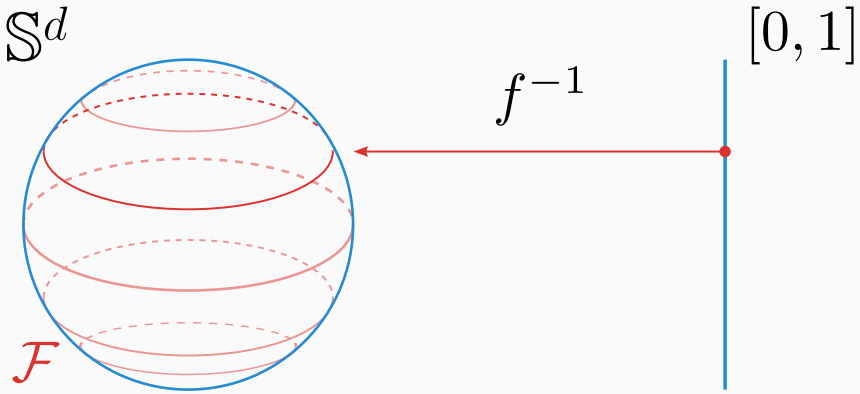


Figure 2: f maps each point of the sphere to its *latitude*.

An equivalence class: the concept of *foliation*



Foliation $\mathcal{F} = \{f^{-1}(y) \mid y \in [0,1]\}$.

What is a foliation?

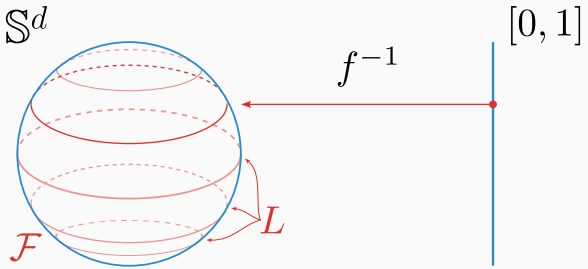
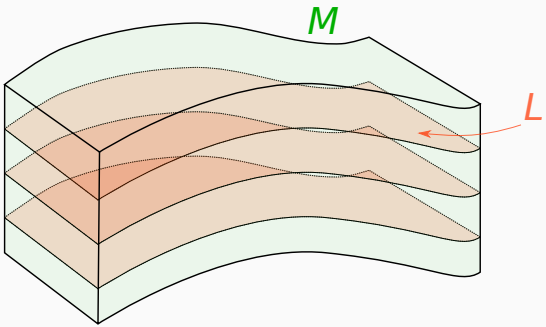


Figure 3: Foliation $\mathcal{F} = \{f^{-1}(y) \mid y \in [0,1]\}$.

Definition (Foliation)

Given a manifold $\mathcal{M}$, a *foliation* $\mathcal{F}$ is a partition of $\mathcal{M}$ into connected immersed submanifolds L called *leaves*.

What is a foliation?



Definition (Foliation)

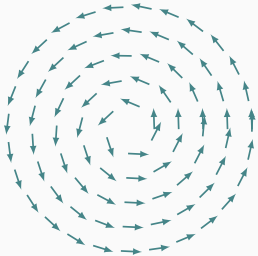
Given a manifold $\mathcal{M}$, a *foliation* $\mathcal{F}$ is a partition of $\mathcal{M}$ into connected immersed submanifolds L called *leaves*.

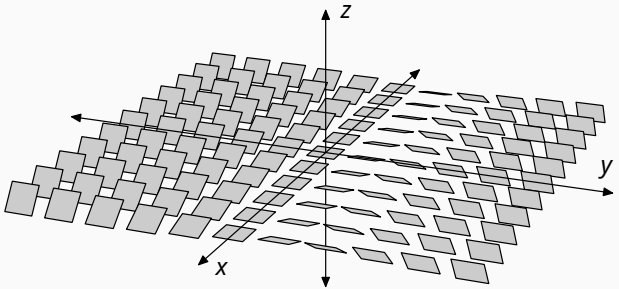
How can we construct foliations?

Theorem (Frobenius)

Let $D \subset T\mathcal{M}$ be a smooth distribution of constant rank, then the two following propositions are equivalent:

1. *Involutivity: for every two sections $X, Y \in \Gamma(D)$, $[X, Y]$ is also a section of D.*
2. *Integrability: there exists $\mathcal{F}$ foliation such that for all leaf $L \in \mathcal{F}$, $T_pL = D_p$ for all $p \in L$.*



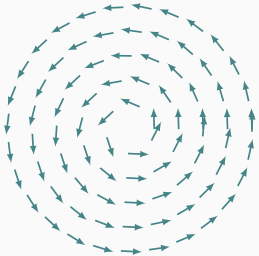


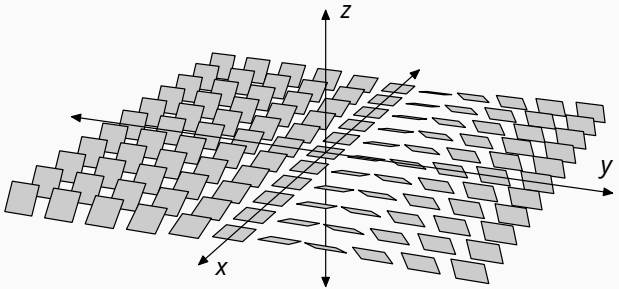
How can we construct foliations?

Theorem (Frobenius)

Let $D \subset T\mathcal{M}$ be a smooth distribution of constant rank, then the two following propositions are equivalent:

1. *Involutivity: for every two sections $X, Y \in \Gamma(D)$, $[X, Y]$ is also a section of D.*
2. *Integrability: there exists $\mathcal{F}$ foliation such that for all leaf $L \in \mathcal{F}$, $T_p L = D_p$ for all $p \in L$.*



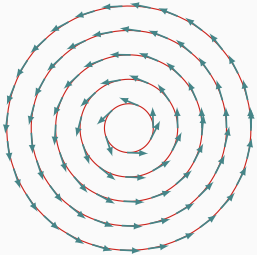


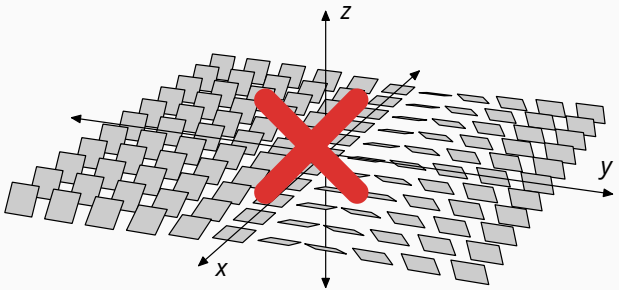
How can we construct foliations?

Theorem (Frobenius)

Let $D \subset T\mathcal{M}$ be a smooth distribution of constant rank, then the two following propositions are equivalent:

1. *Involutivity: for every two sections $X, Y \in \Gamma(D)$, $[X, Y]$ is also a section of D.*
2. *Integrability: there exists $\mathcal{F}$ foliation such that for all leaf $L \in \mathcal{F}$, $T_p L = D_p$ for all $p \in L$.*





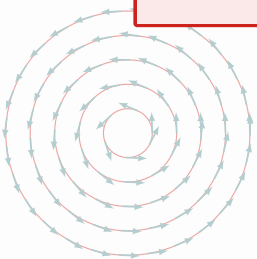
How can we construct foliations?

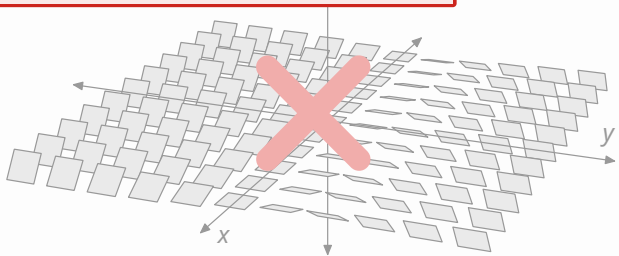
Theorem (Frobenius)

Let $D \subset T\mathcal{M}$ be a smooth distribution of constant rank, then the two following propositions are equivalent:

1. *Involutivity: for every two sections $X, Y \in \Gamma(D)$, $[X, Y]$ is also a section of D.*
2. *Integrability: there exists $\mathcal{F}$ foliation such that for all leaf $L \in \mathcal{F}$, $T_p L = D_p$ for all $p \in L$.*

A foliation (2.) is equivalent to its tangent spaces (1.).

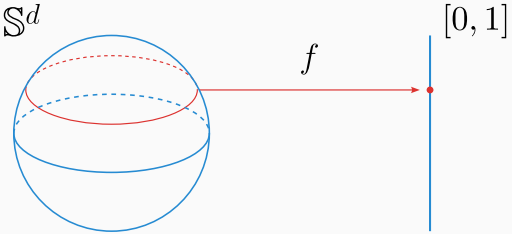


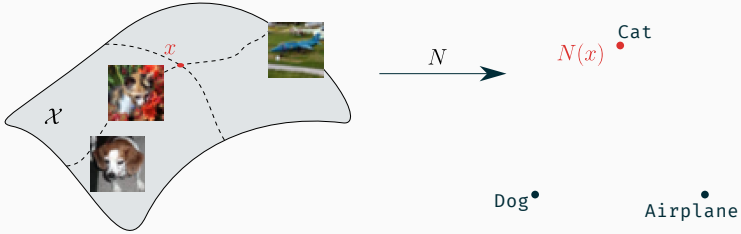


A Geometrical Framework for the Analysis of Neural Networks

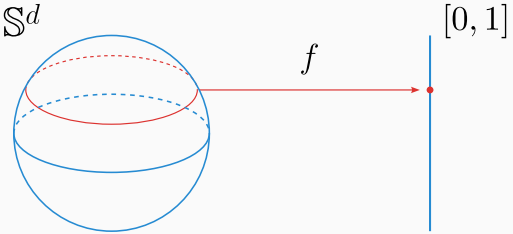
(Chap. 2)

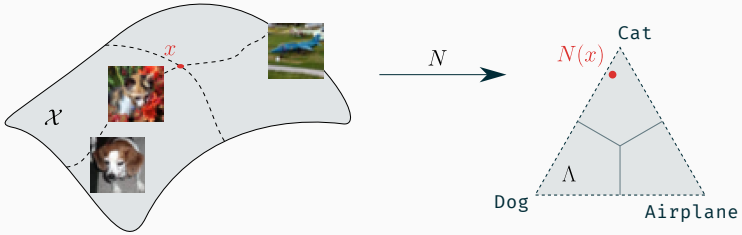
Back to Neural Networks





Discrete classification.





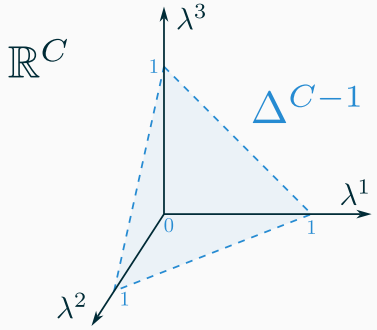
Soft classification[a].

[a] Szegedy et al. (2016)

What probability family for a classifier?

Definition (Probability simplex)

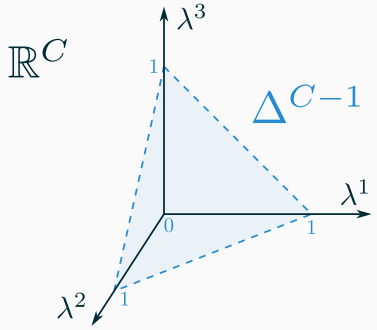
$$\Lambda = \Delta^{C-1} := \left\{ \lambda \in \mathbb{R}^C,\ \sum_{k=1}^{C} \lambda^k = 1,\ \lambda^k > 0\ \forall k \right\} \subset \mathbb{R}^C,$$



What probability family for a classifier?

Definition (Probability simplex)

$$\Lambda = \Delta^{C-1} := \left\{ \lambda \in \mathbb{R}^C,\ \sum_{k=1}^{C} \lambda^k = 1,\ \lambda^k > 0\ \forall k \right\} \subset \mathbb{R}^C,$$



$$\begin{cases} \mathcal{Y} = \{y_1, \ldots, y_C\}, \\ p(Y = y_k \mid \lambda) = \lambda^k := (N_\theta(x))^k. \end{cases}$$

$$\mathcal{X} \xrightarrow{N_\theta} \Lambda \xrightarrow{\cong} \mathcal{P} \xrightarrow{\text{sample}} \mathcal{Y}\ .$$

How do we measure the sensitivity of a probability?

$$D_{\mathrm{KL}}\left(P_{\lambda} \parallel P_{\lambda+v}\right)$$

How do we measure the sensitivity of a probability?

$$D_{\mathrm{KL}}\left(P_\lambda \parallel P_{\lambda+v}\right) = \frac{1}{2} v^t F(\lambda) v + o\left(\|v\|^2\right)$$

with $F(\lambda)$ a common tool of Information Geometry (see Amari (2016) and Nielsen (2020)):

Definition (Fisher information)

$$F_{i,j}(\lambda) := \mathbb{E}_{Y|\lambda}\left[\left(\frac{\partial}{\partial \lambda_i} \log p\left(Y \mid \lambda\right)\right)\left(\frac{\partial}{\partial \lambda_j} \log p\left(Y \mid \lambda\right)\right)\right].$$

How do we measure the sensitivity of a probability?

$$D_{\mathrm{KL}}\left(P_\lambda \parallel P_{\lambda+v}\right) = \frac{1}{2} v^t F(\lambda) v + o\left(\|v\|^2\right)$$

with $F(\lambda)$ a common tool of Information Geometry (see Amari (2016) and Nielsen (2020)):

Definition (Fisher information)

$$F_{i,j}(\lambda) := \mathbb{E}_{Y|\lambda}\left[\left(\frac{\partial}{\partial \lambda_i} \log p\left(Y \mid \lambda\right)\right)\left(\frac{\partial}{\partial \lambda_j} \log p\left(Y \mid \lambda\right)\right)\right].$$

Setting $\lambda := \theta$ the parameters of the network gives the Fisher Information Metric[a]:

$$F_{i,j}(\theta) := \mathbb{E}_{Y|\theta}\left[\left(\frac{\partial}{\partial \theta_i} \log p\left(Y \mid \theta\right)\right)\left(\frac{\partial}{\partial \theta_j} \log p\left(Y \mid \theta\right)\right)\right]. \tag{FIM}$$

[a] Amari (1998) and Sun and Nielsen (2025)

The Data Information Metric

Setting $\lambda := x$ the input of the network gives the Data Information Metric[1]:

$$D_{i,j}(x) := \mathbb{E}_{Y|x}\left[\left(\frac{\partial}{\partial x_i}\log p\left(Y \mid x\right)\right)\left(\frac{\partial}{\partial x_j}\log p\left(Y \mid x\right)\right)\right]. \qquad \text{(DIM)}$$

[1] Tron, Fioresi, et al. (2024)

The Data Information Metric

Setting $\lambda := x$ the input of the network gives the Data Information Metric[1]:

$$D_{i,j}(x) := \mathbb{E}_{Y|x}\left[\left(\frac{\partial}{\partial x_i}\log p\left(Y \mid x\right)\right)\left(\frac{\partial}{\partial x_j}\log p\left(Y \mid x\right)\right)\right]. \qquad \text{(DIM)}$$

Remark
The DIM is the pullback metric on $\mathcal{X}$ of the Fisher metric on $\mathcal{P}$ by the network N_θ.

[1] Tron, Fioresi, et al. (2024)

But why the DIM?

Pros of using the Data Information Metric:

- It measures distances between the output probabilities regarding **variations** in the input space.

But why the DIM?

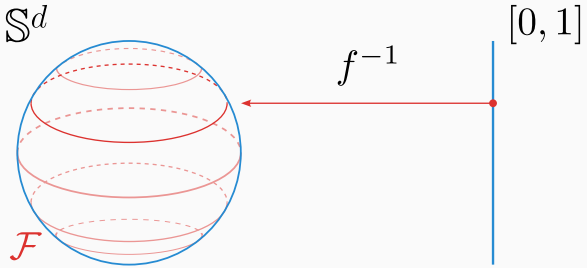
Pros of using the Data Information Metric:

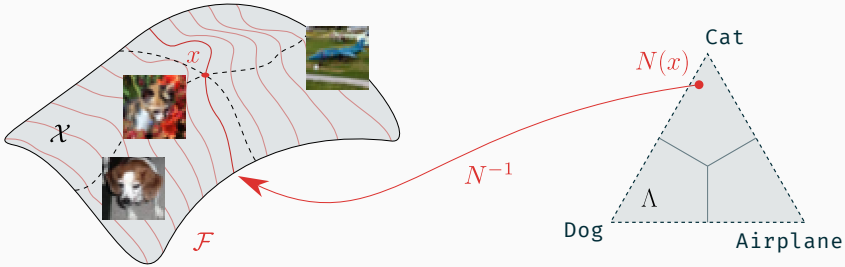
- It measures distances between the output probabilities regarding **variations** in the input space.
- Its link with the KL-divergence makes it easier to understand.

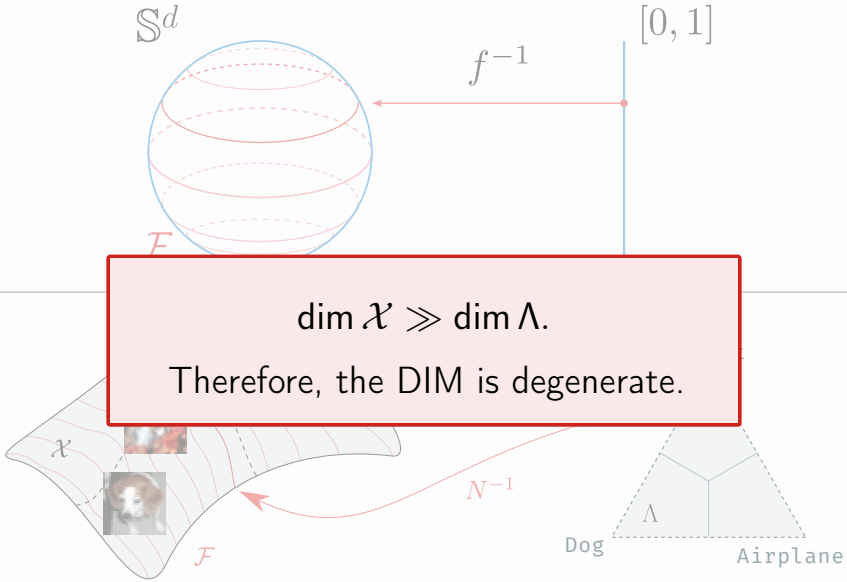
But why the DIM?

Pros of using the Data Information Metric:

- It measures distances between the output probabilities regarding **variations** in the input space.
- Its link with the KL-divergence makes it easier to understand.
- It yields a (symmetric) distance on the input space: the geodesic distance.


$\mathbb{S}^d$
$[0, 1]$
f^{-1}
$\mathcal{F}$


x
$\mathcal{X}$
$\mathcal{F}$
N^{-1}
$N(x)$
Cat
Λ
Dog
Airplane



$\dim \mathcal{X} \gg \dim \Lambda.$

Therefore, the DIM is degenerate.

The kernel foliation

Proposition

$\ker D$ *is an involutive distribution on* $\mathcal{X}$ *of codimension bounded by* $C - 1$*, and is thus integrable into a foliation called the kernel foliation.*

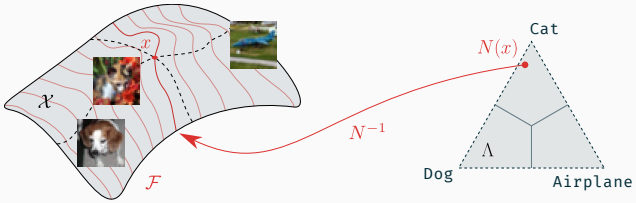
The kernel foliation

Proposition

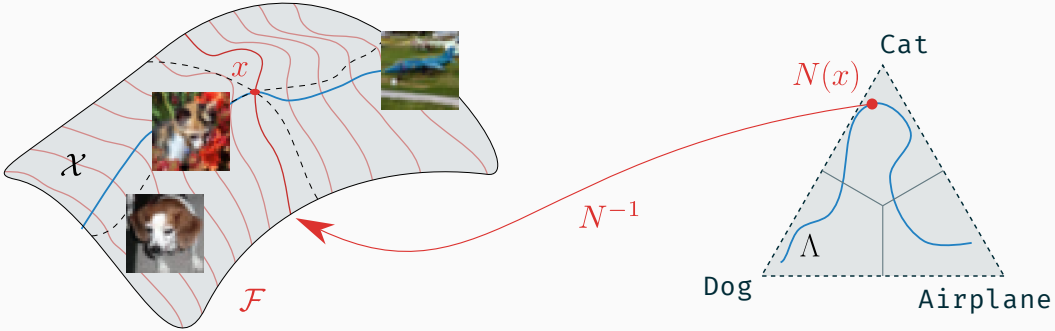
$\ker D$ *is an involutive distribution on* $\mathcal{X}$ *of codimension bounded by* $C - 1$*, and is thus integrable into a foliation called the kernel foliation.*

Proposition

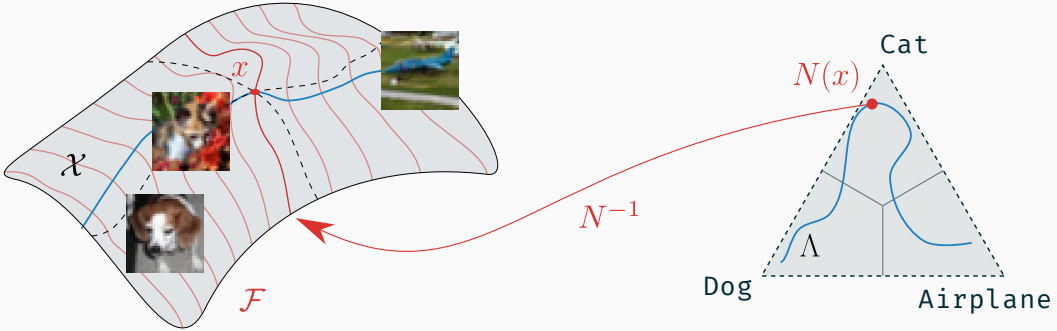
The leaves of the kernel foliation are the connected components of the inverse image by N_θ *of the points in* Λ*.*



Transverse metric to the kernel foliation



Transverse metric to the kernel foliation



To study robustness, we need to focus on $\mathcal{F}^{\perp}$, on which the DIM is full rank, but which might not be a foliation.

Key points

- We provided a geometrical framework to study the input/output relationship in a neural network.
- To study the robustness of a neural network, one needs to compare $d_{\mathrm{geo}}(x, y)$ and $d_{\mathrm{obs}}(x, y)$.

Key points

- We provided a geometrical framework to study the input/output relationship in a neural network.
- To study the robustness of a neural network, one needs to compare $d_{\text{geo}}(x, y)$ and $d_{\text{obs}}(x, y)$.

Next chapter

Tackle the limitations of this framework in the context of practical neural networks.

Non-Smoothness and Rank of the DIM: the Case of ReLU Networks

(Chap. 4)

A first problem: non-smoothness of ReLU networks

Definition (ReLU network)

$$N_\theta : x \in \mathcal{X} \mapsto \text{SoftMax} \circ \underbrace{f^L_{\theta^{(L)}} \circ \cdots \circ \sigma \circ f^1_{\theta^{(1)}}}_{=S \text{ the score}}(x) \in \Delta^{C-1}.$$

$f^k_{\theta^{(k)}}(a) = W^{(k)}a + b^{(k)}$ (linear layers)

$\sigma = \text{ReLU}$, or other piecewise-linear maps (activation functions/non-linearities)

S is a piecewise-affine map (score)

A first problem: non-smoothness of ReLU networks

Definition (ReLU network)

$$N_\theta : x \in \mathcal{X} \mapsto \mathrm{SoftMax} \circ \underbrace{f^L_{\theta^{(L)}} \circ \cdots \circ \sigma \circ f^1_{\theta^{(1)}}}_{=S \text{ the score}}(x) \in \Delta^{C-1}.$$

$f^k_{\theta^{(k)}}(a) = W^{(k)} a + b^{(k)}$ (linear layers)

$\sigma = \mathrm{ReLU}$, or other piecewise-linear maps (activation functions/non-linearities)

S is a piecewise-affine map (score)

A first problem: non-smoothness of ReLU networks

Definition (ReLU network)

$$N_\theta : x \in \mathcal{X} \mapsto \text{SoftMax} \circ \underbrace{f^L_{\theta^{(L)}} \circ \cdots \circ \sigma \circ f^1_{\theta^{(1)}}}_{=S \text{ the score}}(x) \in \Delta^{C-1}.$$

$f^k_{\theta^{(k)}}(a) = W^{(k)}a + b^{(k)}$ (linear layers)

$\sigma = \text{ReLU}$, or other piecewise-linear maps (activation functions/non-linearities)

S is a piecewise-affine map (score)

A first problem: non-smoothness of ReLU networks

Definition (ReLU network)

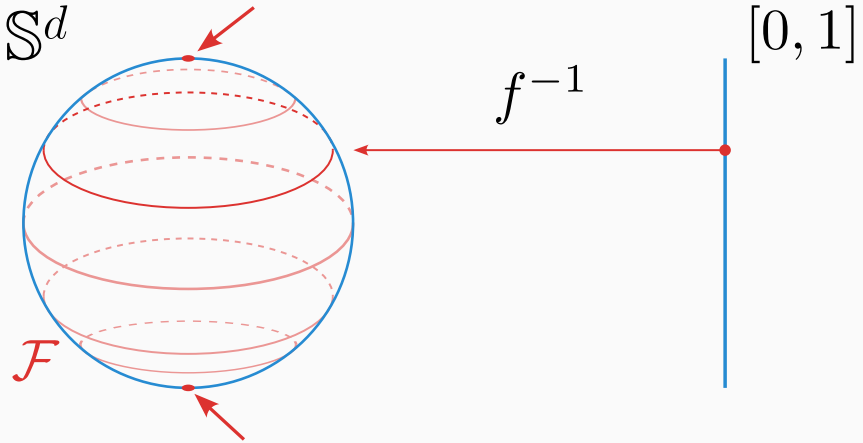
$$N_\theta : x \in \mathcal{X} \mapsto \text{SoftMax} \circ \underbrace{f^L_{\theta^{(L)}} \circ \cdots \circ \sigma \circ f^1_{\theta^{(1)}}}_{=S \text{ the score}}(x) \in \Delta^{C-1}.$$

$f^k_{\theta^{(k)}}(a) = W^{(k)}a + b^{(k)}$ (linear layers)

$\sigma = \text{ReLU}$, or other piecewise-linear maps (activation functions/non-linearities)

S is a piecewise-affine map (score)

A second problem: non-constant rank



In practice: the DIM is not of constant rank

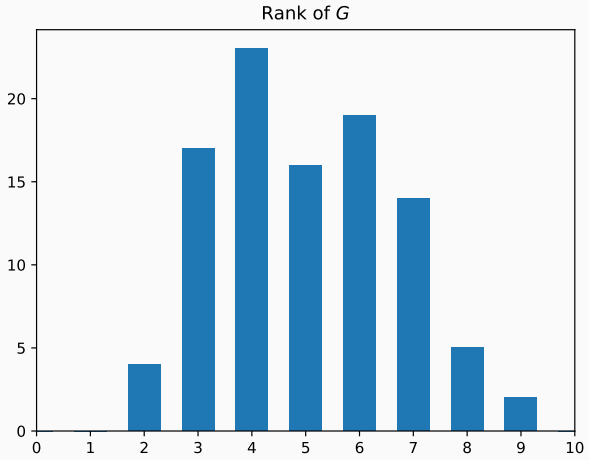


Figure 6: Distribution of the rank of the DIM, evaluated on 100 data points (MNIST).

Singularity and non smoothness: where?

Theorem (Tron and Fioresi (2024))

Consider the distribution:

$$\mathcal{D}_x := (\ker D_x)^{\perp} \subset T_x\mathcal{X}.$$

For a ReLU neural network as defined previously, then $\mathcal{D}$ is singular and non-smooth only on a closed null subset of $\mathbb{R}^d$ contained in a union of hypersurfaces.

Singularity and non smoothness: where?

Theorem (Tron and Fioresi (2024))

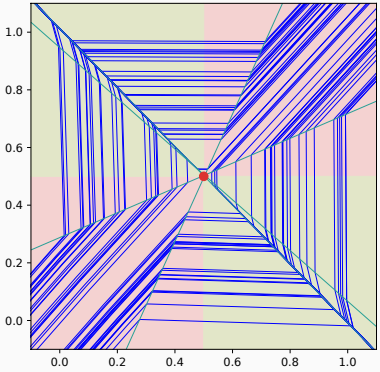
Consider the distribution:

$$\mathcal{D}_x := (\ker D_x)^{\perp} \subset T_x\mathcal{X}.$$

For a ReLU neural network as defined previously, then $\mathcal{D}$ is singular and non-smooth only on a closed null subset of $\mathbb{R}^d$ contained in a union of hypersurfaces.

Frobenius theorem is locally satisfied almost everywhere
$\implies$ the transverse foliation exists almost everywhere.

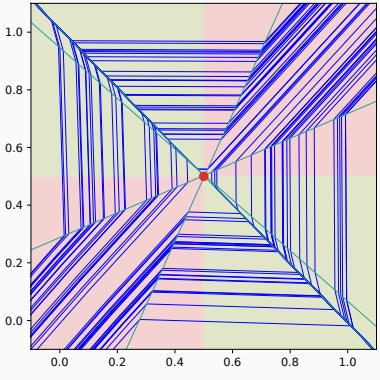
A concrete toy example: the Xor problem



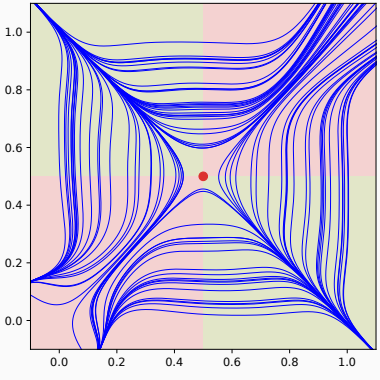
(a) ReLU.

Figure 7: The blue lines are a sample of the data foliation defined by the distribution $\mathcal{D}$ for a `Xor` network. The two classes of the `Xor` problem are represented in red and green squares underneath. The red dot is a singular point for the foliation. In (a), the green lines are the non-smooth points.

A concrete toy example: the Xor problem



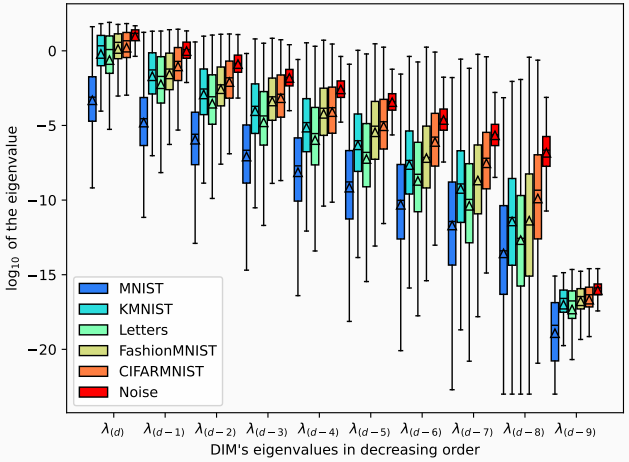
(a) ReLU.



(b) GeLU.

Figure 7: The blue lines are a sample of the data foliation defined by the distribution $\mathcal{D}$ for a `Xor` network. The two classes of the `Xor` problem are represented in red and green squares underneath. The red dot is a singular point for the foliation. In (a), the green lines are the non-smooth points.

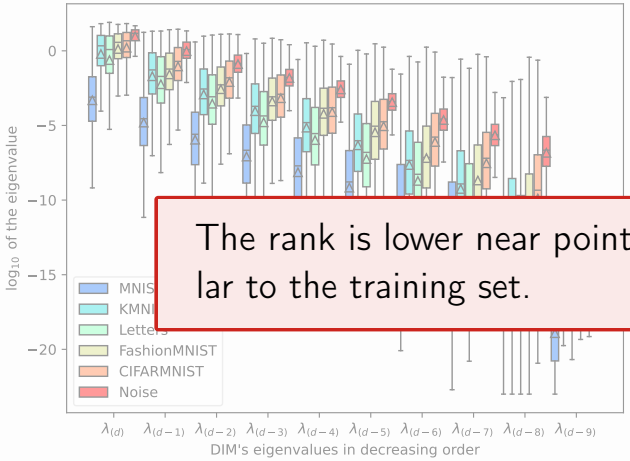
Where is the rank dropping when the network is trained on MNIST?



- N_θ trained on MNIST,
- D_x DIM computed on MNIST-like datasets,
- eigenvalues sorted.

Figure 8: DIM eigenvalues sorted by decreasing order evaluated on 10K points for each dataset.

Where is the rank dropping when the network is trained on MNIST?



The rank is lower near points similar to the training set.

- N_θ trained on MNIST,
- D_x DIM computed MNIST-like sets,
- nvalues sorted.

Figure 8: DIM eigenvalues sorted by decreasing order evaluated on 10K points for each dataset.

Proof of concept: dataset distance and knowledge transfer

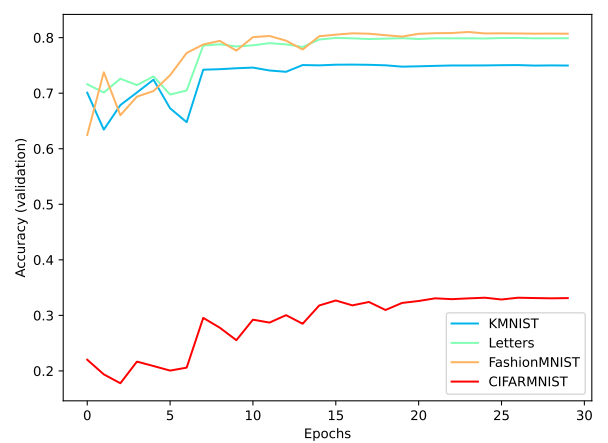


Figure 9: Accuracy after transfer learning starting from the weights of a ReLU network trained on MNIST (98% of accuracy) and retraining only the last linear layer.

Key points

- For ReLU networks, the transverse foliation exists almost everywhere.
- The rank of the DIM is closely related to the data the network was trained on.

Key points

- For ReLU networks, the transverse foliation exists almost everywhere.
- The rank of the DIM is closely related to the data the network was trained on.

Next chapter

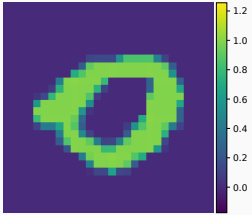
Apply this theoretical framework to study the robustness of a neural network in practice.

Application To Adversarial Attacks: the Importance of Curvature

What is our goal?

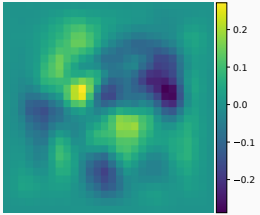
Given a budget ε, to construct the most harmful **adversarial attack**.

Test point
Predicted class: 0
Confidence: 100.0%



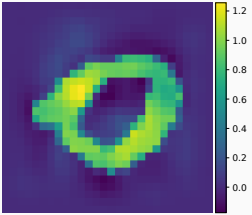
$+\varepsilon$

Attack vector



$=$

Attacked point
Predicted class: 7
Confidence: 91.5%



Can we use geometry?

Definition (Adversarial Attack)

An *adversarial attack* of budget $\varepsilon > 0$ is a solution of:

$$\max_{x_a \in \mathcal{X}} d_{\mathrm{nn}}\left(N_\theta\left(x_o\right), N_\theta\left(x_a\right)\right) \qquad \text{subject to} \qquad d_{\mathrm{obs}}\left(x_o, x_a\right) \leq \varepsilon. \qquad \text{(AAP)}$$

Can we use geometry?

Definition (Adversarial Attack)

An *adversarial attack* of budget $\varepsilon > 0$ is a solution of:

$$\max_{x_a \in \mathcal{X}} d_{\text{geo}}(x_o, x_a) \qquad \text{subject to} \qquad d_{\text{obs}}(x_o, x_a) \leq \varepsilon. \qquad \text{(AAP)}$$

Can we use geometry?

Definition (Adversarial Attack)

An *adversarial attack* of budget $\varepsilon > 0$ is a solution of:

$$\max_{x_a \in \mathcal{X}} d_{\text{geo}}(x_o, x_a) \qquad \text{subject to} \qquad d_{\text{obs}}(x_o, x_a) \leq \varepsilon. \qquad \text{(AAP)}$$

Can we use geometry?

Definition (Adversarial Attack)

An *adversarial attack* of budget $\varepsilon > 0$ is a solution of:

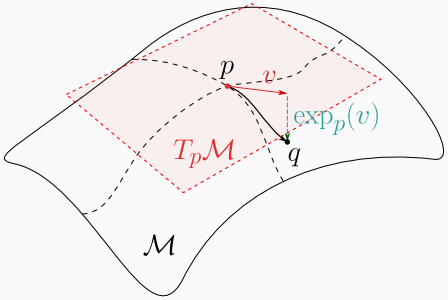
$$\max_{x_a \in \mathcal{X}} d_{\text{geo}}(x_o, x_a) \qquad \text{subject to} \qquad d_{L^2}(x_o, x_a) \leq \varepsilon. \qquad \text{(AAP)}$$

Reparametrization through the Riemannian exponential.

Definition (Adversarial Attack)

With $x_a = \exp_{x_o}(v)$, and $v = \log_{x_o}(x_a) \in T_{x_o}\mathcal{X}$,

$$\max_{v \in T_{x_o}\mathcal{X}} \|v\|_{\mathcal{X}}^2 \qquad \text{subject to} \qquad \left\|\exp_{x_o}(v) - x_o\right\|_2 \leq \varepsilon. \tag{AAP}$$



Reparametrization through the Riemannian exponential.

Definition (Adversarial Attack)

With $x_a = \exp_{x_o}(v)$, and $v = \log_{x_o}(x_a) \in T_{x_o}\mathcal{X}$,

$$\max_{v \in T_{x_o}\mathcal{X}} \|v\|_{\mathcal{X}}^2 \qquad \text{subject to} \qquad \left\|\exp_{x_o}(v) - x_o\right\|_2 \leq \varepsilon. \tag{AAP}$$

Question:

How do we solve this optimisation problem?

A One-Step Attack

By approximating the geodesic to the first order: $\exp_{x_o}(v) \approx x_o + v$, the problem becomes:

$$\max_{v \in T_{x_o}\mathcal{X}} \underbrace{\|v\|_{\mathcal{X}}^2}_{=v^T D_{x_o} v} \quad \text{subject to} \quad \|v\|_2 \leq \varepsilon.$$

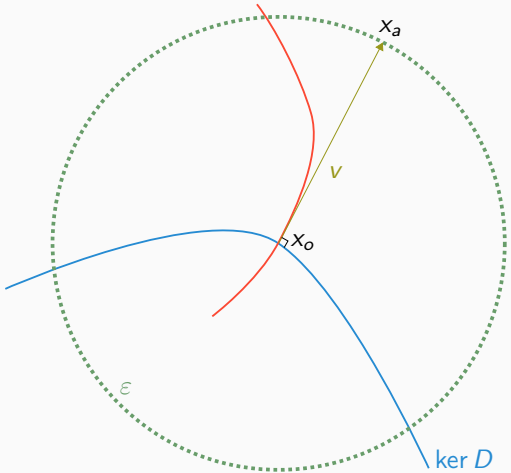
A One-Step Attack

By approximating the geodesic to the first order: $\exp_{x_o}(v) \approx x_o + v$, the problem becomes:

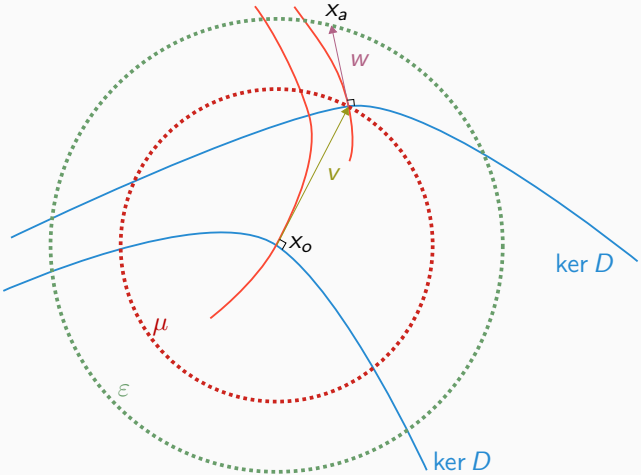
$$\max_{v \in T_{x_o}\mathcal{X}} \underbrace{\|v\|_{\mathcal{X}}^2}_{=v^T D_{x_o} v} \quad \text{subject to} \quad \|v\|_2 \leq \varepsilon.$$

Solution:
v is an eigenvector of D_{x_o} associated to the highest eigenvalue, and with Euclidean norm ε. This is the method presented by Zhao et al. (2019).

A One-Step Attack with a figure

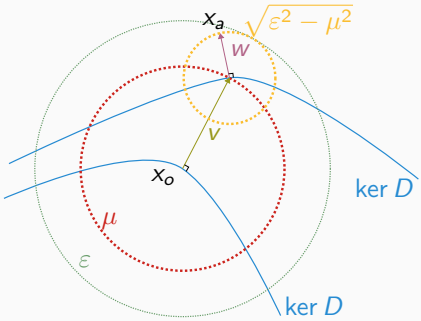


How can we improve? With a Two-Step Attack!



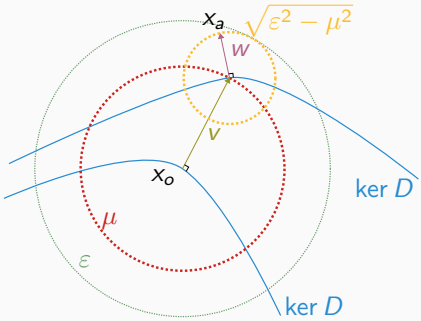
How can we improve? With a Two-Step Attack!

$$\max_{w} \|w\|_{\mathcal{X}}^2 \text{ subject to } \begin{cases} \|v\|_2 + \|w\|_2 \leq \varepsilon \\ \|v\|_2 = \mu < \varepsilon \\ v \text{ eigenvector of } D_{x_o} \end{cases}$$



How can we improve? With a Two-Step Attack!

$$\max_{w} \|w\|_{\mathcal{X}}^2 \text{ subject to } \begin{cases} \|v\|_2 + \|w\|_2 \leq \varepsilon \\ \|v\|_2 = \mu < \varepsilon \\ v \text{ eigenvector of } D_{x_o} \end{cases}$$



Solution:
w is an eigenvector of D_{x_o+v} associated to the highest eigenvalue, and with Euclidean norm $\sqrt{\varepsilon^2 - \mu^2}$.

Is two-step better in practice?

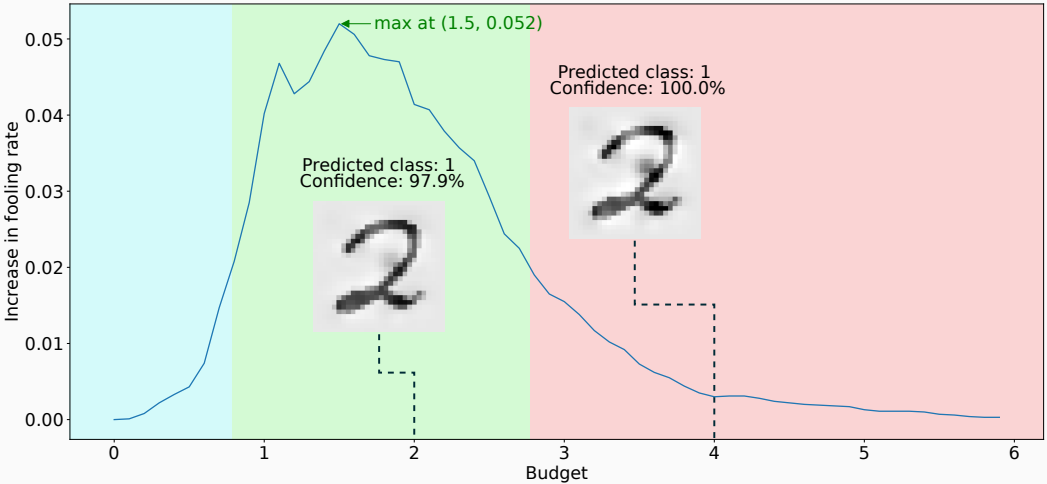


Figure 10: Difference between TSSA and OSSA.

Are we better than SOTA attacks?
The case of AutoAttack by Croce and Hein (2020).

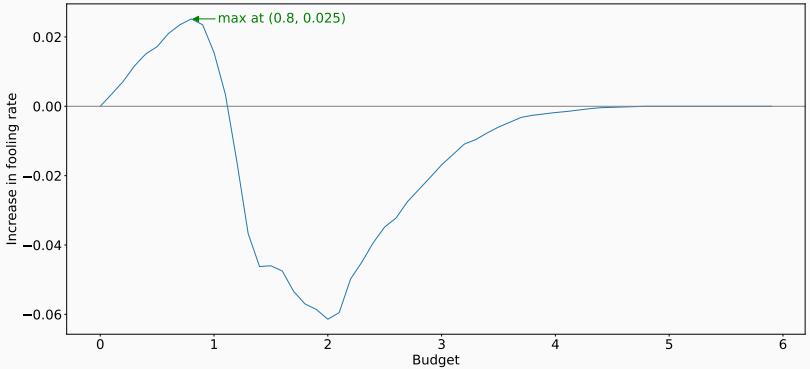


Figure 11: Difference between the fooling rate on MNIST of the TSSA (fr_{TSSA}) and the one of AutoAttack (fr_{AA}) with respect to the Euclidean budget ($\text{fr}_{\text{TSSA}} - \text{fr}_{\text{AA}}$).

Conclusion

Application To Adversarial Attacks: the Importance of Curvature

Key points

- Curvature plays a role in the efficiency of adversarial attacks.
- Our geometric framework successfully enlightened one of the component of robustness.

Key points

- Curvature plays a role in the efficiency of adversarial attacks.
- Our geometric framework successfully enlightened one of the component of robustness.

Application to robustness

- Regularize training through geometry to force low curvature models, aiming for more robustness.
- Use our adversarial examples for adversarial training.

Conclusion and perspectives

Conclusion and perspectives

Conclusions

- We propose a geometrical framework based on the DIM that tries to answer many questions:
  - Robustness ;
  - Explainability ;
  - Dataset distances ;
- We implemented this framework in Python, openly available online[2].

Perspectives

- This work is still preliminary and the framework can be improved.
- Many questions remains:
  - What transverse?
  - Different attacks with geodesics?
  - What link to the manifold hypothesis?
  - What are piecewise smooth foliations?
  - Variations/regularisations of the DIM?
  - What changes during training?

[2] https://github.com/eliot-tron/CurvNetAttack

Conclusion and perspectives

Conclusions

- We propose a geometrical framework based on the DIM that tries to answer many questions:
  - Robustness ;
  - Explainability ;
  - Dataset distances ;
- We implemented this framework in Python, openly available online[2].

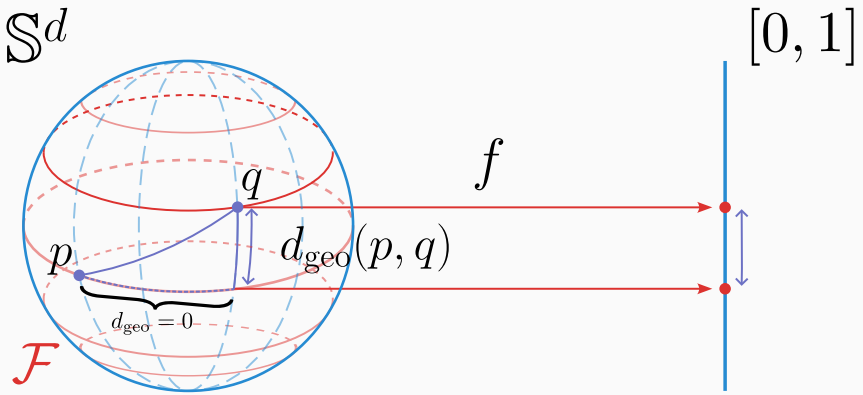
Perspectives

- This work is still preliminary and the framework can be improved.
- Many questions remains:
  - What transverse?
  - Different attacks with geodesics?
  - What link to the manifold hypothesis?
  - What are piecewise smooth foliations?
  - Variations/regularisations of the DIM?
  - What changes during training?

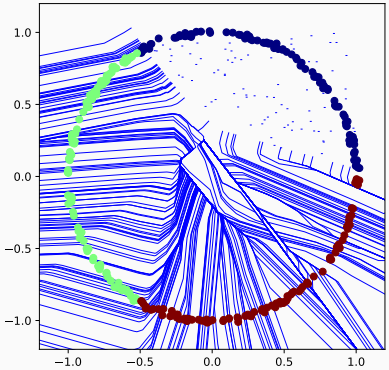
[2] https://github.com/eliot-tron/CurvNetAttack

Thank you for your attention.

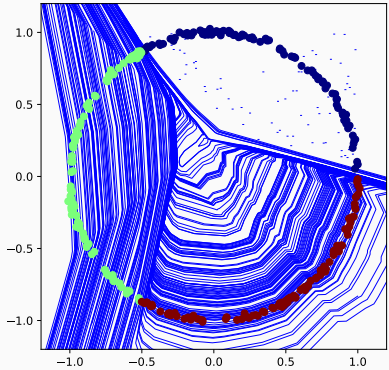
Geodesic distance and kernel



Our geometry is not the manifold hypothesis geometry



(a) Transverse foliation



(b) Kernel foliation

Other benefits of using the TSSA?

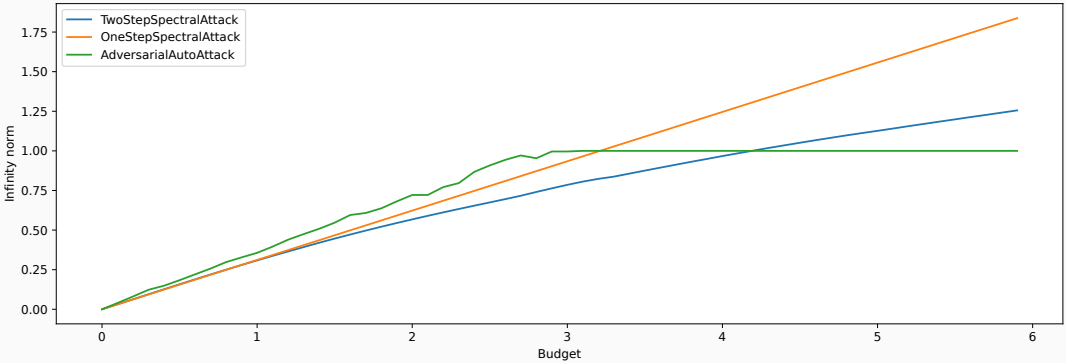


Figure 13: $\|\cdot\|_\infty$ of the attack with respect to its Euclidean norm $\|\cdot\|_2$ on MNIST.