Beyond the scope of **Manifold Learning**: the importance of the **Data Foliation** to understand **classifiers** and **datasets**.

Manifold Learning via Foliations and Knowledge Transfer

Eliot TRON, Rita FIORESI

1 Modeling

- A ReLU Neural Network (i.e. $\sigma = \mathrm{ReLU}$)

$$N_\theta \;:\; \mathbb{R}^d \longrightarrow \Delta^{C-1} := \left\{p \in \mathbb{R}^C \mid \textstyle\sum_{i=1}^C p_i = 1,\; p_i > 0\right\}$$
$$x \longmapsto \mathrm{SoftMax} \circ L_{(W_N, b_N)} \circ \sigma \circ \cdots \circ \sigma \circ L_{(W_1, b_1)}(x)$$

- associated with the probability distribution $p_\theta(y \mid x, \theta) = (N_\theta(x))_y$
- equipped with the Data Information Matrix (DIM):

$$D(x, \theta) := \mathbb{E}_{y \sim p}[\nabla_x \log p(y|x, \theta) \cdot (\nabla_x \log p(y|x, \theta))^T].$$

2 Manifold Learning

Manifold Hypothesis:

> "The manifold hypothesis posits that many high-dimensional data sets that occur in the real world actually lie along a low-dimensional latent manifold $\mathcal{M}$ inside that high-dimensional space." – Wikipedia

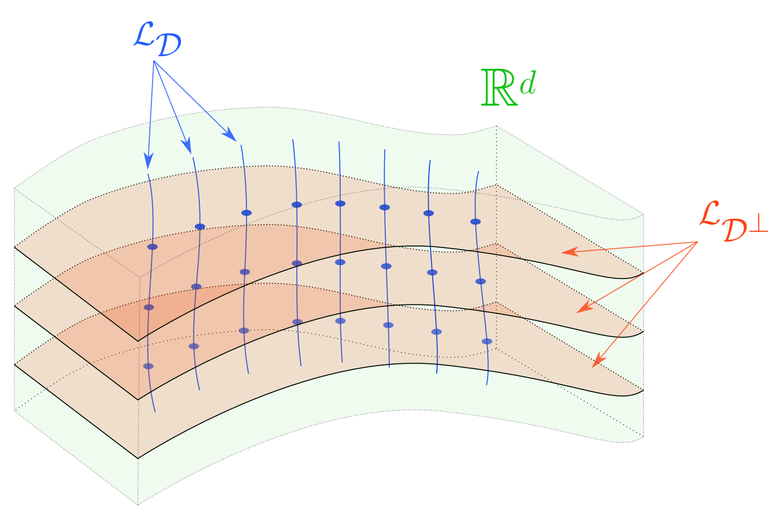
The goal of *Manifold Learning* is to learn $\mathcal{M}$ or any meaningful geometric structure correlated to the sampled data points that are given to us.

3 Singular Foliations

Definition 3.1. We define a distribution $\mathcal{D}$ from the columns of the DIM as such: $\mathbb{R}^d \ni x \mapsto \mathcal{D}_x := \mathrm{span}\{\nabla_x p_i(y|x, w),\ i = 1, \dots c\}$. In the case of a ReLU network, this distribution is involutive on the points where it is well-defined.

Theorem 3.1. *Consider the distribution $\mathcal{D}$ for a ReLU Neural Network. Then, its singular points and non-smooth points are a closed null subset of $\mathbb{R}^d$ contained in the union of hypersurfaces.*

Corollary 3.1.1. *Frobenius theorem gives the existence of a data foliation associated to $\mathcal{D}$ almost everywhere on $\mathbb{R}^d$.*



4 Results

Remark. *The data foliation is spanned by the columns of the DIM. Therefore, the eigenvalues of $D(x, \theta)$ give great information on the nature of the leaf at x (e.g. its dimension).*

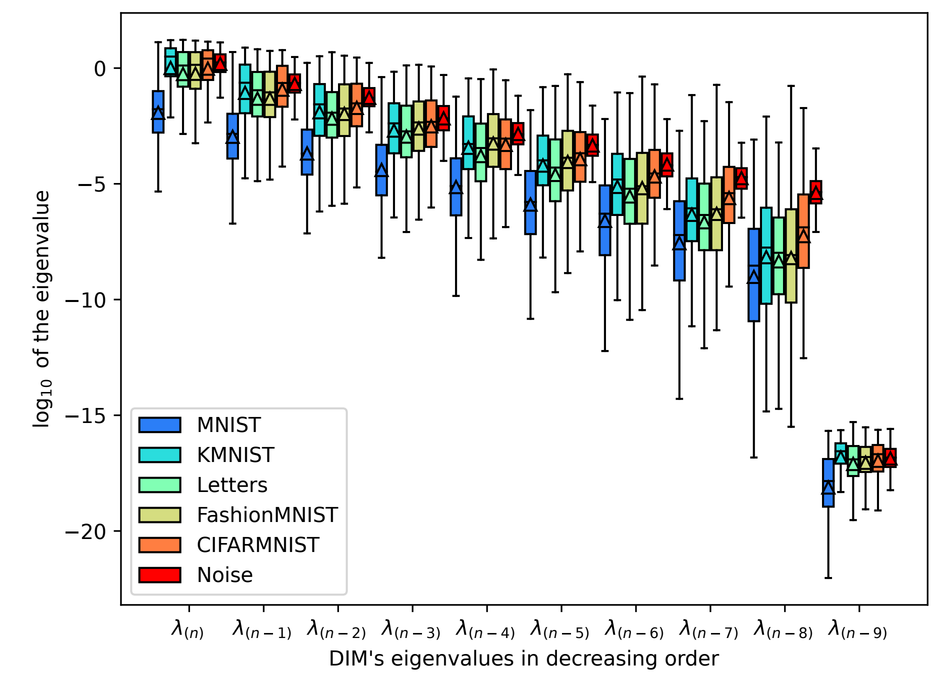


Figure 1: DIM eigenvalues sorted by decreasing order evaluated on 250 points for each dataset.

Interpretation: The DIM, and thus the Data Foliation, is correlated with the dataset N_θ was trained on (lower eigenvalues on average).

Knowledge Transfer

We train N_θ on the MNIST dataset (pictures of digits 0 to 9), then freeze the weights $W_1, b_1, \dots, W_{N-1}, b_{N-1}$ and retrain only W_N, b_N on a new dataset.

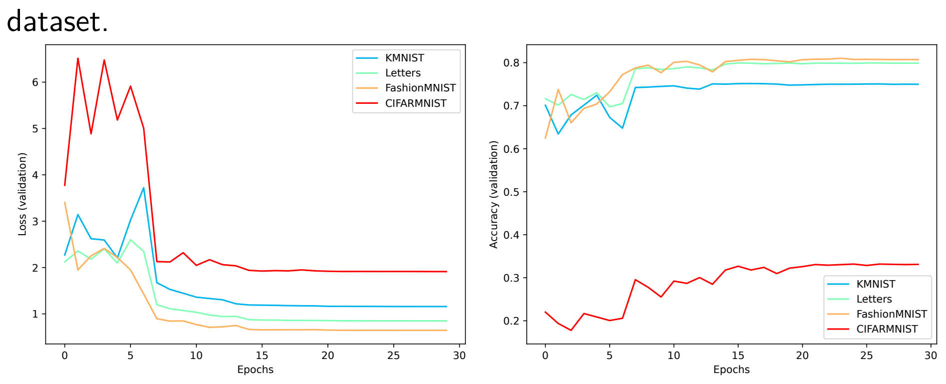


Figure 2: Loss and accuracy after transfer learning starting from the weights of a ReLU network trained on MNIST (98% of accuracy) and retraining only the last linear layer.

Interpretation: The final accuracy can be correlated with the median of the lowest non-zero eigenvalue ($\lambda_{(n-8)}$). The lower this is, the higher the accuracy. Thus, the rank of the DIM seems to correlate with the similarity between the data sets.

Extra figures

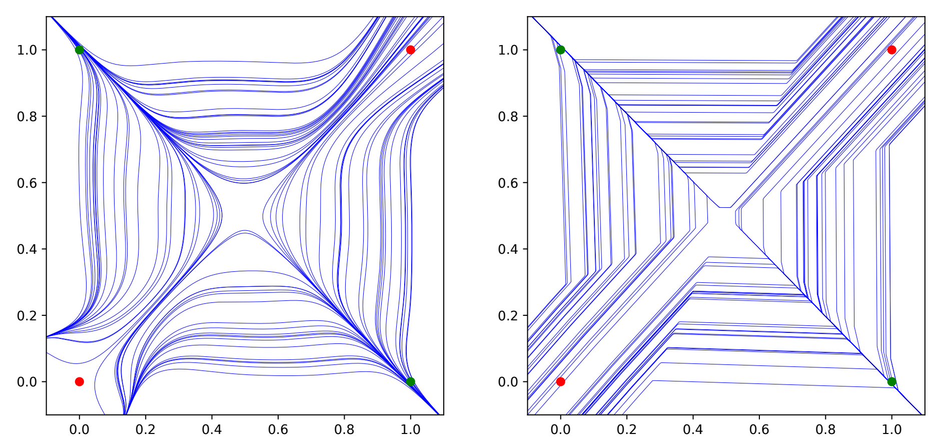


Figure 3: The Data Foliation defined by the distribution $\mathcal{D}$ for the Xor problem (left: GeLU, right: ReLU).

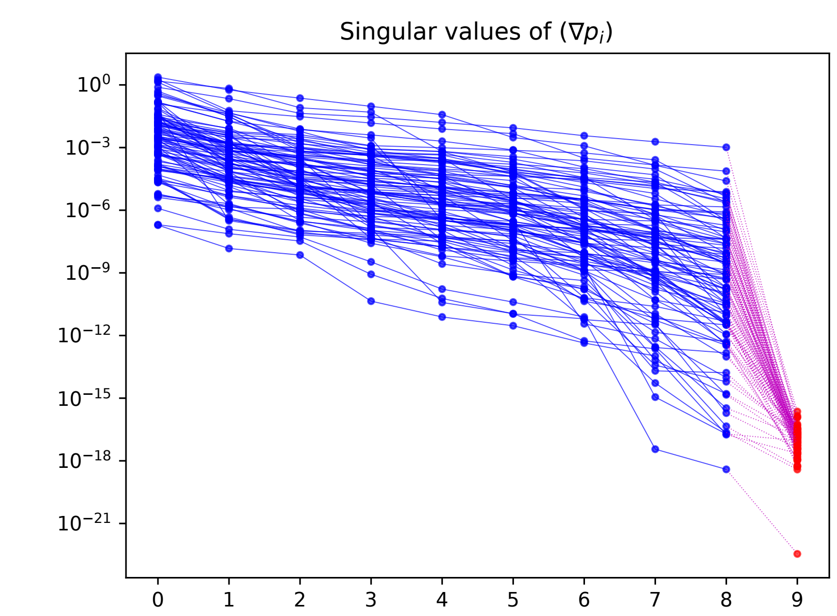


Figure 4: Singular values of $(\nabla p_i)_i$ ranked from highest to lowest, on 100 data points (MNIST). Each line corresponds to one picture.

Table 1: Parameters for Knowledge Transfer (logarithmic scale)

Dataset	Highest evalue	Lowest evalue	Val. Acc.
MNIST	-1.78	-8.58	98%
KMNIST	0.49	-7.75	75%
Letters	0.11	-7.99	80%
Fashion-MNIST	0.14	-8.08	81%
CIFARMNIST	0.41	-6.90	33%
Noise	0.24	-5.36	NA

Table 2: Involutivity of the distribution $\mathcal{D}$

Non linearity	dim $\mathcal{D}_x$	dim $\mathcal{V}_x^\mathcal{D}$
ReLU	9	9
GeLU	9	44.84
Sigmoid	9	45

$$\mathcal{V}_x^D := \mathrm{Span}\left\{X,\ [Y, Z] \mid X, Y, Z \in \mathcal{D}_x^3\right\}$$
$$= \mathrm{Span}\left\{\nabla_x \log p_i, [\nabla_x \log p_j, \nabla_x \log p_k] \mid i, j, k = 1, \dots, c\right\}.$$

RÉPUBLIQUE FRANÇAISE

Liberté
Égalité
Fraternité

ENAC