Information Geometry in Deep Neural Networks

Statistical Treatment and Analysis of the Data

Lucio De Simone

Department of Physical Sciences, Earth and Environment

Contents

- What is Information Geometry? “A method of exploring the world of information by means of modern geometry”[1], basically the application of differential geometry to statistics
- A brief history of Information Geometry...
  - C. R. Rao (1945) $\rightarrow$ Fisher matrix $=$ Riemannian metric
  - Further contributions in the following decades $\rightarrow$ H. Jeffreys, B. Efron, N. N. Chentsov, S. Kullback, among many others...
- Maturity has been reached by the work of *S. Amari* (1983)
- In 2018, Springer created the journal “Information Geometry”

Differential Geometry in a Nutshell

- Differential geometry lies on the concept of manifold: an m-dimensional manifold $\mathcal{M}$ is a topological space such that each point $p \in \mathcal{M}$ admits a neighborhood and an homeomorphism to $\mathbb{R}^m$
  $\rightarrow$ E.g. Stereographic projection
- From this simple statement, we can naturally define connection coefficients, covariant derivatives, the curvature tensor, parallel transport, and other related concepts.
- General relativity is the best application of this
- But, let's switch to statistics...

- Definition of Statistical Manifold:

$$\mathcal{M} = \{p_\xi = p(x;\xi) \,|\, \xi = (\xi^1, ..., \xi^m) \subset \mathbb{R}^m\}$$

  with the "natural" homeomorphism $\varphi(p_\xi) = \xi$

- How to measure the discrepancy between two points $p_\xi, p_{\xi'}$? We define a divergence $D[\xi : \xi']$ (not necessarily symmetric). For example:
  - Kullback-Leibler divergence

$$D_{KL}[\xi : \xi'] = \sum_i p(x_i;\xi) \log \left[\frac{p(x_i;\xi)}{p(x_i;\xi')}\right]$$

  - Bregman divergence (for a convex function $\psi(\xi)$)

$$D_\psi[\xi : \xi'] = \psi(\xi) - \psi(\xi') - \nabla\psi(\xi') \cdot (\xi - \xi')$$

Information Monotonicity

f-divergences

- On the manifold, we require *two* invariance principles
  1) invariance under coordinate trasformation
  2) *Information Monotonicity* [2]: let $t = t(x)$ be a general mapping between the sample spaces $\mathcal{X}$ and $\mathcal{Y}$, then

  $$D[\bar{p}(t;\xi) : \bar{p}(t;\xi')] \leq D[p(x;\xi) : p(x;\xi')]$$

  The equality holds if and only if $t(x)$ is a *sufficient statistics*
- An important class of divergences is the f-divergence [3]

  $$D_f[\xi : \xi'] = \int_{\mathcal{X}} p(x;\xi) f\left(\frac{p(x;\xi')}{p(x;\xi)}\right) dx$$

  where f is a convex function with $f(1) = 0$
  - *Any* f-divergence satisfies the information monotonicity [4]
  - Conversely, any decomposable information monotonic divergence is written in the form of f-divergence [5]

- The f-divergence satisfies the following relations
    1) For $\bar{f}(u) = f(u) + c(u-1)$ with $c \in \mathbb{R}$, we have $D_{\bar{f}} = D_f$
    2) For $f \to cf$ with $c > 0$, we have $D_{cf} = c\, D_f$
- From the first symmetry, we fix $f'(1) = 0$. In order to set the scale, we assume $f''(1) = 1$. The resulting f-divergence is called a *standard* f-Divergence
- *Chentsov's theorem*[2]: Any standard f-divergence gives the same Riemannian metric, the *Fisher information metric* given by

$$g_{ij} = \mathbb{E}\left[\partial_i \log p(x;\xi) \partial_j \log p(x;\xi)\right]$$

where $\partial_i = \frac{\partial}{\partial \xi^i}$

A curiosity...

AdS^N from a Multivariate Gaussian Distribution

- If we take the distribution

$$p(\{x^i\};\{\mu^i\},\{\Lambda_{ij}\}) = \frac{exp\big(-\frac{1}{2}\Lambda_{ij}(x^i-\mu^i)(x^j-\mu^j)\big)}{\sqrt{(2\pi)^{N/2}|\Lambda^{-1}|}}$$

  and compute the Fisher metric we get

$$ds^2 = \Lambda_{ij}d\mu^i d\mu^j + \frac{1}{2}\Lambda^{ik}\Lambda^{jl}d\Lambda_{ij}d\Lambda_{kl}$$

  where $dim(\mathcal{M}) = N(N+3)/2$

- For an isotropic distribution ($\Lambda_{ij} = \sigma^2\delta_{ij}$), the metric is formally equivalent to the AdS^N space (after a Wick rotation and a conformal constant factor), i.e.

$$ds^2 = \frac{1}{\sigma^2}\big(\delta_{ij}d\mu^i d\mu^j + 2N(d\sigma)^2\big)$$

Let's complicate a bit...

Amari-Chentsov triplet $\{\mathcal{M}, g_{ij}, T_{ijk}\}$

- So far, we considered the object g_{ij} as a Riemannian metric. To get a full and coherent picture of a manifold, we need to relate it to a connection. Typically, one requires $\langle X, Y\rangle = \langle \Pi X, \Pi Y\rangle$
- Here, we requires $\langle X, Y\rangle = \langle \Pi X, \Pi^* Y\rangle$ where Π and Π^* are related to the connections Γ_{ijk} and Γ^*_{ijk}
- The quantity $T_{ijk} = \Gamma^*_{ijk} - \Gamma_{ijk}$ is called the *Amari-Chentsov* tensor and it can be demonstrated that $\Gamma_{ijk} = \Gamma^0_{ijk} - \frac{1}{2}T_{ijk}$ and $\Gamma_{ijk} = \Gamma^0_{ijk} + \frac{1}{2}T_{ijk}$ where Γ^0_{ijk} is the Levi-Civita connection
- The triplet $\{\mathcal{M}, g_{ij}, T_{ijk}\}$ is called *Amari-Chentsov structure*

α-Geometry

- From a standard f-divergence we have the Fisher information metric and

$$T_{ijk}^{(\alpha)} = \alpha T_{ijk}$$

with $\alpha = 2f'''(1) + 3$ and
$T_{ijk} = \mathbb{E}\big[\partial_i \log p(x;\xi)\partial_j \log p(x;\xi)\partial_k \log p(x;\xi)\big]$

- Hence, we have a family of α-connections $\Gamma_{ijk}^{(\alpha)}$, $\Gamma_{ijk}^{(-\alpha)}$ which are dually coupled to the Fisher metric
- The same geometry is derived from the α-divergence[6]

$$D^{(\alpha)}[\xi;\xi'] = \frac{4}{1-\alpha^2}\left(1 - \int_{\mathcal{X}} p(x;\xi)^{\frac{1-\alpha}{2}} p(x;\xi')^{\frac{1+\alpha}{2}} dx\right)$$

- Anyway, the application of the tensor T_{ijk} is still unknown...

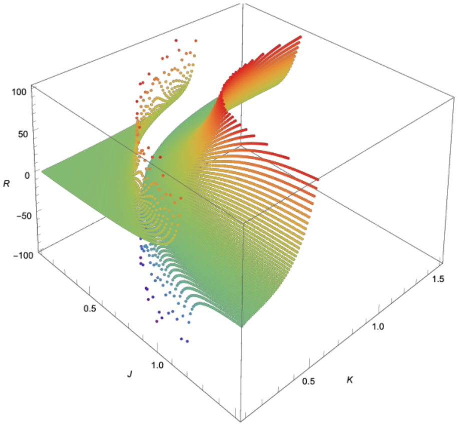
An application of α-Geometry

2D anisotropic Ising model[7] pt. 1

- Let's consider the 2D anisotropic Ising model $H(\sigma) = -J \sum_{i,j=1}^{N} \sigma_{i,j}\sigma_{i+1,j} - K \sum_{i,j=1}^{N} \sigma_{i,j}\sigma_{i,j+1}$
- At the equilibrium, we have $P(\sigma) = Z^{-1} e^{-\beta H(\sigma)}$
- Recall that, the canonical distribution arises by minimizing the Kullback-Leibler divergence ($\alpha = 1$) with the constraint on the energy
- Since $\ln Z$ is the potential, by computing the curvature with $\alpha = 1$ we find $R^{(1)} = 0$
- BUT...with $\alpha = 0$ (Levi-Civita connection), the curvature does not vanish...

An application of α-Geometry

2D anisotropic Ising model[7] pt. 2



This curvature $R^{(0)}$ correctly captures the phase transition and the divergence is the *Hellinger* distance

$D[p:q] = \sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2$ ...why this?

Deep Neural Networks

A very compact introduction

- Deep learning is based on neural networks. What are neural networks?
- Basically, given an input layer with N neurons $x = (x^1, \ldots, x^N)$, L hidden layers with n_l neurons for $l = 1, L$, and an output layer with M neurons $y = (y^1, \ldots, y^M)$, a neural network computes the numbers

$$z_i^{(l)} = \sum_{j=1}^{n_{l-1}} w_{ij}^{(l)} \varphi^{(l-1)}(z_j^{(l-1)}) + b_i^{(l)},\; i = 1, n_l,\; l = 1, L$$

  where $y_i = \varphi^{(out)}\Big(\sum_{j=1}^{n_L} w_{ij}^{out} \varphi^{(L)}(z_j^{(L)}) + b_i^{out} \Big)$
- The computational complexity increases with the number of hidden layers and related parameters w and b
- Chat-GPT 3 uses about 175 billion of parameters

- We want the output $y_i = y_i(x; w, b)$ to be as close as possible to a desired result $\tilde{y}_i$. The loss function $\mathcal{L}(w, b)$ quantifies the discrepacy, for instance

$$\mathcal{L}(w, b) = \frac{1}{N_{data}} \sum_x ||y_i(x; w, b) - \tilde{y}_i||^2$$

- As an example, let's consider a program that recognizes handwritten digits. Here, the input x would be the grayscale values of the pixels, while the output y would be an array of 10 probabilities, each corresponding to a digit from 0 to 9.
- The "Learning" essentially involves the minimization of $\mathcal{L}$

Minimization of $\mathcal{L}$

Stochastic Gradient Descent

- Parameters $\xi_0 = (w_0, b_0)$ are randomly generated at time $t = 0$. Then, they are updated according to

$$\xi_{t+1} = \xi_t - \eta_t \nabla \mathcal{L}(\xi_t)$$

  where η_t is the *learning rate* (generally depending by the *epoch* t). This is the so called *batch learning procedure*
- Typically, when the data set is big, we can estimate the gradient using a small sample of randomly chosen training inputs
- Since $\mathcal{L}(\xi) = \sum_x \mathcal{L}(y(x;\xi))$, the so called *on-line learning procedure* modifies ξ_t according to

$$\xi_{t+1} = \xi_t - \eta_t \nabla \mathcal{L}(y(x;\xi_t))$$

- Everything seems perfect, but... gradient descent could get stuck in local minima

- Learning takes place in a parameter space that is not Euclidean in general
- In this framework we have the *natural gradient descent*[8]

$$\xi_{t+1} = \xi_t - \eta_t G^{-1}(\xi_t) \nabla \mathcal{L}(\xi_t)$$

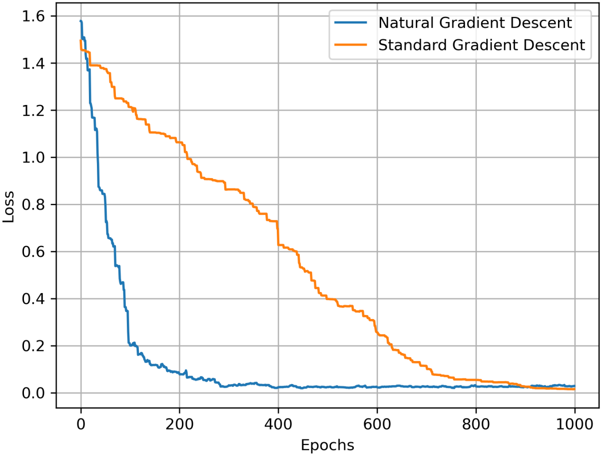
- Anyway, the choice of η_t is crucial. A good choice is to use an *adaptive learning rate* given by the *stochastic approximation* $\sum_t \eta_t > \infty$ $\sum_t \eta_t^2 < \infty$ (for instance $\eta_t = \mu/t$)
- In the on-line procedure and with $\eta_t = \mu/t$, the natural gradient descent is Fisher efficient, i.e. the Cramér-Rao bound is attained asymptotically

Let's go to practice...

Noisy Networks pt. 1

- Imagine to have an input signal x, distributed according to some $q(x)$, and a teacher signal given by $y = \varphi(x;\xi) + \epsilon$, where ϵ is some random noise (typically gaussian). The training sample is $D = \{(x_i, y_i), i = 1, T\}$
- The joint probability is $p(x, y) = q(x)P(y|x) = q(x)P_\epsilon(y - \varphi(x;\xi))$ and we can define an instantaneous loss as $\mathcal{L}(x_i, y_i; \xi) = -\log P_\epsilon$
- Minimizing $\mathcal{L}$ is equivalent to maximizing the log-likelihood
- The Fisher metric is $g_{ij}(\xi) = \mathbb{E}_q[\partial_i \varphi(x;\xi)\partial_j \varphi(x;\xi)]$. We could approximate it as $g_{ij}(\xi) \approx \frac{1}{T}\sum_t \partial_i \varphi(x_t;\xi)\partial_j \varphi(x_t;\xi)$

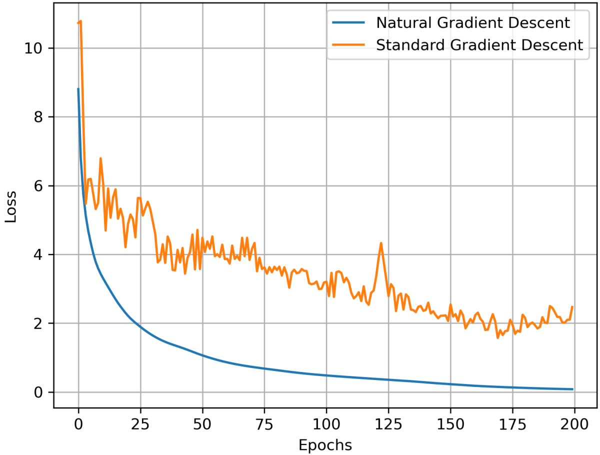
This is the case with $dim\_x = 20$ input neurons, no hidden layer and one output neuron.



Let's go to practice...

Noisy Networks pt. 3 (Jupyter Code: *Noisy Network (1 hidden layer)*)

This is the case with $dim\_x = 20$ input neurons, 1 hidden layer with $hidden\_dim = 10$ neurons and one output neuron.



Conclusions

- Information Geometry is a promising research field. The Ising model example suggests the possibility of going beyond the "canonical" statistical mechanics
- There are approaches that generalise the entropy, for instance, the Tsallis Entropy and the Rényi Entropy
- AI is conquering the world, and Information Geometry provides it with more efficient ways to do so...

Thank you for your attention!

Lucio De Simone

l.desimone3@student.unisi.it

[1] S. Amari. *Information Geometry and Its Applications*. Applied Mathematical Sciences. Springer Japan, 2016. ISBN: 9784431559771.

[2] N. N. Chentsov. "Statistical decision rules and optimal inference". In: 1982. URL: https://api.semanticscholar.org/CorpusID:122486850.

[3] I. CSISZAR. "Information-type measures of difference of probability distributions and indirect observation". In: *Studia Scientiarum Mathematicarum Hungarica* 2 (1967), pp. 229–318. URL: https://cir.nii.ac.jp/crid/1571417125811646464.

[4] Imre Csiszár. "Information measures: A critical survey". In: *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*. 1974, pp. 73–86.

[5] Shun-ichi Amari. "-Divergence Is Unique, Belonging to Both -Divergence and Bregman Divergence Classes". In: *Information Theory, IEEE Transactions on* 55 (Dec. 2009), pp. 4925–4931. DOI: 10.1109/TIT.2009.2030485.

[6] Shun-ichi Amari and Hiroshi Nagaoka. "Methods of information geometry". In: 2000. URL: https://api.semanticscholar.org/CorpusID:116976027.

[7] Johanna Erdmenger, Kevin Grosvenor, and Ro Jefferson. "Information geometry in quantum field theory: lessons from simple examples". In: *SciPost Physics* 8.5 (May 2020). ISSN: 2542-4653. DOI: 10.21468/scipostphys.8.5.073. URL: http://dx.doi.org/10.21468/SciPostPhys.8.5.073.

[8] Shun-ichi Amari and S.C. Douglas. "Why natural gradient?" In: vol. 2. June 1998, 1213–1216 vol.2. DOI: 10.1109/ICASSP.1998.675489.