The Fisher-Rao Loss for Learning under Label Noise

Henrique K. Miyamoto

hmiyamoto@ime.unicamp.br

jointly with Fábio C. C. Meneghetti and Sueli I. R. Costa

University of Campinas (Unicamp), Brazil

Partially supported by FAPESP 21/04516-8

September 20, 2022

Context: supervised learning

- Training a classifier (e.g., a neural network) can be done by empirical risk minimisation of a *loss function*.
- Choosing a suitable loss function is important, as it affects the performance of the resulting classifier and the training dynamics.
  - The study and design of loss functions has been a topic of interest [Gol13, Fro15, Jan17, Dem20, Hui21].
- Important case: *label noise*, i.e., some labels of the dataset are incorrect. An efficient way to mitigate this problem is to use loss functions that are inherently robust to label noise [Gho15].
- We study the *Fisher-Rao loss function*, derived by an information-geometric approach, especially in the case of label noise.

Context: information geometry and learning

- The celebrated natural gradient method exploits the geometry of the so-called neuromanifolds, parametrised by the network parameters [Ama98].
- Fisher-Rao distances have been used in unsupervised learning for shape clustering, clustering financial returns and image segmentation [Gat17, Tay19, Pin20].
- They have been used as a regulariser term for adversarial learning, and to study the geometry of the latent space of generative models [Pic22, Arv22].
- Here we use the Fisher-Rao distance of the manifold of discrete distributions as a loss function on its own in a standard classification framework.

Information geometry preliminaries

Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a σ-finite measure space and P a probability measure on it. A statistical model

$$M := \left\{ p_\theta \mid \theta = (\theta^1, \ldots, \theta^n) \in \Theta \subset \mathbb{R}^n \right\}$$

is a parametric family of densities $p_\theta = \frac{\mathrm{d}P}{\mathrm{d}\mu} \colon \mathcal{X} \to \mathbb{R}_+$. If M is smoothly parametrised by $\theta \in \Theta$ and satisfies certain regularities conditions, then becomes a smooth manifold, known as *statistical manifold* [Ama00].

It is possible to equip M with a Riemannian structure with the *Fisher metric*, given in matrix form as $G_\theta = [g_{ij}(\theta)]_{ij}$, with

$$g_{ij}(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial \theta^i} \log p_\theta \right) \left(\frac{\partial}{\partial \theta^j} \log p_\theta \right) \right].$$

The Fisher metric provides a 'natural' choice of geometry, since it is essentially the unique metric that is invariant under sufficient statistics, [Ay17].

Information geometry preliminaries

A curve $\gamma\colon [0,1] \to \Theta$ defines a curve $p_{\gamma(t)}$ in M. Its length can be computed as

$$l(\gamma) \coloneqq \int_0^1 \sqrt{\|\dot{\gamma}(t)\|_G}\,\mathrm{d}t = \int_0^1 \sqrt{\dot{\gamma}(t)^\top G_{\gamma(t)} \dot{\gamma}(t)}\,\mathrm{d}t.$$

The *Fisher-Rao* distance is defined as the infimum of the length of piecewise smooth paths linking p_{θ_1} and p_{θ_2} (geodesic length):

$$d_{\mathrm{FR}}(p_{\theta_1}, p_{\theta_2}) \coloneqq d_{\mathrm{FR}}(\theta_1, \theta_2) \coloneqq \inf_{\gamma} \left\{ l(\gamma) \mid \gamma(0) = \theta_1, \gamma(1) = \theta_2 \right\}.$$

Closed-form expressions for the Fisher-Rao distance are only known for particular cases [Atk81].

Manifold of discrete distributions

Let $\mathcal{X} = \{1, 2, \ldots, K\}$ and $\delta^i \colon \mathcal{X} \to \{0, 1\}$ given by $\delta^i(j) = \delta_{ij}$.
The statistical manifold

$$M = \left\{ p = \sum_{i=1}^{K} p_i \delta^i \mid p_i \in [0,1]\,,\ \sum_{i=1}^{K} p_i = 1 \right\}$$

is in correspondence with the probability simplex

$$\Delta^{K-1} = \left\{ \boldsymbol{p} = (p_1, \ldots, p_K) \mid p_i \in [0,1]\,,\ \sum_{i=1}^{K} p_i = 1 \right\}$$

and both can be parametrised by the set

$$\Theta = \left\{ \theta = (\theta^1, \ldots, \theta^{K-1}) \mid \theta^i \geq 0,\ \sum_{i=1}^{K-1} p_i \geq 1 \right\},$$

with $p_i = \theta^i$, $1 \leq i \leq K-1$ and $p_K = 1 - \sum_{i=1}^{K-1} \theta^i$.

Manifold of discrete distributions

The Fisher metric in this manifold is given by

$$g_{ij}(\xi) = \frac{\delta_{ij}}{\theta^i} + \frac{1}{1 - \sum_{k=1}^{n-1} \theta^k}.$$

An easier way to obtain the geodesics is through the isometry

$$\begin{aligned} \pi\colon \quad M &\to S^{n-1}_{2,+} \\ p = \textstyle\sum_i p_i \delta^i &\mapsto (2\sqrt{p_1}, \ldots, 2\sqrt{p_n}) =: (z_1, \ldots, z_n) \end{aligned}$$

from the statistical manifold with the Fisher metric to the positive part of the radius-two sphere $S^{n-1}_{2,+}$ with the Euclidean metric.

Thus the Fisher metric in M is essentially the spherical metric.

Manifold of discrete distributions

The geodesics on the sphere are arcs of great circles. Thus the distance between two points $\boldsymbol{z}_p, \boldsymbol{z}_q$ on $S_{2,+}^{n-1}$ is double the angle α between them:

$$2\alpha = 2\arccos\left\langle \frac{\boldsymbol{z}_p}{2}, \frac{\boldsymbol{z}_q}{2} \right\rangle = 2\arccos\left(\sum_{i=1}^{n} \sqrt{p_i q_i}\right).$$

Therefore the Fisher-Rao distance on this manifold is

$$d_{\mathrm{FR}}(p, q) = 2\arccos\left(\sum_{i=1}^{n} \sqrt{p_i q_i}\right).$$

An immediate approximation is given by the chordal distance

$$\|\boldsymbol{z}_p - \boldsymbol{z}_q\|_2 = 2\left(\sum_{i=1}^{n}(\sqrt{p_i} - \sqrt{q_i})\right)^{1/2} = 2d_{\mathrm{H}}(p, q),$$

which happens to be double the *Hellinger distance*.

Supervised learning

Each feature vector $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^n$ belongs to exactly one class $y \in \mathcal{Y} := \{1, \dots, K\}$, and the data follows distribution $(\boldsymbol{x}, y) \sim \mathcal{D}$.
A classifier (e.g., neural network) $f \colon \mathcal{X} \to \mathbb{R}^K$ assigns a vector of scores $\boldsymbol{s} = (s_1, \dots, s_K) := f(\boldsymbol{x})$, which induces a decision $\hat{y} = \arg\max_{1 \leq i \leq K} s_i$.
By applying the softmax function σ, we obtain a conditional probability $P(y|\boldsymbol{x})$ represented by $\boldsymbol{p} = (p_1, \dots, p_K) := \sigma\left((s_1, \dots, s_K)\right)$, with $p_i = e^{s_i} / \sum_{j=1}^{K} e^{s_j}$.
The *risk* associated with a loss function $L \colon \mathcal{Y} \times \mathbb{R}^K \to \mathbb{R}_+$ is

$$R_L := R_L(f) := \mathbb{E}_{\mathcal{D}}\left[L\left(y, f(\boldsymbol{x})\right)\right].$$

Given a training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, the associated *empirical risk* is

$$\bar{R}_L := \bar{R}_L(f) := \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, f(\boldsymbol{x}_i)\right).$$

Training a classifier consists in solving $\min_f \bar{R}_L(f)$.

Loss functions

We denote $\boldsymbol{e}^{(y)} := (0, \ldots, 0, \underbrace{1}_{y\text{-th}}, 0, \ldots, 0) \in \mathbb{R}^K$.

Common loss functions:

- **Mean squared error (MSE):**

$$L_{\mathrm{MSE}}\left(y, f(\boldsymbol{x})\right) := \|\boldsymbol{e}^{(y)} - (\sigma \circ f)(\boldsymbol{x})\|_2^2 = \|\boldsymbol{p}\|_2^2 - 2p_y + 1$$

- **Mean absolute error (MAE):**

$$L_{\mathrm{MAE}}\left(y, f(\boldsymbol{x})\right) := \frac{1}{2}\|\boldsymbol{e}^{(y)} - (\sigma \circ f)(\boldsymbol{x})\|_1 = 1 - p_y$$

- **Cross entropy (CE):**

$$L_{\mathrm{CE}}\left(y, f(\boldsymbol{x})\right) := -\sum_{i=1}^{K} e_i^{(y)} \log[(\sigma \circ f)(\boldsymbol{x})]_i = -\log p_y$$

Loss functions

Other loss functions:

- **Cross q-entropy (q-CE)** [Zha18]:

$$L_{q\text{-CE}}(y, f(\boldsymbol{x})) := -\sum_{i=1}^{K} e_i^{(y)} \log_q [(\sigma \circ f)(\boldsymbol{x})]_i = -\log_q p_y,$$

with the Tsallis q-logarithm, for $q \in [0, 1]$:

$$\log_q(x) := \begin{cases} \frac{x^{1-q}-1}{1-q}, & q \neq 1 \\ \log(x), & q = 1 \end{cases}, \quad x > 0.$$

$q = 1$ corresponds to the CE loss, and $q = 0$ is the MAE loss.

Loss functions

- **Fisher-Rao distance:**

$$L_{\mathrm{FR}}(y, f(\boldsymbol{x})) \coloneqq \frac{1}{4}\left(d_{\mathrm{FR}}(\boldsymbol{e}^{(y)}, (\sigma \circ f)(\boldsymbol{x}))\right)^2$$
$$= \left(\arccos\sqrt{p_y}\right)^2$$

- **Hellinger distance:**

$$L_{\mathrm{H}}(y, f(\boldsymbol{x})) \coloneqq \left(d_{\mathrm{H}}(\boldsymbol{e}^{(y)}, (\sigma \circ f)(\boldsymbol{x}))\right)^2$$
$$= 2\left(1 - \sqrt{p_y}\right)$$

It corresponds to the q-CE loss for $q = 1/2$.

Loss functions

Proposition

The loss functions L_{FR}, L_{CE} e L_{H} are related:

1. $L_{\mathrm{FR}}(y, f(\boldsymbol{x})) = L_{\mathrm{H}}(y, f(\boldsymbol{x})) + O\big(L_{\mathrm{H}}^2(y, f(\boldsymbol{x}))\big)$;
2. $L_{\mathrm{FR}}(y, f(\boldsymbol{x})) = L_{\mathrm{CE}}(y, f(\boldsymbol{x})) + O\big(L_{\mathrm{CE}}^2(y, f(\boldsymbol{x}))\big)$.

Moreover:

3. $L_{\mathrm{H}}(y, f(\boldsymbol{x})) \leq L_{\mathrm{FR}}(y, f(\boldsymbol{x})) \leq L_{\mathrm{CE}}(y, f(\boldsymbol{x}))$.

Label noise

The classifier does not have access to a set of *clean* samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, but instead to a *noisy* dataset $\{(\boldsymbol{x}_i, \tilde{y}_i)\}_{i=1}^N$. In the case of *uniform label noise* of rate $\eta \in [0,1]$, the noisy data follows $(\boldsymbol{x}, \tilde{y}) \sim \mathcal{D}_\eta$, given by

$$\Pr(\tilde{y}_i = j | y_i = k) = \begin{cases} 1-\eta, & j = k, \\ \frac{\eta}{K-1}, & j \neq k. \end{cases}$$

Definition

Let f^* and $\hat{f}$ be the global minimisers of $R_L(f) \coloneqq \mathbb{E}_\mathcal{D}\left[L(y, f(\boldsymbol{x}))\right]$ and $R_L^\eta(f) \coloneqq \mathbb{E}_{\mathcal{D}_\eta}\left[L(\tilde{y}, f(\boldsymbol{x}))\right]$, respectively. The risk minimisation under loss function L is said to be *noise tolerant* if the classifier $\hat{f}$ has the same probability of misclassification as that of f^*.

$\rightsquigarrow$ Classifiers trained with clean and noisy data achieve the same classification accuracy.

Robustness to label noise

Theorem (Suffcicient condition for robustness [Gho17])

A loss function L is tolerant under uniform label noise with $\eta < \frac{K-1}{K}$, if $\sum_{i=1}^{K} L(i, f(\boldsymbol{x})) = C$, $\forall \boldsymbol{x} \in \mathcal{X}$, $\forall f$, for some constant C.

The MAE loss satisfies this condition, whereas MSE and CE do not:

$$\sum_{i=1}^{K} L_{\mathrm{MAE}}(i, f(\boldsymbol{x})) = \sum_{i=1}^{K} (1 - p_i) = K - 1$$

$$\sum_{i=1}^{K} L_{\mathrm{MSE}}(i, f(\boldsymbol{x})) = \sum_{i=1}^{K} \left(\|\boldsymbol{p}\|_2^2 - 2p_i + 1\right) = K\left(\|\boldsymbol{p}\|_2^2 + 1\right) - 2$$

$$\sum_{i=1}^{K} L_{\mathrm{CE}}(i, f(\boldsymbol{x})) = \sum_{i=1}^{K} (-\log p_i) = \sum_{i=1}^{K} \log \frac{1}{p_i}$$

$\rightsquigarrow$ If the sum in the condition is bounded, it is still possible to derive some theoretical guarantees.

Robustness to label noise

Theorem (Performance degradation under uniform label noise)

Let f^* and $\hat{f}$ be the global minimisers of $R_L(f)$ e $R_L^\eta(f)$, respectively. For the Fisher-Rao loss L_{FR}, under uniform label noise with $\eta < \frac{K-1}{K}$:

$$0 \leq R^\eta_{L_{\mathrm{FR}}}(f^*) - R^\eta_{L_{\mathrm{FR}}}(\hat{f}) \leq A_{\mathrm{FR}}$$
$$B_{\mathrm{FR}} \leq R_{L_{\mathrm{FR}}}(f^*) - R_{L_{\mathrm{FR}}}(\hat{f}) \leq 0$$

with

$$A_{\mathrm{FR}} \coloneqq A_{\mathrm{FR}}(K,\eta) \coloneqq \eta \left(\frac{\pi^2}{4} - \frac{K}{K-1} \left(\arccos \frac{1}{\sqrt{K}} \right)^2 \right)$$

$$B_{\mathrm{FR}} \coloneqq B_{\mathrm{FR}}(K,\eta) \coloneqq \eta \frac{K \left(\arccos \frac{1}{\sqrt{K}} \right)^2 - \frac{\pi^2}{4}(K-1)}{K-1-\eta K}.$$

Robustness to label noise

Not only label noise has a limited impact, but also it becomes negligible as K grows, for fixed η :

$$\lim_{K\to\infty} A_{\mathrm{FR}}(K,\eta) = \lim_{K\to\infty} \eta \left(\frac{\pi^2}{4} - \frac{K}{K-1} \left(\arccos \frac{1}{\sqrt{K}} \right)^2 \right) = 0,$$

$$\lim_{K\to\infty} B_{\mathrm{FR}}(K,\eta) = \lim_{K\to\infty} \eta \frac{K \left(\arccos \frac{1}{\sqrt{K}} \right)^2 - \frac{\pi^2}{4}(K-1)}{K-1-\eta K} = 0$$

Robustness to label noise

Table: Bounds $A(K,\eta)$ and $B(K,\eta)$ for different loss functions.

Loss function	$A(K,\eta)$	$B(K,\eta)$
Mean squared error (MSE)	η	$-\eta\frac{K-1}{K-1-\eta K}$
Mean absolute error (MAE)	0	0
Cross entropy (CE)	$+\infty$	$-\infty$
Cross q-entropy [Zha18]	$\eta\frac{K^q-1}{(1-q)(K-1)}$	$\eta\frac{1-K^q}{(1-q)(K-1-\eta K)}$
Fisher-Rao	$\eta\left(\frac{\pi^2}{4}-\frac{K}{K-1}\left(\arccos\frac{1}{\sqrt{K}}\right)^2\right)$	$\eta\frac{K\left(\arccos\frac{1}{\sqrt{K}}\right)^2-\frac{\pi^2}{4}(K-1)}{K-1-\eta K}$
Hellinger ($q=1/2$)	$\eta\frac{2(\sqrt{K}-1)}{K-1}$	$\eta\frac{2(1-\sqrt{K})}{(K-1-\eta K)}$

Robustness to label noise

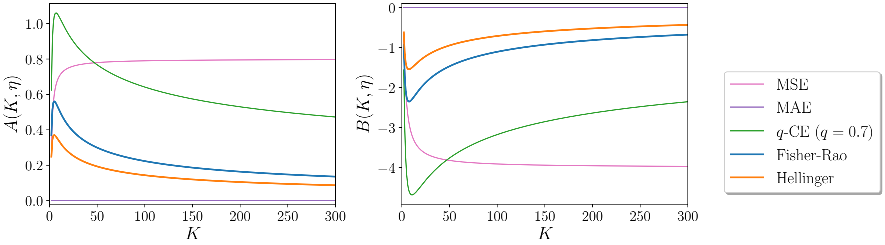


Figure: Bounds $A(K, \eta)$ and $B(K, \eta)$ as function of K, with $\eta = 0.8 - 1/K$.

Robustness to label noise:

$$\text{MAE} \geq \text{Hellinger} \geq \textbf{Fisher-Rao} \geq q\text{-CE } (q = 0.7) \geq \text{CE}$$

Learning speed

One would not like to trade label noise robustness for learning speed.
In gradient-like methods, the network parameters $\boldsymbol{w}$ are updated proportionally to the gradient of the empirical risk:

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \gamma \nabla_{\boldsymbol{w}} \bar{R}_L,$$

with $\nabla_{\boldsymbol{w}} \bar{R}_L = \frac{1}{N} \sum_{i=1}^{N} \nabla_{\boldsymbol{w}} L(y_i, f(\boldsymbol{x}_i))$.

MAE, CE, q-CE, Fisher-Rao and Hellinger losses can be written in the form

$$L(y, f(\boldsymbol{x})) = h(p_y),$$

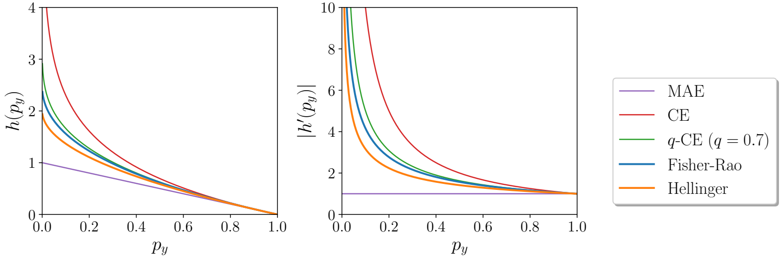
for a C^1 non-increasing function $h : [0,1] \to \mathbb{R}$, with $h(1) = 0$. In this case:

$$\nabla_{\boldsymbol{w}} L(y, f(\boldsymbol{x})) = h'(p_y) \nabla_{\boldsymbol{w}} [(\sigma \circ f)(\boldsymbol{x})]_y.$$

Learning speed

Table: Functions $h(p_y)$ and their derivatives $|h'(p_y)|$.

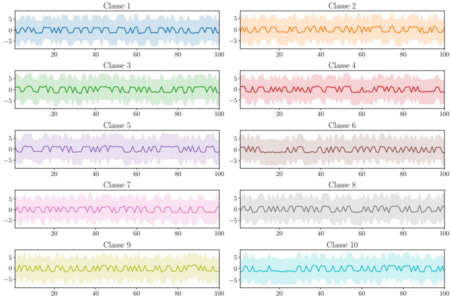
Loss function	$h(p_y)$	$\lvert h'(p_y)\rvert$
Mean absolute error (MAE)	$1 - p_y$	1
Cross entropy (CE)	$-\log p_y$	$\frac{1}{p_y}$
Cross q-entropy	$-\log_q p_y$	$\frac{1}{(p_y)^q}$
Fisher-Rao	$\left(\arccos\sqrt{p_y}\right)^2$	$\frac{\arccos\sqrt{p_y}}{\sqrt{p_y(1-p_y)}}$
Hellinger ($q = 1/2$)	$2\left(1 - \sqrt{p_y}\right)$	$\frac{1}{\sqrt{p_y}}$



Synthetic data

Data generated by Gaussian distributions centred on the vertex of a hypercube.

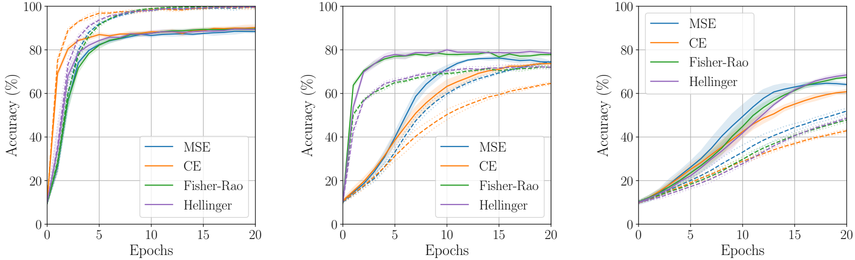
- 100-dimensional vectors divided in 10 classes.
- 8,000 training examples and 2,000 test examples.
- MLP network with there layers (80, 40, 20 neurons).
- ReLU, stochastic gradient.



Synthetic data

Table: Test accuracy (%).

Loss function	$\eta = 0$	$\eta = 0.3$	$\eta = 0.5$
Mean square error (MSE)	88.39 ($\pm$0.70)	74.43 ($\pm$0.41)	64.08 ($\pm$0.70)
Cross entropy (CE)	**90.21 ($\pm$1.27)**	73.68 ($\pm$0.99)	60.78 ($\pm$1.15)
Fisher-Rao	**89.64 ($\pm$0.80)**	**77.83 ($\pm$0.71)**	**67.38 ($\pm$0.46)**
Hellinger	89.36 ($\pm$1.18)	**78.43 ($\pm$0.66)**	**68.49 ($\pm$1.07)**



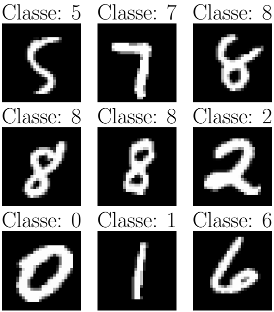
$\eta = 0$ $\eta = 0.3$ $\eta = 0.5$

Experimental results

MNIST

Grey-scale images of handwritten digits.

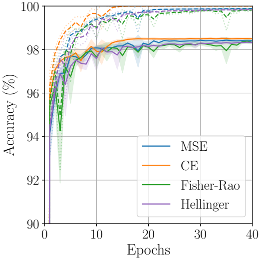
- 28×28 images divided in 10 classes.
- 60,000 training examples and 10,000 test examples.
- MLP network with two layers (300, 100 neurons).
- ReLU, stochastic gradient.



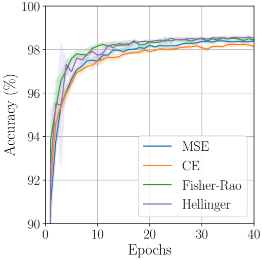
MNIST

Table: Test accuracy (%).

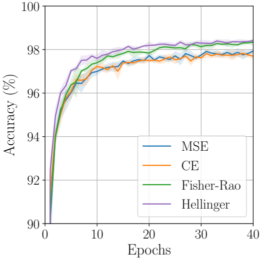
Loss function	$\eta = 0$	$\eta = 0.3$	$\eta = 0.5$
Mean square error (MSE)	**98.41 (±0.09)**	98.40 (±0.10)	97.93 (±0.07)
Cross entropy (CE)	**98.50 (±0.04)**	98.14 (±0.06)	97.69 (±0.16)
Fisher-Rao	98.32 (±0.07)	**98.44 (±0.05)**	**98.34 (±0.14)**
Hellinger	98.33 (±0.05)	**98.53 (±0.03)**	**98.40 (±0.06)**



$\eta = 0$



$\eta = 0.3$



$\eta = 0.5$

Conclusion and perspectives

- We have studied the use of a loss function based on the Fisher-Rao distance of the manifold of discrete distributions.
- It provides natural trade-off between robustness to (uniform) label noise and learning speed, as seen in theoretical results and illustrative examples.

Future perspectives:

- Extensive experiments, including more complex datasets and architectures (in progress).

References

S. Amari and H. Nagaoka. *Methods of Information Geometry*
Providence, RI, USA: American Mathematical Society, 2000

S. Amari. “Natural gradient works efficiently in learning”
Nueral Comput., vol. 10, no. 2, pp. 251–276, 1998

G. Arvanitidis *et al.* “Pulling back information geometry”
Proc. 25th Int. Conf. Artif. Intell. Stat. (AISTATS), pp. 4872–4894, 2022

C. Atkinson and A.F.S. Mitchell. “Rao’s distance measure”
Sankhya: The Indian J. Stat., Ser. A, vol. 43, no. 3, pp. 345–365

N. Ay, J. Jost, H. Vân Lê and L. Schwachhöfer. *Information Geometry*
Cham, Switzerland: Springer, 2017

O. Calin and C. Udrişte. *Geometric Modeling in Probability and Statistics*
Cham, Switzerland: Springer, 2014

O. Calin. *Deep Learning Architectures: A Mathematical Approach*
Cham, Switzerland: Springer, 2020

A. Demirkaya, J. Chen and S. Oymak. “Exploring the role of loss functions in multiclass classification”
54th Annu. Conf. Inf. Sci. Syst. (CISS), pp. 1–5, 2020

C. Frogner *et al.* “Learning with a Wasserstein loss”
Proc. 28th Int. Conf. Neural Inf. Process. Syst., pp. 2053–2061, 2015

S.A. Gattone *et al.* “A shape distance based on the Fisher-Rao metric and its application for shapes clustering”
Physica A: Statist. Mechanics Appl., vol. 487, pp. 93–102, 2017

References

A. Ghosh, N. Manwani and P. S. Sastry. "Making risk minimization tolerant to label noise"
Neurocomputing, vol. 160, pp. 93–107, 2015

A. Ghosh, H. Kumar and P. S. Sastry. "Robust loss functions under label noise for deep neural networks"
Proc. 31st AAAI Conf. Artif. Intell., pp. 1919–1925, 2017

P. Golik, P. Doetsch and H. Ney. "Cross-entropy vs. squared error training: a theoretical and experimental comparison"
Proc. Interspeech 2013, pp. 1756–1760, 2013

L. Hui and M. Belkin. "Evaluation of neural architectures trained with square loss vs. cross-entropy in classification tasks"
Proc. 9th Int. Cof. Learn. Representations (ICLR), 2021

K. Janocha abd W. M. Czarnecki. "On loss functions for deep neural networks in classification"
Schedae Informaticae, vol. 25, 2017

M. Picot *et al.*. "Adversarial robustness via Fisher-Rao regularization"
IEEE Trans. Pattern Anal. Mach. Intell., (early access), 2022

J. Pinele, J.E. Strapasson and S.I.R. Costa, "The Fisher-Rao distance between multivariate normal distributions: Special cases, bounds and applications"
Entropy, vol. 22, no. 4, 2020

S. Taylor. "Clustering financial return distributions using the Fisher information metric"
Entropy, vol. 21, no. 2, 2019

Z. Zhang and M.R. Sabuncu. "Generalized cross entropy loss for training deep neural networks with noisy label"
Proc. 32nd Conf. Neural Inf. Process. Syst. (NIPS), 2018