1 Introduction

A recurrent neural network is a function $f(e, \theta, \mathcal{W})$ where $e \in R^N$ is the input of the network, $\theta \in \mathbb{R}^K$ is the state of the network, and $\mathcal{W}$ are the weights of the network. In what follows, the weights $\mathcal{W}$ are assumed to be fixed and we will write $f(e, \theta, \mathcal{W}) = f(e, \theta)$. In other words, we aim at investigating a already trained network. We hope that this study will shed light on some aspects that will help design robust recurrent neural networks.

At iteration k, the network produces an output $s_k = f(e_k, \theta_{k-1}) \in \mathbb{R}^M$. Let us define the recurrence relation $\theta_{k-1} = \phi(s_{k-p}, s_{k-(p-1)} \ldots, s_{k-2}, s_{k-1})$. In the following, we restrict ourselves to the special case where ϕ is the concatenation: $\theta_k = \begin{pmatrix}\pi(\theta_{k-1})^T & s_k^T\end{pmatrix}^T = \begin{pmatrix}s_{k-(p-1)}^T & \cdots & s_{k-1}^T & s_k^T\end{pmatrix}^T$ where π is the projection on the $(p-1) \times M$ last components. Thus, θ_{k-1} and θ_k share the following $(p-1) \times M$ components: $s_{k-(p-1)}, \ldots, s_{k-1}$. Moreover, we have $K = p \times M$.

Let X be a dynamical system (i.e., a vector field) living on a smooth (i.e., C^∞) manifold $\mathcal{M}$ that is assumed to be unobservable. Let ψ be a measure function (i.e., any smooth function from $\mathcal{M}$ to $\mathbb{R}$). For example, when studying the dynamics of an aircraft we have $\mathcal{M} = \mathbb{R}^{12}$ since the state space of the complete dynamical model of an aircraft has six degrees of freedom (three positions and three angles) along with their time derivatives. In this setting, ψ could be a measure of the altitude. Let us come back to the general case. The sequence of the network's inputs (without noise) are: $\begin{pmatrix}e_{k-(p-1)} & e_{k-(p-2)} & \ldots & e_{k-1} & e_k\end{pmatrix} = \begin{pmatrix}\psi(x_{k-(p-1)}) & \psi(x_{k-(p_2)}) & \ldots & \psi(x_{k-1}) & \psi(x_k)\end{pmatrix}$. Takens' theorem tells us that the network is able to predict $s_k = \psi(x_{k+1})$ from θ_{k-1} if $p > 2\text{dim}(\mathcal{M})$. In practice, the inputs are noisy (and the network is only imperfectly trained). The external input e_k allows to introduce the first p measures $\psi(x_k)$ and to compensate the noise of previous inputs. Nevertheless, one can legitimately wonder if the network will be able to produce a series of correct predictions (its training, i.e., the weights, being fixed) or if it will be drowned in the noise of the input, for example if the variance of the predictions diverges. Moreover, one would like to quantify this level of robustness against noise and to know how this level of robustness varies as the initial conditions over θ vary, or as the noiseless inputs vary.

In order to apply the formalism of information geometry to recurrent neural networks, one must introduce randomness. In this perspective, one can decompose the general case into three special cases:

1. The network defines a deterministic differential equation. The initial condition is a random variable. The output is a stochastic process. It is then possible to study the behavior of the moments of this process over time. In this case, the output space is endowed with the Fisher information metric ${}_{\mathcal{N}}g$. Strictly speaking, there is no input space (the output is directly reinjected into the network).

2. The network defines a stochastic differential equation. The initial condition is a random variable. Moreover, at each instant t, the network receives in input a random variable whose law is known, e.g., a Wiener process (to model a purely random noise). One can then study the curve of the parameters of the output process. The input and output spaces are identical and both endowed with the Fisher information metric ${}_{\mathcal{N}}g$.

3. This is the most general case. The network defines a stochastic differential equation. The initial condition is a random variable. At each instant t, the network receives in input a random variable whose law is unknown. One can observe the curves of the parameters of the output process in the output space. The output space is endowed with the Fisher information metric ${}_{\mathcal{N}}g$. The input space is endowed with the pullback metric $f^* {}_{\mathcal{N}}g$.

In this three cases, the output space is endowed with the metric tensor field of the Fisher information ${}_{\mathcal{N}}g$. To do so, we assume that the network defines a Gaussian law in each point of the output space. The general model is illustrated in Figure 1.

In this document, we begin with a simplified model defined by the following assumptions:

- We consider the continuous case i.e., θ follows the differential equation $\dot{\theta}(t) = f(\theta(t))$ instead of the difference equation $\theta_{k+1} - \theta_k = f(\theta_k) - \theta_k$.

- f is purely recurrent i.e., the external input e is a Gaussian white noise ($e(t)dt = dw(t, \omega)$) with the same dimension as θ. Thus, the input space and the output space are identical.

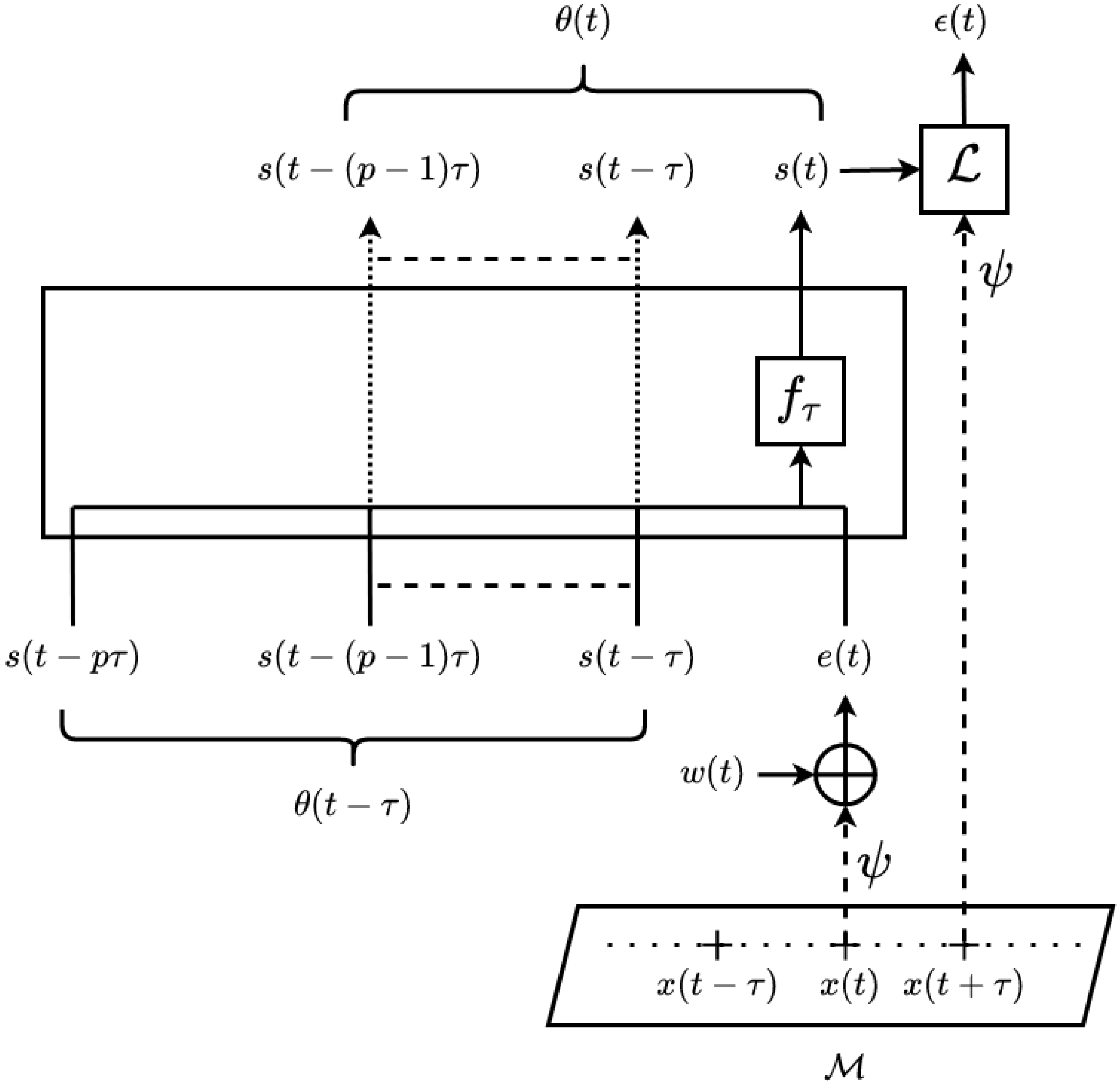

Figure 1: The objective is to learn the dynamic of a system living on a manifold $\mathcal{M}$. We have a measure function ψ. However, a random noise $w(t)$ is added to the measures. This noise can be a Gaussian noise or an adversarial attack. Thus, $e(t) = \psi(x(t)) + w(t)$. The prediction of the model is $s(t) = f_\tau(e(t), \theta(t-\tau))$ where τ is a time step. When $\tau \longrightarrow 0$, the difference equation on θ can be converted into a differential equation. A loss function is used to train the network by computing the error $\epsilon(t) = \mathcal{L}(s(t), \psi(x(t+\tau)))$.

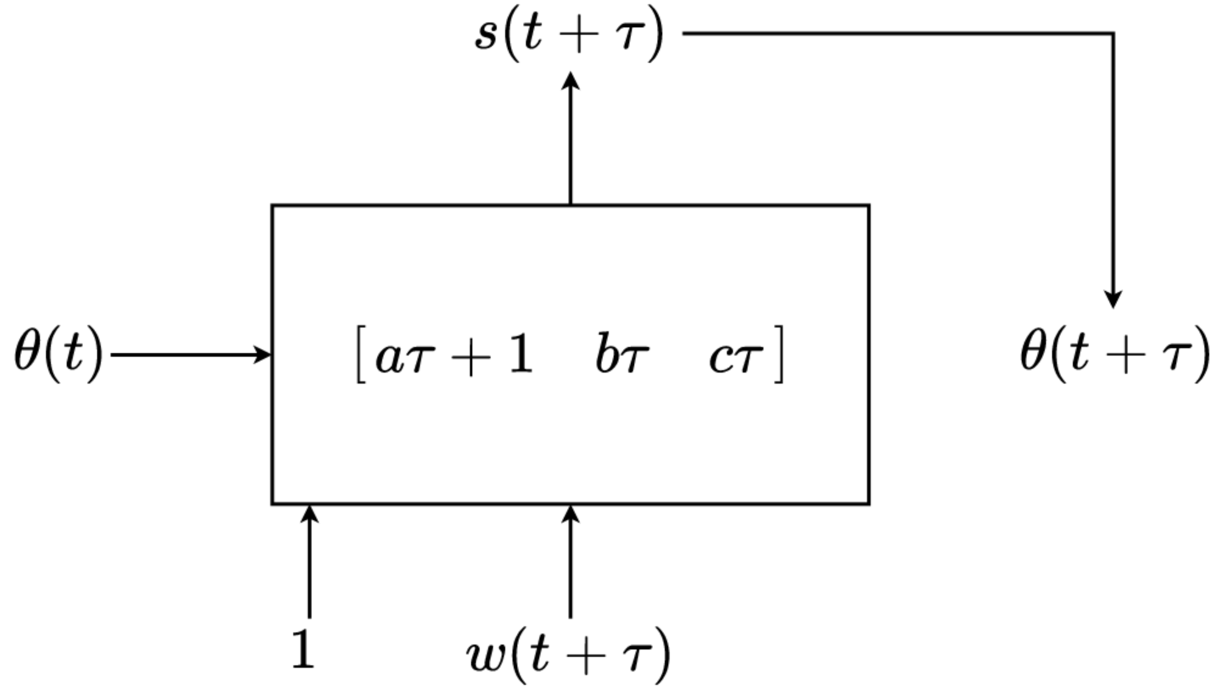

Figure 2: Simplified model. We assume that $p = 1$ (hence $\theta(t) = s(t)$), and $e(t) = w(t)$. Moreover, we have $\theta(t+\tau) = f_\tau(\theta(t), w(t+\tau)) = (a\tau+1)\theta(t) + b\tau + c\tau w(t)$. We add a constant 1 in input to include the bias in the matrix of f_τ.

- $\theta \in \mathbb{R}$ i.e., $M = 1$ and $p = 0$. Thus, $\theta = s$ such that we work directly on s.
- f is linear.

The simplified model is shown in Figure 2.

In order to study the robustness of the network, we assume that θ is a stochastic process solution of the following differential equation:

$$d\theta(t,\omega) = f(\theta(t,\omega))dt + cdw(t,\omega),$$

where w is a standard Wiener process (i.e., Brownian motion) defined by:

- w has independent increments i.e., for all $0 \le t_1 < t_2 < t_3 < t_4$, $w(t_2) - w(t_1)$ is independent from $w(t_4) - w(t_3)$.
- For all n and $t_1, \ldots, t_n$, one has that $\begin{pmatrix} w(t_1) & \ldots & w(t_n) \end{pmatrix}$ is a Gaussian vector.
- The trajectories of w are almost surely continuous.

Moreover, the trajectories of w are almost nowhere differentiable. For a given t, one has $w(t) \sim \mathcal{N}(0, t^2)$. It can be shown that $\theta(t) \sim \mathcal{N}(\mu(t), \sigma^2(t))$. The goal of this first step is to study the metric properties of the curves of θ in the space (μ, σ^2) as a function of the initial conditions $\theta_0 = (\mu(0), \sigma^2(0))$. Since θ follows a Gaussian law, the space (μ, σ^2) is a Poincaré half-plane (hyperbolic space). Since f is linear, the curves can be described explicitly. We expect two kinds of behaviors depending on the eigenvalues of f. If f is a contraction mapping (non-positive eigenvalue), the curves will exponentially converge to a limit value σ_* proportional to the standard deviation of θ. This means that the preceding noise has been attenuated and only the current noise affects the standard deviation of θ. If f has a positive eigenvalue, the curves will diverge exponentially along σ. It is then impossible to make prediction on the value of θ since its standard deviation diverges.

In the simplified case treated here, the input space has same dimension as the output space (since both spaces are identical). Hence, the pullback metric is not degenerate and there is no foliation in the input space. Therafter, we will consider the following obectives:

- Study of the scalar linear case with a deterministic differential equation.
- Relax the assumption $M = 1$. We will consider the case $\theta \in \mathbb{R}^K$ where $K = M$ is arbitrary, and $p = 0$. It will no longer be possible to compute lengths and curvature explicitly (it will be impossible in the general non-linear case anyway). The goal will then be to bound the length and the curvature of the curve of the output process using known linear processes. A common method is to assume that the network is Lipschitz. Under this assumption, it is possible to bound the output of the network by two linear processes.

These studies do not take into account the correlation between consecutive time steps of the input process. At a later stage, we will relax the assumption $p = 0$ in order to study the curves of a network taking as input a sequence of p consecutive time steps (as in Takens' theorem). These curves will live in a product of hyperbolic spaces. The dimension of the input space is strictly greater than the dimension of the output space, the pullback metric is then degenerated (its a pseudometric) which leads to the definition of several foliations:

- A foliation where the leaves are the kernels of the metric. This means that when we move along these leaves in the input space, there is no movement in the output space.
- A foliation transverse (or normal) to the preceding leaves. A displacement along these leaves corresponds to a maximal displacement in the output space.

Our objective is then to study the interaction of these foliations with the curves of θ. By bounding the output process with known processes, it is possible to quantify the robustness of the network to noise. In particular, we can show that the standard deviation does not diverge in finite time. This noise can originate from an aspect of the studied dynamical system not included in the model, or unknown (e.g., the wind). It may originate from measurement errors of transmission errors. It can also be intentionally injected (e.g., adversarial attack). If the standard deviation is asymptotically bounded then the network is robust: despite the noise, the observed output will always be sufficiently correlated with the expected output.

2 Litterature review

$\longrightarrow$ *Geometrical studies of robustness for the prediction of time series. See ICML or NeurIPS.*

2.1 Information Geometry and Data Manifold Representations

Let us focus on Ke Sun's PhD thesis published in 2015 [1]. The objective of this thesis is to create links between information geometry and machine learning. More precisely, Sun is interested in unsupervised learning, in a family of methods called "manifold learning".

2.1.1 Information Geometry

A statistical manifold $\mathcal{M}^m$ of dimension m is a space of probability distributions endowed with a Riemannian manifold structure. Why is this point of view important and useful? Because every method of statistical learning consists in evolving a statistical model (i.e., some $\theta \in \mathcal{M}$), this is why the study of $\mathcal{M}$ is useful for statistics and machine learning. Let us give two examples. The geometric distance between the current model θ_0 and a submanifold defined by some constraints defines a cost function for a learning as well as an intuitive interpretation of *learning as projection* of θ_0 on this submanifold.[1] An optimization algorithm taking into account the geometry of $\mathcal{M}$ may be more efficient, this is the idea behind *natural gradient.*

[1] In statistics, estimation consists of projecting models according to data. The models are the objects of study.

In the following, we assume that $\mathcal{M}$ is an *exponential family* where the densities $p(x|\theta)$ are:

$$p(x|\theta) = \exp(\theta^T t(x) - \phi(\theta)),$$

such that $\int p(x|\theta) d\nu(x) = 1$ with θ the canonical parameter, $t(x)$ a sufficient statistics with respect to θ, $\phi(\theta)$ a strictly convex potential function, and $\nu(x)$ a base measure. Since ϕ is strictly convex, there is a bijection between θ and $\eta = \frac{\partial\psi}{\partial\theta}$. One has:

$$\frac{\partial}{\partial\theta} \int p(x|\theta) d\nu(x) = \int \exp(\theta^T t(x) - \psi(\theta))(t(x) - \eta) d\nu(x) = 0,$$

hence:

$$\eta = \int p(x|\theta) t(x) d\nu(x) = \mathbb{E}[t(x)],$$

where the expectation is taken with respect to $p(x|\theta)$. This is why η is called the expectation parameter. Both parameters θ and η can be used as global coordinate systems on $\mathcal{M}$. They are linked by the Legendre transformation:

$$\eta = \frac{\partial\psi}{\partial\theta} \text{ et } \theta = \frac{\partial\psi^*}{\partial\eta},$$

with $\psi^* = \int p(x|\theta) \ln p(x|\theta) d\nu(x)$ the negative entropy, which is the dual potential function strictly convex with respect to η. Notice that:

$$\psi^*(\theta) = \int p(x|\theta)(\theta^T t(x) - \psi(\theta)) d\nu(x) = \theta^T \eta - \psi(\theta),$$

which gives the fundamental relation $\psi^* - \theta^T \eta + \psi = 0$.

2.1.2 Fisher Information Metric

A Riemannian metric is a covariant tensor field of rank 2. The Fisher information metric (FIM), also known as the Fisher-Hotelling-Rao metric, is given by:

$$g(\theta) = \int p(x|\theta) \frac{\partial \ln p(x|\theta)}{\partial\theta} \frac{\partial \ln p(x|\theta)}{\partial\theta^T} d\nu(x).$$

It transforms itself in the following way:

$$g(\eta(\theta)) = \left(\frac{\partial\theta}{\partial\eta}\right)^T g(\theta) \frac{\partial\theta}{\partial\eta}.$$

The FIM is invariant under change of coordinate system, hence it is a intrinsic property of $\mathcal{M}$. Up to a multiplicative factor, the FIM is the only invariant Riemannian metric over $\mathcal{M}$. The FIM can also be written:

$$g(\theta) = -\int p(x|\theta) \frac{\partial^2}{\partial\theta\partial\theta^T} \ln p(x|\theta) d\nu(x).$$

In the case of an exponential family, the FIM becomes:

$$g(\theta) = \frac{\partial^2\psi}{\partial\theta\partial\theta^T},$$

which is the Hessian matrix of the potential function which can be rewritten:

$$g(\theta) = \frac{\partial\eta}{\partial\theta}.$$

Similarly, one has $g(\eta) = \frac{\partial\theta}{\partial\eta} = \frac{\partial^2\psi^*}{\partial\eta\partial\eta^T}$.

The FIM allows to measure information. The information volume can be defined by $d\theta = \sqrt{|g(\theta)|} d\theta$ where $|\cdot|$ is the determinant. According to the Cramér-Rao bound, $|g[\theta)|$ is the amount of information contained in an unique observation with respect to θ (more formally, it is the inverse of the smallest variance of an unbiased estimator of θ built using one observation, thus it is the precision of the estimation). The smaller $|g[\theta)|$ is, the more one had to increase the number of observations to estimate θ. The information volume is invariant under coordinate change. The information capacity of a family of models $\mathcal{M}_h \subset \mathcal{M}$ is given by its volume. It is the total amount of information, or the "number" of distinct probability distributions, contained in $\mathcal{M}_h$.

2.1.3 Information Divergence

The infinitesimal distance between two points η and $\eta + d\eta$ is $\sqrt{d\eta^T g(\eta)\eta}$. However, the macroscopic distance has no close form in general. This is why information geometry relies on divergences. A divergence is a smooth measure of distortion verifying:

- $D(\eta_1||\eta_2) \geq 0$,
- $D(\eta_1||\eta_2) = 0$ iff. $\eta_1 = \eta_2$,
- $D(\eta + d\eta||\eta) \approx \frac{1}{2} d\eta^T g(\eta) d\eta$.

A divergence is not necessarily symmetric, neither verify the triangle inequality.

The Bregman divergence is induced by the negative entropy:

$$\begin{aligned}
D(\eta_1||\eta_2) &= \psi^*(\eta_1) - \psi^*(\eta_2) - \frac{\partial \psi^*}{\partial \eta^T}(\eta_2)(\eta_1 - \eta_2), \\
&= \psi^*(\eta_1) - \psi^*(\eta_2) - \theta_2^T(\eta_1 - \eta_2), \\
&= \psi^*(\eta_1) - \eta_1^T \theta_2 + \psi(\theta_2), \\
&= \int p(x|\eta_1) \ln \frac{p(x|\eta_1)}{p(x|\eta_2)} d\nu(x),
\end{aligned}$$

which is the Kullback-Leibler (KL) divergence. The KL divergence is thus a special case of the Bregman divergence. Another class of invariant divergences is defined from the principle of monotonous information. It is the family of f-divergences:

$$D(\eta_1||\eta_2) = \int p(x|\eta_1) f(\frac{p(x|\eta_2)}{p(x|\eta_1)} d\nu(x),$$

where f is a convex function verifying $f(1) = 0$. The KL divergence is obtained with $f(t) = -\ln(t)$.

2.1.4 Natural Gradient

Given a coordinate system θ (not necessarily the canonical parameter), the natural gradient is defined by:

$$\mathrm{grad} f = \left(g^{-1}(\theta) \frac{\partial f}{\partial \theta} \right)^T \frac{\partial}{\partial \theta}.$$

The natural gradient is invariant under coordinate change. In the case of exponential family with θ and η the canonical and expectation parameters, one has:

$$\mathrm{grad} f = \left(\frac{\partial f}{\partial \eta} \right)^T \frac{\partial}{\partial \theta} = \left(\frac{\partial f}{\partial \theta} \right)^T \frac{\partial}{\partial \eta}.$$

2.1.5 Gaussian Manifold

The density of a multidimensional normal law of mean μ and covariance matrix Σ is:

$$\begin{aligned}
G(x|\mu, \Sigma) &= \exp\left(-\frac{m}{2} \ln(2\pi) - \frac{1}{2} \ln|\Sigma| - \frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) \right), \\
&= \exp\left(x^T \theta^{(1)} + \mathrm{tr}(\theta^{(2)} x x^T) - \psi(\theta^{(1)}, \theta^{(2)}) \right),
\end{aligned}$$

with $\theta^{(1)} = \Sigma^{-1}\mu$, $\theta^{(2)} = -\Sigma^{-1}/2$, and

$$\begin{aligned}\psi(\theta^{(1)}, \theta^{(2)}) &= \frac{m}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma| + \frac{1}{2}\mu^T\Sigma^{-1}\mu, \\ &= \frac{m}{2}\ln(2\pi) - \frac{1}{2}\ln|-2\theta^{(2)}| - \frac{1}{4}\left(\theta^{(1)}\right)^T\left(\theta^{(2)}\right)^{-1}\theta^{(1)}.\end{aligned}$$

The set of these normal laws defines a Gaussian manifold $\mathcal{G}_m$ which is an exponential family of dimension $\dim(\mathcal{G}_m) = m + m(m+1)/2 = m(m+3)/2$. The expectation parameters are $\eta^{(1)} = \mathbb{E}[x] = \mu$ and $\eta^{(2)} = \mathbb{E}[xx^T] = \Sigma + \mu\mu^T$. The set of diagonal normal laws is an embedded submanifold of dimension $2m$. The KL divergence is:

$$D(\eta_1||\eta_2) = \frac{1}{2}\mathrm{tr}(\Sigma_1\Sigma_2^{-1}) - \frac{1}{2}\ln|\Sigma_1\Sigma_2^{-1}| + \frac{1}{2}(\mu_1 - \mu_2)^T\Sigma_2^{-1}(\mu_1 - \mu_2) - \frac{m}{2}.$$

2.2 Deep Learning and Information Geometry for Time-Series Classification

Let us focus on Daniel Brooks' PhD thesis published in 2020 [2]. In this thesis, Brooks focuses on supervised learning for the prediction of time series. His goal is to take advantage of the specific structure of time series data in order to improve supervised learning methods on such data.

2.2.1 Introduction

Data representation is important for the robustness of deep learning models to adversarial attacks. Where the human can infer concepts and causal relations, the model only sees float tensors. The choice of the mathematical space where the data live is thus extremely important in order to design reliant and robust models. Convolutional networks are a paradigmatic example of an architecture adapted to its data. Indeed, convolution is equivariant by translation and leads to better performances with less parameters and with more robustness. Then, the thesis pursues two objectives leading to two different approaches:

- A more intrinsic representation of data ⟶ data preprocessing.
- Models with a notion of robustness ⟶ model design.

In the area of Digital Signal Processing, most representations are linked with the Fourier transform. Concerning models, the trend for the last years is to focus on smaller models in order to favor the reasoning capacity instead of pure learning. Another trend is the emergence of the xAI (explainable AI) research program. More generally, the idea is to design a representation which is robust/invariant to a given class of transformations. For example, normalization of model's parameters leads to invariance to scaling (e.g., Batch Normalization, Layer Normalization etc.). Other approaches to regularize a network take inspiration of Shannon entropy, using optimal transportation, or using information geometry. This last approach is developed in the thesis. The idea of information geometry applied to machine learning is to model the data by the distribution that generates them, instead of seeing data as points in an Euclidean space. Notice that, for now, information geometry applied to learning is far less developed than its Euclidean counterpart.

Among the famous applications of information geometry to machine learning, one can mentions:

- The Fisher kernel method.
- The natural gradient.

Although there are two very different applications of information geometry, they share the idea to rely on theoretical aspects (of information geometry) in order to improve more classical machine learning methods. Another approach could consist of seeing specific data as living naturally on a statistical manifold and to design learning methods above this structure. This is this last idea that is developed in the thesis. In practice, the study is restricted to exponential families, and more specifically to Gaussian families.

The thesis focuses on neural networks over covariance matrices and uses methods from computer vision for other application domains including micro-Doppler radar.

2.2.2 Information Geometry

The Kullback-Leibler (KL) divergence can be defined from the Shannon entropy, then it can be shown that the second-order Taylor expansion of the KL divergence reduces to the Fisher information matrix. The Fisher information matrix can thus be seeing as a Riemannian metric. This is true for any exponential family. Here, we focus on the family of multidimensional centered Gaussian distributions of dimension n, denoted $\mathcal{S}_*^+$. The element of this family are entirely characterized by their covariance matrix Σ. The parameter Σ is a coordinate system for the manifold of this family. Given a covariance matrix $P \in \mathcal{S}_*^+$ (i.e., a positive semidefinite (PSD) matrix), the tangent space at P is the set of symmetric matrices.

Let S_1 and S_2 be two symmetric matrices of the tangent space at P. The dot product induced by the Fisher information metric is $\langle S1, S2 \rangle_P = \text{tr}(S_1 P^{-1} S_2 P^{-1})$. The geodesic distance between two PSD matrices P_1 et P_2, called "affine invariant Riemannian metric" or Rao distance, is then be defined as: $\delta(P_1, P_2) = \frac{1}{2}\|\log(P_1^{-1/2} P_2 P_1^{-1/2})\|_F$ where $\|\cdot\|_F$ is the Frobenius norm and log is the matrix logarithm defined on PSD matrices by: $\log(P) = U\text{diag}(\log \lambda_i)U^T$. Although being relevant from the point of view of information geometry, this distance is hard to compute (due to the computation of eigenvalues). There are several variants trying to approach this computation, or relying on another theoretical approach, e.g., the Fisher-Bures metric, the Bregman divergence, or the optimal transportation. While the KL divergence is a local metric, the Rao distance is a global metric then can be used in learning algorithm.

The tangent space (symmetric matrices) and the manifold (PSD matrices) can be linked by the exponential map. Given a symmetric matrix S from the tangent space at the PSD matrix G, the exponential map is $\exp_G(S) = G^{1/2}\exp(G^{-1/2} S G^{-1/2}) G^{1/2} \in S_*^+$ where the matrix exponential is defined by $\exp(A) = U\text{diag}(\exp \lambda_i)U^T$. Notice that the singular value decomposition can be used to efficiently derive the eigenvalues of a PSD matrix.

2.2.3 Riemannian machine learning

Let us give some examples of machine learning algorithms where the data are PSD matrices. First, it is possible to adapt the k nearest neighbors algorithm by replacing the Euclidean distance by the Rao distance. It is also possible to adapt the minimum distance to mean (MDM) algorithm using the Rao distance. Moreover, the Euclidean barycenters can be replaced by Riemannian barycenters defined by $\mathfrak{B} = \text{Bar}(\{P_i\}) = \arg\min_{G \in S_*^+} \sum_i^N \delta(G, P_i)^2$. The method thus obtained is called the Minimum Riemannian Distance to Riemannian Mean (MRDRM). It is also possible to weight the P_i. If the metric δ is induced by a global dot product, a closed formula can be obtained for $\mathfrak{B}$. This is no longer possible (except some special cases) if δ comes from a Riemannian metric. In this later case, one may use the Karcher algorithm. This is an iterative algorithm. Starting with an estimate of the barycenter $\mathfrak{B}(t)$, the points P_i are pulled back in the tangent space at $\mathfrak{B}(t)$ with $\log_{\mathfrak{B}(t)}$. Then, the barycenter of these points in the tangent space. Finally, this new barycenter is projected into the manifold with $\exp_{\mathfrak{B}(t)}$, providing the new estimate $\mathfrak{B}(t+1)$.

2.3 GeoSeq2Seq: Information Geometric Sequence-to-Sequence Networks

In [3], Bay and Sengupta apply the Fisher kernel method to sequence-to-sequence recurrent neural networks. In a sequence-to-sequence RNN, an encoder network uses the input sequence to produce a *context vector*. Then, this context vector is used by a decoder network to produce the prediction of the model. The idea of the Fisher kernel mehtod is to encode the context vector further and use this further encoding as input for the decoder network. There are two different encoding of the context vector:

- Fisher encoding. First, a set a N context vectors (each of dimension D) is extracted: $X = (w_1, \ldots, w_N : w \in \mathbb{R}^D)$. Then, we can consider the log-likelihood of this sample $L(\Theta) = \sum_{i=1}^N \log(\pi(w_i))$ where $\pi(w_i)$ is a Gaussian Mixture Model (GMM) with K component $\pi(w_i) = \sum_{k=1}^K \omega_k \mathcal{N}(w_i; \mu_k, \Sigma_k)$. The parameters $\Theta = \{(\omega_k, \mu_k, \Sigma_k)_k\}$ are learnable. The vector μ_k is D-dimensional while Σ_k is a $D \times D$ diagonal matrix. Finally, we can define the Fisher encoding as $W(X, \Theta) = (\frac{\partial L}{\partial \mu_1}, \ldots, \frac{\partial L}{\partial \mu_K}, \frac{\partial L}{\partial \Sigma_1}, \ldots, \frac{\partial L}{\partial \Sigma_K})$ where each element has dimension D, such that the entire Fisher encoding has dimension $2KD$.

- Vector Locally Aggregated Descriptors (VLAD). In this method, the context vectors $X = (w_1, \ldots, w_N : w \in \mathbb{R}^D)$ are clustered (using a clustering algorithm such as K-means). Let q_{ik} be the assignment of

w_i to the kth cluster such that $0 \leq q_{ik} \leq 1$ and $\sum_{k=1}^{K} q_{ik} = 1$. Then, for each cluster, we can compute a residual $v_k = \sum i = 1^N q_{ik}(w_i - \mu_k) \in \mathbb{R}^D$ where μ_k is the cluster center. The final encoded context vector is obtained by stacking toghether these residuals, hence we get a context vector of dimension DK.

Now, the "GeoSeq2Seq" model works as follows. First, a vanilla Seq2Seq model is trained. Then, for each one of the N training example, we obtain a context vector $w_i \in \mathbb{R}^D$ (in general, $D = 256$ or 512 etc.) These contect vectors w_i are used to learn the parameters of the GMM (for Fisher encoding) or the cluster centers μ_k and assignements q_{ik} (for VLAD). Once learned, we can generate the Fisher encoding or the VLAD encoding for each training example and use these new encodings as inputs to train a decoder network. Previous works have shown that it is possible to train the Fisher kernel parameters in an end-to-end manner for convolutional neural network, but it has not been investigated (as 2018) for Seq2Seq models.

The model is then tested on the shortest route problem using the A^* algorithm to build the groung truth for the dataset. It is claimed that the probablistic representation of the context vector supersedes the non-probabilistic one by 10-15 % on prediction accuracy. VLAD has a higer accuracy then Fisher encoding despite the fast that Fisher encoding takes covariances into account (contrary to VLAD). This might be explained by the fact that the learned covariance matrices have high condition number. Finally, the model is tested on the copying task and the associative recall task. These are sequence-to-sequence task that have been developed to test the Neural Turing Machines. The GeoSeq2Seq model achieves perfect accuracy on the copying task. However, it was accurate on the associative task only for small sequences. Longer sequences result in small errors, even if the general shape of the element was still identifiable.

3 Random Ordinary Differential Equation

Consider the following random differential equation:

$$\begin{aligned} \dot{\theta}(t) &= a\theta(t) + b, \\ \theta(0) &\sim \mathcal{N}(\theta_0, \sigma_0^2). \end{aligned}$$

Its solution is:

$$\theta(t) = \left(\theta(0) + \frac{b}{a}\right) e^{at} - \frac{b}{a}.$$

Thus, $\theta(t)$ is Gaussian for all t, i.e., $\theta(t) \sim \mathcal{N}(\mu(t), \sigma^2(t))$ with:

$$\begin{aligned} \mu(t) &= \left(\theta_0 + \frac{b}{a}\right) e^{at} - \frac{b}{a}, \\ \sigma^2(t) &= \sigma_0^2 e^{2at}. \end{aligned}$$

Assume that $\theta_0 \neq 0$ and $\sigma_0^2 \neq 0$.
If $a < 0$, we have $\mu(t) \longrightarrow -\frac{b}{a}$ and $\sigma^2(t) \longrightarrow 0$.
Denote $r(t) = (\mu(t), \sigma^2(t))$. We have $\dot{r}(t) = (a\theta_0 + b)e^{at}\partial_\mu + 2a\sigma_0^2 e^{2at}\partial_{\sigma^2}$ hence $|\dot{r}(t)| = \frac{\sqrt{(a\theta_0+b)^2+2\sigma_0^2 a^2}}{\sigma_0}$.
Thus:

$$L(\theta_0, \sigma_0, t) = \int_0^t |\dot{r}(s)| ds = \frac{\sqrt{(a\theta_0 + b)^2 + 2\sigma_0^2 a^2}}{\sigma_0} t.$$

$\longrightarrow$ *recompute the curvature using the right formula.*

4 Ornstein–Uhlenbeck process

Consider the following stochastic differential equation (SDE) scalar, linear, autonomous, and with additive noise:

$$\begin{aligned} d\theta(t) &= (a.\theta(t) + b)dt + c.dw(t), \\ \theta(0) &= \theta_0 \text{ p.s.} \end{aligned}$$

where $\theta : \mathbb{R}^+ \times (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a stochastic process, w is a standard Wiener process, and $a, b, c, \theta_0 \in \mathbb{R}$. Assume that $a \neq 0$. The SDE is a Langevin equation:

$$d\theta(t) = k(\alpha - \theta(t))dt + \beta dw(t),$$

with $k = -a, \alpha = -\frac{b}{a}, \beta = c$. Its solution is an Ornstein-Uhlenbeck (OU) process. This process is stationary, Gaussian, Markov and continuous in probability. One has:

$$\begin{aligned}\mu(t) &= \alpha + (\theta_0 - \alpha)e^{-kt}, \\ &= -\frac{b}{a} + \left(\theta_0 + \frac{b}{a}\right)e^{at}, \\ \sigma^2(t) &= \frac{\beta^2}{2k}(1 - e^{-2kt}), \\ &= \frac{c^2}{2a}(e^{2at} - 1).\end{aligned}$$

If $a < 0$, then $\sigma^2(t) \longrightarrow -\frac{c^2}{2a}$ and $\mu(t) \longrightarrow -\frac{b}{a}$.
If $a > 0$, then $\sigma^2(t) \longrightarrow +\infty$ and:

- $\mu(t) \longrightarrow +\infty$ si $\theta_0 > -\frac{b}{a}$,
- $\mu(t) \longrightarrow -\infty$ si $\theta_0 < -\frac{b}{a}$,
- $\mu(t) \longrightarrow -\frac{b}{a}$ si $\theta_0 = -\frac{b}{a}$.

Let ${}_{\mathcal{N}}g$ be the metric tensor field of the Fisher information associated to the family of unidimensional normal laws parameterized by their expectation μ and their variance σ^2. Consider the parameter space $(\mathbb{R} \times \mathbb{R}^+_*, {}_{\mathcal{N}}g)$ and the global coordinate system $(\mathbb{R} \times \mathbb{R}^+_*, (\mu, \sigma^2))$. Denote $r(t) = (\mu(t), \sigma^2(t))$. We aim at computing the length of the curves $L(\theta_0, t_0, t) = \int_{t_0}^{t} |\dot{r}(s)| ds$ for all $t > t_0 > 0$ as well as the curvature of the curves.

Let us recall the definition of the Fisher information matrix. Let $\{p(x, \Theta)\}_\Theta$ be a family of probability densities parameterized by a parameter vector $\Theta \in \mathbb{R}^K$ (notice that for the Fisher information to be well defined, the support of $p(x, \Theta)$ must be independent of Θ). Denote the score by $s_\Theta(x) = \nabla_\Theta \ln p(x, \Theta) \in \mathbb{R}^K$, which is the gradient of the likelihood with respect to the parameters. The Fisher information matrix is defined by:

$${}_{\mathcal{N}}I = \mathbb{E}_\Theta[s_\Theta(x) s_\Theta(x)^T].$$

It is a symmetric positive semidefinite matrix of dimension $K \times K$. It can be interpreted as the covariance matrix of the score (since the expectation of the score is zero). More precisely, given two indices i and j, one has:

$$\begin{aligned}{}_{\mathcal{N}}I_{ij} &= \int_\Omega \frac{\partial \ln p(x, \Theta)}{\partial \Theta_i} \frac{\partial \ln p(x, \Theta)}{\partial \Theta_j} p(x, \Theta) dx, \\ &= \int_\Omega \frac{1}{p(x, \Theta)} \frac{\partial p(x, \Theta)}{\partial \Theta_i} \frac{\partial p(x, \Theta)}{\partial \Theta_j} dx,\end{aligned}$$

where Ω is the sample space in which x lives. Under mild regularity assumptions, one has:

$${}_{\mathcal{N}}I = -\mathbb{E}_\Theta[\nabla^2_\Theta(\ln p(x, \Theta))],$$

where $\nabla^2_\Theta(\ln p(x, \Theta))$ is the Hessian matrix of the likelihood with respect to the parameters.

In our case, consider the parametrization $r = (\mu, \sigma^2)$ (i.e., the coordinate system $(\mathbb{R} \times \mathbb{R}^+_*, (\mu, \sigma^2))$) with:

$$p(x, (\mu, \sigma^2)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right).$$

It can be shown that:

$${}_{\mathcal{N}}I = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

The corresponding metric tensor field is:

$$ {}_{\mathcal{N}}g = \frac{1}{\sigma^2}(d\mu)^2 + \frac{1}{2\sigma^4}(d\sigma^2)^2. $$

One has $\dot{r}(t) = \begin{pmatrix} \dot{\mu}(t) & \dot{\sigma^2}(t) \end{pmatrix}$ with:

$$ \dot{\mu}(t) = (a\theta_0 + b)e^{at}, $$
$$ \dot{\sigma^2}(t) = c^2 e^{2at}. $$

Assume that $c \neq 0$ (since the Fisher information is not defined if $\sigma^2 = 0$). The norm of the velocity is:

$$ |\dot{r}(t)| = \frac{\sqrt{2a((a\theta_0 + b)^2(e^{2at} - 1) + ac^2 e^{2at})}}{|c||1 - e^{2at}|} e^{at}, $$

hence

$$ L(\theta_0, t_0, t) = \int_{t_0}^{t} \frac{\sqrt{2a((a\theta_0 + b)^2(e^{2as} - 1) + ac^2 e^{2as})}}{|c||1 - e^{2as}|} e^{as} ds. $$

Notice that the integration starts at $t_0 > 0$. Indeed, at $t = 0$, $\theta(0)$ is almost surely constant so that the Fisher information is not defined (and the velocity diverges).
The geodesic curvature of a curvre r is:

$$ \kappa(t) = \frac{\sqrt{|\dot{r}(t)|^2 |D_t \dot{r}(t)|^2 - \langle D_t \dot{r}(t), \dot{r}(t) \rangle^2}}{|\dot{r}(t)|^3}. $$

The covariant derivative of $\dot{r}$ along r has the following expression in coordinates:

$$ D_t \dot{r}(t) = (\ddot{r}^k(t) + \dot{r}^i(t)\dot{r}^j(t)\Gamma_{ij}^k(r(t)))\partial_k. $$

The Christoffel symbols have the following expression:

$$ \Gamma_{ij}^k = \frac{1}{2} {}_{\mathcal{N}}g^{kl}(\partial_i {}_{\mathcal{N}}g_{jl} + \partial_j {}_{\mathcal{N}}g_{il} - \partial_l {}_{\mathcal{N}}g_{ij}). $$

After computation, we obtain $\Gamma_{11}^1 = \Gamma_{22}^1 = \Gamma_{12}^2 = \Gamma_{21}^2 = 0$, $\Gamma_{11}^2 = 1$, $\Gamma_{22}^2 = -\frac{1}{\sigma^2}$, et $\Gamma_{12}^1 = \Gamma_{21}^1 = -\frac{1}{2\sigma^2}$, hence :

$$ D_t \dot{r}(t) = -a(a\theta_0 + b)e^{at} \frac{e^{2at} + 1}{e^{2at} - 1} \partial_\mu + e^{2at} \left((a\theta_0 + b)^2 - \frac{2ac^2}{e^{2at} - 1} \right) \partial_{\sigma^2}. $$

The norm of the covariant derivative is:

$$ |D_t \dot{r}(t)| = \frac{\sqrt{2}|a|e^{at}}{c^2(e^{2at} - 1)^2} \sqrt{ac^2(a\theta_0 + b)^2(e^{2at} + 1)^2(e^{2at} - 1) + e^{2at}\left((a\theta_0 + b)^2(e^{2at} - 1) - 2ac^2\right)^2}. $$

Furthermore, we have:

$$ \langle D_t \dot{r}(t), \dot{r}(t) \rangle = -\frac{2a^2 e^{2at}}{c^2(e^{2at} - 1)^2} \left((a\theta_0 + b)^2 + \frac{2ac^2 e^{2at}}{e^{2at} - 1} \right). $$

We get the following expression for the curvature[2]:

$$ \kappa(t) = |e^{2at} - 1||a\theta_0 + b| \sqrt{\frac{a(e^{2at} - 1)[(a\theta_0 + b)^2 + c^2 a]^2}{2[a(e^{2at} - 1)(a\theta_0 + b)^2 + c^2 a^2 e^{2at}]^3}} $$

Thus, we have $\kappa(t) \geq 0$ (with $a \neq 0$ and $c \neq 0$).

[2] Notice that $a(e^{2at} - 1) > 0$ for all $a \neq 0$ and $t > 0$.

- If $a < 0$ and $a\theta_0 + b \neq 0$, we have :

$$\lim_{t \to +\infty} \kappa(t) = \frac{|(a\theta_0 + b)^2 + c^2 a|}{\sqrt{2}|a||a\theta_0 + b|^2}.$$

- If $a > 0$, we have :

$$\lim_{t \to +\infty} \kappa(t) = \frac{|a\theta_0 + b|}{a\sqrt{2((a\theta_0 + b)^2 + c^2 a)}}.$$

- If $a\theta_0 + b = 0$ then $\kappa(t) = 0$ for all t.

5 Linear SDE of dimension n

Before studying the case of a state $\theta(t) \in \mathbb{R}^n$ made of successive delays (as in Takens' embedding), we focus on a dimension n SDE with $\theta(t) \in \mathbb{R}^n$ without any other assumption on the relations between the components of $\theta(t)$. Furthermore, we will consider a linear SDE in order to obtain closed formulas. Thus, this study will constitute a reasonable step, since it can be seen as a linearization of the conjugate dynamics obtained from Takens' embedding.

Consider the following stochastic differential equation (SDE):

$$d\theta(t) = (A\theta(t) + b)dt + dw(t),$$

with $\theta(t) \in \mathbb{R}^n$, $A \in \mathcal{M}_n(\mathbb{R})$, $b \in \mathbb{R}^n$ and $w(t)$ a Wiener process of dimension n with a diffusion matrix Q and assumed independent with $\theta(t_0)$. The solution of this equation is a Gaussian process $\theta(t) \sim \mathcal{N}(\mu(t), \Sigma(t))$. Once endowed with the Fisher information metric, the solution lives on the manifold $M = \mathbb{R}^n \times \mathcal{S}_n^+$ where $\mathcal{S}_n^+$ is the set of symmetric positive definite matrices of size $n \times n$. The tangent space at point p is $T_pM = \mathbb{R}^n \times \mathcal{S}_n \sim \mathbb{R}^n \times \mathbb{R}^{n(n+1)/2}$ where $\mathcal{S}_n$ is the set of symmetric matrices of size $n \times n$.

We might be interested in the sectional curvature defined as follows. Let $p \in M$. Let Π be a plane of T_pM spanned by two tangent vectors X and Y. The plane section induced by Π, denoted by S_Π, is defined by: $S_\Pi = \exp_p \Pi$. The plane section S_Π is a submanifold of M of dimension 2 containing p (it is the set of points reached by the geodesics starting at p with a tangent vector in Π). The sectional curvature (of M associated to Π) is defined as the Gauss curvature of the surface S_Π at point p, endowed with the induced metric. It is denoted by $K(X, Y)$ or $K(\Pi)$. The sectional curvature can be expressed as $K(X, Y) = \frac{Rm(X,Y,Y,X)}{|X|^2|Y|^2 - \langle X, Y \rangle^2}$ where Rm is the Riemann tensor.

The Gauss curvature of S_Π is defined as follows. If II is the second fundamental form, the form operator s can be defined as: for all vector fields X and Y, $\langle X, sY \rangle = \langle II(X, Y), N \rangle$ where N is a unit normal vector field of the surface S_Π (the choice of an orientation defined a unique unit normal vector field). The Gauss curvature of S_Π (i.e., the sectional curvature) is $K(\Pi) = \det s = \kappa_1\kappa_2$ with κ_1, κ_2 the eigenvalues of s called the principal curvatures.

We may also be interested in the geodesic curvature of the curve $r : t \mapsto (\mu(t), \Sigma(t))$. If the velocity of the curve has unit norm, the geodesic curvature is defined as $\kappa(t) = |D_t \dot{r}(t)|$. If the the velocity of the curve has not unit norm, the curve must be reparametrized to get a velocity with unit norm. It is also possible to use this formula: $\kappa(t) = \frac{\sqrt{|\dot{r}(t)|^2 |D_t \dot{r}(t)|^2 - \langle D_t \dot{r}(t), \dot{r}(t) \rangle^2}}{|\dot{r}(t)|^3}$.

5.1 Solution of the SDE

We rely on the book of Särkkä and Solin [4].

The solution $\theta(t)$ is a Gaussian process, i.e., $\theta(t) \sim \mathcal{N}(m(t), \Sigma(t))$ whose moments verify the following differential equations:

$$\frac{dm}{dt} = Am + b,$$
$$\frac{d\Sigma}{dt} = AP + PA^T + Q.$$

The solutions are:

$$m(t) = \exp(A(t-t_0))m(t_0) + b\int_{t_0}^{t} \exp(A(t-\tau))d\tau,$$

$$P(t) = \exp(A(t-t_0))P(t_0)\exp(A(t-t_0))^T + \int_{t_0}^{t} \exp(A(t-\tau))Q\exp(A(t-\tau))^T d\tau.$$

This SDE is equivalent to the following discrete system (in the sense that their solutions coincide at t_k):

$$x(t_{k+1}) = \exp(A(t_{k+1}-t_k))x(t_k) + b_k + q_k,$$

with $q_k \sim \mathcal{N}(0, \int_0^{t_{k+1}-t_k} \exp(A(t_{k+1}-t_k-\tau))Q\exp(A(t_{k+1}-t_k-\tau))^T d\tau)$.
Remark: we have[3] $A\int_{t_0}^{t} \exp(A(t-\tau))d\tau = \exp(A(t-t_0)) - I$, thus if A is nonsingular, we have $\int_{t_0}^{t} \exp(A(t-\tau))d\tau = A^{-1}(\exp(A(t-t_0)) - I)$.

5.2 Geometry of multidimensional Gaussian laws

We rely on the following references: a recent paper of Costa et al. [5] on the explicit computation of the Fisher-Rao distance for unidimensional normal distributions (and multidimensional distributions for some special cases), a paper of Skovgaard [6] on the Riemannian geometry of multidimensional normal laws, a paper of Atkinson and Mitchell [7] on the Fisher-Rao distance, and two papers of Calvo and Oller [8] [9], the former dealing with a lower bound of the Fisher-Rao distance for multidimensional normal ditributions, the later with the computation of geodesics for these distributions.

The densities have the following expression: $p(x;\mu,\Sigma) = (2\pi)^{-n/2}(\det\Sigma)^{-1/2}\exp(-(x-\mu)^T\Sigma^{-1}(x-\mu)/2)$.

5.2.1 Unidimensional Gaussian law

In the scalar case, the Fisher-Rao distance has a closed form:

$$d_{F^*}((\mu_1,\sigma_1),(\mu_2,\sigma_2)) = \sqrt{2}\log\frac{|(\frac{\mu_1}{\sqrt{2}},\sigma_1)-(\frac{\mu_2}{\sqrt{2}},-\sigma_2)|+|(\frac{\mu_1}{\sqrt{2}},\sigma_1)-(\frac{\mu_2}{\sqrt{2}},\sigma_2)|}{|(\frac{\mu_1}{\sqrt{2}},\sigma_1)-(\frac{\mu_2}{\sqrt{2}},-\sigma_2)|-|(\frac{\mu_1}{\sqrt{2}},\sigma_1)-(\frac{\mu_2}{\sqrt{2}},\sigma_2)|}.$$

It is the distance in the Poincaré half-plane (with a scaling on the first parameter μ).

5.2.2 Scalar covariance matrix

If $\Sigma = \sigma^2 I$, the parameter space $\mathbb{H}_F^{n+1}$ of dimension $n+1$ can be parameterized by (μ,σ). The FIM is:

$$_{\mathcal{N}}g = \mathrm{diag}\left(\frac{1}{\sigma^2},\ldots,\frac{1}{\sigma^2},\frac{2n}{\sigma^2}\right).$$

The parameter space $\mathbb{H}_F^{n+1}$ can be transformed into the Poincaré half-space $\mathbb{H}^{n+1}$ through the map $(\mu,\sigma) \mapsto (\mu/\sqrt{2n},\sigma)$. The metric is then $g = \mathrm{diag}\left(\frac{1}{\sigma^2},\ldots,\frac{1}{\sigma^2},\frac{1}{\sigma^2}\right)$. The Fisher-Rao distance can be computed explicitly (the formula can be found in [5]). The geodesics are contained in orthogonal hyperplanes of the hyperplane $\sigma = 0$ and are either straight lines ($\mu =$constant), either half-ellipsis of eccentricity $1/\sqrt{2}$ centered in the hyperplane $\sigma = 0$ (in the Poincaré half-space, these ellipsis become circles). The curvature is constant and equal to $-\frac{1}{n(n+1)}$.

5.2.3 Diagonal covariance matrix

In the case $\Sigma = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_n^2)$, the parameter space (independent normal laws) is the intersection of n half-spaces in $\mathbb{R}^{2n}$. By denoting the parameters $(\mu_1,\sigma_1,\ldots,\mu_n,\sigma_n)$, the FIM is:

$$_{\mathcal{N}}g = \mathrm{diag}\left(\frac{1}{\sigma_1^2},\frac{2}{\sigma_1^2},\ldots,\frac{1}{\sigma_n^2},\frac{2}{\sigma_n^2}\right).$$

[3] through the power series expansion of the exponential.

The metric is then the product metric and it is possible to obtain a closed form for the Fisher-Rao distance. The curvature is constant equal to $-\frac{1}{2(2n-1)}$.

5.2.4 General case

The general case is way harder. Indeed, the sectional curvature of M is not constant. In the case $n = 2$, the distributions can be parameterized with $(\sigma_1, \sigma_2, \mu_1, \mu_2, u)$ where σ_1^2 and σ_2^2 are the eigenvalues of Σ, et u is the angle between the canonical basis and the basis of eigenvectors. With this parametrization, the FIM can be computed explicitly but not the Fisher-Rao distance. Numerical methods can be used to approximate the Fisher-Rao distance, e.g., using a symmetrized variant of the KL divergence to approximate the Fisher-Rao distance between not too distant points. Bounds on the Fisher-Rao distance can also be obtained by embedding M isometrically in the manifold of positive definite matrices endowed with the Siegel metric (see next paragraph §5.3). An expression for the Fisher-Rao distance can be obtained when the mean is fixed.

5.3 Bounds on the Fisher-Rao distance

5.3.1 Lower bound

We rely here on Calvo and Oller [8]. Consider the manifold of symmetric positive definite matrices $\mathcal{S}_{n+1}^+$ endowed with the Siegel metric:

$$ds^2 = \frac{1}{2}\text{tr}((P^{-1}dP)^2),$$

with $P \in \mathcal{S}_{n+1}^+$. This metric is invariant by the action $P \mapsto WSW^T, W \in GL_{n+1}(\mathbb{R})$. The associated distance is:

$$d(P_1, P_2) = \frac{1}{\sqrt{2}}\|\ln(P_1^{-1/2}P_2P_1^{-1/2})\|_F = \sqrt{\frac{1}{2}\sum_{i=1}^{n+1}(\log\lambda_i)^2},$$

already introduced in paragraph §2.2.2, with λ_i the eigenvalues of $P_1^{-1/2}P_2P_1^{-1/2}$. In the following, this distance will be called the Siegel distance. We are going to embed M in $\mathcal{S}_{n+1}^+$ with the following map[4]:

$$\begin{aligned} f : M &\longrightarrow \mathcal{S}_{n+1}^+ \\ (\mu, \Sigma) &\mapsto \begin{bmatrix} \Sigma + \mu\mu^T & \mu \\ \mu^T & 1 \end{bmatrix}. \end{aligned}$$

. It can be shown that:

- f is an embedding.
- $f(M)$ is a submanifold of dimension $\frac{(n+1)(n+2)}{2} - 1$, one less than $\mathcal{S}_{n+1}^+$.
- The induced metric on $f(M)$ is:

  $$ds^2 = \frac{1}{2}\text{tr}((\Sigma^{-1}d\Sigma)^2) + d\mu^T\Sigma^{-1}d\mu.$$

- This metric is the FIM itself. In other words, M and $f(M)$ are isometric. Moreover, the FIM is invariant under the action of the affine group $(X \in M \mapsto QX + c, Q \in GL_n(\mathbb{R}))$.
- $f(M)$ is not totally geodesic, i.e., the geodesics of $\mathcal{S}_{n+1}^+$ do not remain on $f(M)$. This result, along with the previous result, means that **the Siegel distance is a lower bound for the Fisher-Rao distance.**
- However, if we consider M_{μ_0} the submanifold of M of normal laws with fixed mean equal to μ_0, then $f(M_{\mu_0})$ is a totally geodesic submanifold with dimension $\frac{n(n+1)}{2}$. This means that, if the mean is fixed, then the Fisher-Rao distance is equal to the Siegel distance. This result was already shown by Burbea [10].

[4] one could have chosen any maps of the form $f_P(\mu, \Sigma) = Pf(\mu, \Sigma)P^T$ thanks to the invariance of the Siegel metric.

There are two interpretations of this embedding[5]. The first interpretation consists in associating a normal law $X \sim \mathcal{N}(\mu, \Sigma)$ with an element of the affine group of dimension n, more precisely the element that sends the standard normal distribution $Z \sim \mathcal{N}(0, I)$ to X. This element is $A_X = \begin{bmatrix} \Sigma^{1/2} & \mu \\ 0^T & 1 \end{bmatrix}$ since one has:

$$\begin{bmatrix} \Sigma^{1/2} & \mu \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} Z \\ 1 \end{bmatrix} = \begin{bmatrix} X \\ 1 \end{bmatrix}.$$

We get the embedding in $\mathcal{S}_{n+1}^{+}$ with $f(X) = A_X A_X^T$.
For the second interpretation, consider a random vector $(X, Y) \sim \mathcal{N}_{n+1}(0, \Psi)$ where X has dimension n and Y has dimension 1. Then, $\Psi \in \mathcal{S}_{n+1}^{+}$. If the covariance of Y is set to 1, then it can be shown that, by denoting $X|Y = 1 \sim \mathcal{N}_n(\mu, \Sigma)$, the matrix Ψ is equal to $f(\mu, \Sigma)$.
For more details, see Burbea [10] or *Symplectic Geometry* of Siegel.

5.3.2 Upper bounds

Now, we focus on the paper of Pinele et al. [11] in which three upper bounds are proposed.

As mentioned in the previous paragraph §5.3.1, the FIM is invariant under the action of the affine group, i.e., for all $(c, Q) \in \mathbb{R}^n \times GL_n(\mathbb{R})$, the following map is an isometry:

$$\begin{aligned} \Psi_{(c,Q)} : M &\longrightarrow M \\ (\mu, \Sigma) &\mapsto (Q\mu + c, Q\Sigma Q^T). \end{aligned}$$

Thus, the Fisher-Rao distance is invariant under affine transformation. In particular, given $\theta_1 = (\mu_1, \Sigma_1)$ and $\theta_2 = (\mu_2, \Sigma_2)$, consider $\theta_0 = (0, I_n)$ and $\theta_3 = (Q\mu_2 + c, Q\Sigma_2 Q^T)$ with $Q = \Sigma_1^{-1/2}$ and $c = -\Sigma_1^{-1/2}\mu_1$. Then, we have $d_F(\theta_1, \theta_2) = d_F(\theta_0, \theta_3)$ where d_F is the Fisher-Rao distance. We diagonalize $A = \Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} = Q\Lambda Q^T$ with Q an orthogonal matrix whose columns are the eigenvectors of A, and λ_i are the diagonal elements of Λ. Let $\mu = Q^T \Sigma_1^{-1/2}(\mu_2 - \mu_1)$. A first upper bound is:

$$UB_1(\theta_1, \theta_2) = \sqrt{\sum_{i=1}^{n} d_{F^*}^2((0, 1), (\mu_i, \lambda_i))},$$

where d_{F^*} was introduced in the paragraph §5.2.1. This upper bound is the distance on the submanifold of diagonal covariance matrices (§5.2.3), whose metric is the product metric of n Poincaré half-planes.

The two other upper bounds are obtained with a well chosen $\bar{\theta}$ in the triangle inequality $d_F(\theta_0, \theta_3) \leq d_F(\theta_0, \bar{\theta}) + d_F(\bar{\theta}, \theta_3)$.

6 Process with delays

In this section, consider $\theta \in \mathbb{R}^p$ verifying the following SDE:

$$d\theta(t) = \left(\begin{bmatrix} a_1 & \dots & \dots & a_p \\ & a_1 & \dots & a_{p-1} \\ & (0) & \ddots & \vdots \\ & & & a_1 \end{bmatrix} \theta(t - \tau) + \begin{bmatrix} 0 & & & \\ a_p & & (0) & \\ \vdots & \ddots & \ddots & \\ a_2 & \dots & a_p & 0 \end{bmatrix} \theta(t - (p+1)\tau) + b\mathbf{1} \right) dt + c\mathbf{1}^T \begin{bmatrix} dw(t) \\ \vdots \\ dw(t - (p-1)\tau) \end{bmatrix},$$

where $\mathbf{1} \in \mathbb{R}^p$ is a vector full of 1, an $\tau > 0$ is a fixed delay. Considering only the first component of θ, the previous equation is equivalent to:

$$ds(t) = \left(\sum_{i=1}^{p} a_i s(t - i\tau) + b \right) dt + c\,dw(t).$$

[5] See James for more details.

7 Objectives

- Check the computation of the curvature in the unidimensional case.
- Geometric study of the linear SDE of dimension n. Literature review: Siegel half-space, computation/bounds of the curvature (calculus of variations, Grönwall lemma).
- Sensitivity of a perturbation on the initial conditions (θ_0).
- Study of the non-linear case with a linearization (Taylor expansion at first order).
- In order to visualize the curvature, plot the curves in the Poincaré disk, or in the Klein disk.

References

[1] K. Sun, *Information Geometry and Data Manifold Representations.* PhD thesis, Université de Genève, 2015.

[2] D. Brooks, *Deep Learning and Information Geometry for Time-Series Classification.* PhD thesis, Sorbonne Université, 2020.

[3] A. Bay and B. Sengupta, "Geoseq2seq: Information geometric sequence-to-sequence networks," *arXiv*, 2018.

[4] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations.* Cambridge University Press, first ed., Apr. 2019.

[5] S. I. Costa, S. A. Santos, and J. E. Strapasson, "Fisher information distance: A geometrical reading," *Discrete Applied Mathematics*, vol. 197, pp. 59–69, Dec. 2015.

[6] L. T. Skovgaard, "A Riemannian Geometry of the Multivariate Normal Model," *Scandinavian Journal of Statistics*, vol. 11, no. 4, pp. 211–223, 1984.

[7] C. Atkinson and A. F. S. Mitchell, "Rao's Distance Measure," *Sankhya: The Indian Journal of Statistics, Series A*, vol. 43, pp. 345–365, Oct. 1981.

[8] M. Calvo and J. M. Oller, "A distance between multivariate normal distributions based in an embedding into the Siegel group," *Journal of Multivariate Analysis*, vol. 35, pp. 223–242, Nov. 1990.

[9] M. Calvo and J. M. Oller, "An Explicit Solution of Information Geodesic Equations for the Multivariate Normal Model," *Statistics & Risk Modeling*, vol. 9, Jan. 1991.

[10] J. Burbea, "Informative Geometry of Probability Spaces:," Tech. Rep. 84-52, Center for Multivariate Analysis, University of Pittsburgh, Dec. 1984.

[11] 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, Apr. 2020.