Signature

1 Signature et apprentissage.

Les séries temporelles multidimensionnelles sont omniprésentes dans les applications réelles et leur traitement nécessite une attention particulière. Les dépendances entre les composantes d'une série, l'aspect séquentiel mais aussi la masse de données sont autant de difficultés qui poussent la recherche vers de nouvelles méthodes pour exploiter au mieux toute l'information contenue dans les séries. On introduit pour cela un objet mathématique bien connu du calcul stochastique, la signature. On cherchera à comprendre pourquoi il suscite un intérêt croissant dans le domaine de l'apprentissage automatique des séries multivariées et on discutera de quelques méthodes parmi les plus récentes pour son intégration pratique dans des algorithmes.

1.1 Path

Définition 1. *Un chemin est une fonction continue de* $[a,b]$ *dans* $\mathbb{R}^d$

Un chemin est classiquement paramétré par le temps, ce que l'on note alors $X : [a,b] \to \mathbb{R}^d$, $t \mapsto \big(X^1(t), ..., X^d(t)\big)$. $X(t)$ ou de manière équivalente $X_t = (X_t^1, ..., X_t^d)$

Définition 2. *Let* $X : [a,b] \to \mathbb{R}$ *un chemin 1-dimensionnel* $f : \mathbb{R} \to \mathbb{R}$ *une fonction. L'intégrale de chemin de* f *contre* X *est définie comme* $\int_a^b f(X_t)dX_t = \int_a^b f(X_t)\dot{X}_t dt$*, où la dernière intégrale étant par exemple l'intégrale de Lebesgue.*

Pour un chemin multidimensionnel $X : [a,b] \to \mathbb{R}^d$ et une fonction $f : \mathbb{R}^d \to \mathbb{R}$, l'intégrale de chemin est définie comme $\int_a^b f(X_t)dX_t = \int_a^b f(X_t)\|\dot{X}_t\|dt$, où $\|.\|$ est la norme euclidienne.

On remarque dans la Définition 2 que, $f(X(.))$ est elle même un chemin de $[a,b]$. On peut aussi définir l'intégrale d'un chemin $X : [a,b] \to \mathbb{R}$ contre un autre chemin $Y : [a,b] \to \mathbb{R}$ comme :

$$\int_a^b Y_t dX_t = \int_a^b Y_t \dot{X}_t dt$$

1.2 Signature d'un chemin

Définition 3. *Soit* $[a,b]$ *et* $X = (X^1, ..., X^d) : [a,b] \to \mathbb{R}^d, t \mapsto X(t) = X_t = (X_t^1, ..., X_t^d)$ *un chemin continu par morceau. La signature de* X *est définie comme la collection des intégrales itérées suivantes.*

$$Sig(X) = \left(\int_{a<t_1<...<t_k<b} ... \int dX_{t_1} \otimes ... \otimes dX_{t_k} \right)_{k\geq 0} = \left(\left(\int_{a<t_1<...<t_k<b} ... \int dX^{i_1}_{t_1} ... dX^{i_k}_{t_k} \right)_{1\leq i_1,...,i_k\leq d} \right)_{k\geq 0}$$

où le terme pour $k = 0$ est 1 par convention.

Cette définition nous pousse à définit l'algèbre tensorielle de $\mathbb{R}^d$. Comme on peut le constater, le niveau 0 de la signature est 1, le premier niveau est un vecteur de $\mathbb{R}^d$, le second niveau $\mathbb{R}^d \otimes \mathbb{R}^d$, ..., et ainsi de suite.

Définition 4. *On appelle algèbre tensorielle de $\mathbb{R}^d$:*

$$T((\mathbb{R}^d)) = \prod_{k=0}^{\infty} (\mathbb{R}^d)^{\otimes k}$$

L'algèbre tensorielle vient avec son produit tensoriel canonique, étendu par bilinéarité : On note pour cela $A = (A_0, A_1, \ldots) \in T((\mathbb{R}^d))$ et $B = (B_0, B_1, \ldots) \in T((\mathbb{R}^d))$.

$$A \otimes B = \left(\sum_{j=0}^{k} A_j \otimes B_{k-j} \right)_{k\geq 0}$$

Ce qui définit bien un élément de l'algèbre tensorielle.

Certains propriétés liés à cet opération seront détaillées par la suite.

Définition 5. *Étant donnés $i_1, ..., i_k$, la quantité donnée par*

$$S^{i_1,...,i_k}_{[a,b]}(X) = \int_{a<t_1<...<t_k<b} ... \int dX^{i_1}_{t_1} ... dX^{i_k}_{t_k}$$

est appelé le k-fold d'intégrales itérées de X selon les indices $i_1, ..., i_k$

Ce que l'on note $S^{i_1,...,i_k}_{[a,b]}(X)$ plus simplement comme $S^{i_1,...,i_k}(X)$ quand il n'y a pas d'ambiguité. Avec cette définition, la signature se décrit comme la suite infinie de toutes ses k-folds d'intégrales itérées :

$$Sig(X) = \{1, S^1(X), ..., S^d(X), S^{1,1}(X), S^{1,2}(X), ..., S^{d,d}(X), ...\}$$

C'est aussi l'ensemble des mots que l'on peut former avec les $\{1, \ldots, d\}$ disponibles. Formulation qui vient de la nature récursive (itérative) de la signature et de ses intégrales $(i, j) \in \{1, ..., d\}^2$:

$$S^{i,j}_{[a,b]}(X) = \int_a^b S^i_{[a,s]}(X) dX^j_s = \int_{a<r<s<b} dX^i_r dX^j_s$$

Dont on déduit la formule de $S^{i_1,...,i_k}(X)$ pour tout choix d'indices $i_1, ..., i_k$.

1.3 Geometric intuition of signature

In this section, let us give the reader an intuition of what the first levels of signature yield. At $k = 1$, each of the terms $S^i(X) = \int_a^b \dot{X}_t^i dt = X_t^i(b) - X_t^i(a)$ is simply the increment of the path on dimension i over the interval $[a, b]$.

Let us now consider two indices $(i, j) \in \{1, ..., d\}^2$.
If $i = j$, we have $S^{i,i}(X) = \int_a^b X_t^i dX_t^i = (X_b^i - X_a^i)^2/2$. In the case where $i \neq j$, we have :
$S^{i,j}(X) = \int_{a<r<s<b} dX_r^i dX_s^j$, which can be seen as the area located below the parameterized curve
$\{(X_t^i, X_t^j), t \in [a, b]\}$, and delimited by the time interval and the initial point of the path (see Figure **??**). Conversely, $S^{j,i}(X)$ is the area located above the curve, and delimited by the time interval and the final point of the path.

At the second order level, another interesting geometric interpretation is given by the Levy area of a curve. Levy area is the signed area enclosed by the path and the chord connecting the end-points.

$$A_{a,b} = S^{i,j}(X)_{a,b} - S^{j,i}(X)_{a,b}$$

One very interesting geometric interpretation we can guess easily at low levels is shuffle product identity.

$$S^i(X)_{a,b} \cdot S^j(X)_{a,b} = S^{i,j}(X)_{a,b} + S^{i,j}(X)_{a,b}$$

That is to say : nonlinear transformation at low levels can be expressed as linear combination of higher levels. It's a foretaste of the universal non-linearity that will introduce later.

Essentially, the iterated integrals each capture different geometrical features of the path, and the collection of them all (i.e. the signature of the path) defines the path in a very efficient way. As we get higher in the order of the iterated integral, we will capture more subtle features of the path. With this definition, signature is a convenient, but infinite collection of scalars. To make it exploitable, we will then truncate the signature at order $N \in \mathbb{N}$, thus defining

$$S_N(X) = \left(\int\limits_{a<t_1<...<t_k<b} ... \int dX_{t_1} \otimes ... \otimes dX_{t_k} \right)_{0 \leq k \leq N} = \{S^I(X)\}_{I \in \{\{1,...,d\}^k, 0 \leq k \leq N\}}$$

still with the convention that $S^0(X) = 1$

1.4 Signature, ODE and CDE.

1.4.1 Integral iteration through ODE.

As a first witness of the mathematical importance of signature, we briefly show how signature comes up in ordinary differential equations. We start with a simple example of integral iteration.

The well-known Cauchy-Lipschitz theorem states on existence of an unique solution of any initial value problem. Using math :

$$y'(t) = f(t, y(t)), \quad y(t_0) = y_0$$

Has an unique solution since f is uniformly Lipschitz continuous in y, in a neighborhood of t_0. The classical proof scheme makes use of integral notation of the ODE then solving a fixed-point problem, relying on Banach fixed-point theorem, eventually on Grönwall's lemma for uniqueness.

$$\int_{t_0}^{t} f(s, y(s))ds = y(t) = y(t_0)$$

If we set Γ as :

$$\Gamma_0(t) = y_0, \quad \Gamma_{k+1}(t) = y_0 + \int_{t_0}^{t} f(s, \Gamma_k(s))ds$$

Γ (known as Picard iteration) would be the candidate of the fixed-point functionnal problem, since a fixed-point yields a solution of the ODE above. Let's consider this simplest initial value problem.

$$y'(t) = 1 + y(t)^2, \quad y(t_0) = 0 = y_0, t_0 = 0$$

Which a solution is $y(t) = \tan(t)$. Now we start practical computations of Picard iterations.

$$\Gamma_0 = y_0 = 0 \Gamma_1(t) = \int_0^t ds = t \Gamma_2(t) = \int_0^t 1 + s^2 ds = t + \frac{t^3}{3} \Gamma_3(t) = \int_0^t 1 + s^2 ds = t + \frac{t^3}{3} + \frac{2t^5}{15} + \frac{t^7}{63}$$

and so on. Indeed, we are computing the Taylor series expansion of the solution.

1.4.2 Signature as integral iteration : solving CDE.

Now we consider a path $X : [a, b] \xrightarrow{\mathbb{R}}{}^{d}$, $Z : [a, b] \xrightarrow{\mathbb{R}} (\mathbb{R}^d, \mathbb{R}^e)$. Where L denotes the space of linear maps. Accordingly with the previous path definition, we define the integral

$$\int_a^b Z_t dX_t$$

which belongs to $\mathbb{R}^e$. Z is said to be a controlled path.
Let's consider a controlled differential equation, that is to say :

$$dY_t = V(Y_t)dX_t, \quad Y_a = y$$

That means, just as before with ODE, in an integral way :

$$Y_t = y + \int_a^t V(Y_s)dX_s$$

For every t in $[a,b]$. Just as with ODE, we consider Picard iteration with the Γ functionnal defined as :

$$\Gamma_k(Y)_t = y + \int_a^t V(Y_s)dX_s$$

Just as with ODE, finding a fixed-point of this functional leads to a solution of the previous CDE. For ease we suppose that V is a linear map (from $\mathbb{R}^e \xrightarrow{\mathbb{R}} (\mathbb{R}^d, \mathbb{R}^e)$). Computing first iterations gives us :

$\Gamma_0(Y)_t = y$

$\Gamma_1(Y)_t = y + \int_a^t dV(X_s)dX_s$

$\Gamma_2(Y)_t = y + \int_a^t V(\Gamma_1(Y)_s)dX_s = y + \int_a^t dX_s + \int_a^t \int_a^s dV(X_u)dV(X_s)$

$\Gamma_3(Y)_t = y + \int_a^t V(\Gamma_2(Y)_s)dX_s = y + \int_a^t dX_s + \int_a^t \int_a^s dV(X_u)dV(X_s) + \int_a^u \int_a^t \int_a^s dV(X_v)dV(X_u)dV(X_s)$

...

and so on. First level of the $V(X)$ signature comes out. Moreover, the signature tends to completely determined the solution. And this procedure fastly converges.

Theorem 1. *Factorial decay. Let $X[0,T] \to V$ be a path. Then $\forall k \geq 1$, one has*

$$|\int_{0<u_1<\ldots<u_k<T} dX_{u_1} \otimes \ldots \otimes dX_{u_k}| \leq \frac{\|X\|^k_{1,[0,T]}}{k!}$$

Fast convergence is crucial : it allows us early truncature, one should have in mind that the calculation grows exponentially with the level of the signature.

From what, we deduce for the previous CDE fast convergence to the solution. We inform that first, this results extend to non-linear drivers V [**?**], and secondly we were limited until now to piecewise differentiable path even though the signature has sense for more general frameworks in the names of path of bounded p-variations and rough path theory, provided some adaptations (in particular, iterated integrals for Young integral are not uniquely defined). The latter leading to a central theory in SDE. [**?**]

This fundamental example of signature being part of a mathematical theory intimates us that the information that the signature carries is of utmost importance.
Since we aim to use the signature in machine learning, we recall that practical uses needs truncated signature and not the actual one.

The next part of this course tends to show the most important properties of the signature. Most of them justify its growing importance in machine learning by catching crucial information, leading to a powerful non-parametric dimension reduction method.

2 Some properties of the signature

2.1 Chen's identity

Recall that within the tensor algebra, we have at our disposal the canonical tensor product. Chen's identity argue that the signature of concatenation of two paths is the tensor product of

individual signature.

Theorem 2. *Chen's identity TO DO*

Chen's identity yields that signature computation suits stream data very well : we don't have to know the full stream to start compute the signature, we can compute following the stream flowing and compute the tensor product between different pieces to retrieve the full signature.

2.2 Time-reversal path and signature

Firstly we need to introduce the notion of time-reversal of a path.

Définition 6. *Let $X : [a, b] \overset{\mathbb{R}}{\to}^{d}$ be a path. We define the time-reversal path $\overleftarrow{X}$ as the path for which $\overleftarrow{X}_t = X_{a+b-t}$*

Then we state that

Theorem 3.

$$S(X) \otimes S(\overleftarrow{X}) = 1$$

That is, the time-reversal of a path is an inverse for the path considering the tensor product.

2.3 Invariance to time reparametrisations

Invariance to time reparametrisations within signature is a straightforward consequence of integral properties and parametrisation. Let X, Y be two paths, ψ a parametrisation.

$$\int_a^b Y_{\psi(t)} dX_{\psi t} = \int_a^b Y_{\psi(t)} \dot{X}_{\psi(t)} \psi'(t) dt = \int_a^b Y_x dX_x$$

These three first properties forbid signature transform to be invertible, even before any truncation, in contrast to Fourier on L^2 transform for example. However, one should know that we can in a certain way construct a stream from its signature [**?**], up to a certain equivalence class : the tree-like equivalence.

2.4 Tree-like path and uniqueness

Définition 7. *Tree-like path A path $X : [0, 1] \to \mathbb{R}^d$ is tree-like if there exists a continuous function $h : [0, 1] \to [0, +\infty[$ such that $h(0) = h(1) = 0$ and such that for all $s, t \in [0, 1], s \leq t$*

$$\|X_s - X_t\| \leq h(s) + h(t) - 2 \inf_{u \in [s,t]} h(u)$$

That definition yields an equivalence relation [**?**] :

Définition 8. *Tree-like equivalence Let X, Y be two paths. There are tree-like equivalent if $X * \overleftarrow{Y}$ is tree like.*

It follows immediately from Chen's identity that any path concatenated with its time-reversal is tree-like. This kind of symmetry is unavoidable in machine learning problems, an efficient way to make the signature still relevant in thoses cases is to consider the expanded path $((t, X_t))_t$

instead of $(X_t)_t$. More generally, any path which has an monotonic component is not tree-like. Now we state the main results of uniqueness of the signature. This result is also valid for finite p-variations path [?].

Theorem 4. *Uniqueness up to tree-like equivalent Let X be a continuous path with finite variations. We have the following equivalence :*

$$X \text{is tree-like} \iff Sig(X) = 1$$

Corollary 1. *For X a continuous path with finite variations. there exists a unique path of minimal length $\bar{X}$ called the reduced path, sharing the same signature.*

2.5 Universal nonlinearity

A very powerful inherent property of signature is that the signature of a continuous transformation of a stream, belonging to a compact set, can be approximated arbitrary close by a linear function of the initial stream signature. As a reminder, a memorable point of signature method as been reached when it has achieved state-of-the art performances in hand writing recognition. In this case, the continuous transformation of a stream can be interpreted as the natural fluctuation of writing. Let's put mathematical words on it.

Theorem 5. *Universal non linearity*
Let ϕ be a continuous transformation, ie a real-valued continuous function on continuous piecewise smooth paths in $\mathbb{R}^d$ and let $\mathbb{K}$ be a compact set of such paths. Furthermore we assume that $X_0 = 0$ for all $X \in K$. Let $\epsilon > 0$. Then there exists a linear functional L such that for all $X \in K$

$$|\phi(X) - L(S(X))| < \epsilon$$

3 Rappels sur les RKHS et «l'astuce du noyau».

On commence par rappeler brièvement les notions et résultats classiques portant sur les RKHS et les noyaux.[?]

Définition 9. *Noyau*
*Soit $\mathcal{X}$ un ensemble non vide. Une fonction k de $\mathcal{X} \times \mathcal{X}$ est appelée **noyau** s'il existe un espace de Hilbert $\mathcal{G}$ et une fonction $\phi : \mathcal{X} \to \mathcal{G}$ telle que :*

$$\forall (x, x^{'}) \in \mathcal{X}^2 : \quad k(x, x^{'}) = \langle \phi(x), \phi(x^{'}) \rangle_{\mathcal{G}}$$

De même, on appelle $\mathcal{G}$ espace de redescription, et ϕ fonction de redescription.

Définition 10. *Reproducing Kernel Hilbert Space RKHS*

Soit $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ un espace de Hilbert. Soit k un noyau défini sur $\mathcal{X}^2$. On dit que $\mathcal{H}$ est un RKHS de noyau k si pour tout $x \in \mathcal{X}$, $k(\cdot, x)$ est dans $\mathcal{H}$ et si on a la propriété dite de reproduction :

$$\forall f \in \mathcal{H} : \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$$

On rappelle brièvement quelques résultats essentiels. Premièrement à tout noyau est associé un unique RKHS (c'est le théorème de Moore-Aronszajn). Ensuite on a une description bien précise de cet espace $\mathcal{H}$ au moyen du noyau.

Théorème 1. *Soit $\mathcal{H}$ un RKHS de noyau k. Alors :*

$$\mathcal{H} = \left\{ \sum_{i=1}^{\infty} \alpha_i k(\cdot, x_i) : \sum_{i=1}^{\infty} \alpha_i^2 k(x_i, x_i) < \infty, x_i \in \mathcal{X}, \alpha_i \in \mathbb{R} \right\}$$

Enfin on présente un dernier résultat qui justifie la pertinence du modèle :

Théorème 2. *du représentant*
Soit k un noyau sur $\mathcal{X}^2$, $X \in (\mathcal{X}, \mathcal{Y})^N$ un échantillon, R un risque empirique quelconque et ψ une fonction positive strictement croissante (fonction de régularisation). Alors toute fonction $\hat{h}$ qui est solution du problème :

$$\min_{h \in \mathcal{H}} R(h) + \lambda \psi(h)$$

admet une représentation portée par les données, ie :

$$\hat{h}(\cdot) = \sum_{i=1}^{N} \alpha_i k(X_i, \cdot)$$

De manière informelle, on peut restreindre la minimisation d'un risque empirique dans un RKHS à un domaine beaucoup plus simple, qui ne dépend que des données.

On dispose d'un choix canonique d'espace de redescription (pourvu de sa fonction de redescription), en posant :

$$\phi(x) = k(\cdot, x)$$

Ainsi l'espace de redescription $\mathcal{G}$ est ici égal à $\mathcal{H}$ et par la propriété de reproduction il découle :

$$\forall (x, x^{'}) \in \mathcal{X}^2, \quad \langle \phi(x), \phi(x^{'}) \rangle_{\mathcal{G}} = \langle k(\cdot, x), k(\cdot, x^{'}) \rangle_{\mathcal{H}} = k(x, x^{'})$$

Dans cette configuration, l'astuce du noyau est la suivante : nos données vivent dans l'espace $\mathcal{X}$. On soupçonne (ou on espère) qu'un espace de redescription existe et que celui-ci permettra une meilleure représentation de nos données. Dés lors que notre algorithme de machine learning préféré n'a besoin que des produits scalaires entre les données pour fonctionner, et que l'on dispose d'un noyau dont le calcul effectif n'est pas trop coûteux, on peut utiliser nos algorithmes sur nos données vivant dans l'espace $\mathcal{H}$ dont la dimension est potentiellement bien plus grande, voire infinie ($\mathcal{H}$ est un espace de fonction).
Par ailleurs le théorème du représentant nous garantie que la solution de notre problème favori s'exprimera à partir de nos données ponctuelles.

En particulier, et c'est ce qui fait l'efficacité des méthodes reposants sur cette astuce, dés lors que le noyau n'est pas linéaire, on est en mesure d'apprendre des classifieurs non-linéaires avec des méthodes linéaires. La redescription est implicite : on ne cherche pas à calculer les coordonnées des données dans la nouvelle représentation, tout passe par la fonction noyau.

Les algorithmes classiques qui sont compatibles avec cette approche sont les SVM, l'ACP, la SVR (ou ridge regression) et bien d'autres.

4 La signature et les méthodes à noyau.

Pour pouvoir tirer profit de l'astuce du noyau sur des données séquentielles, il faut disposer d'un noyau sur l'espace des chemins. C'est le cœur de l'article *Kernels for Sequentially Ordered Data* [**?**] qui propose une construction d'un tel noyau utilisable en pratique pour des chemins à valeurs dans $\mathcal{X}$ partant de n'importe quel noyau de $\mathcal{X}$.

Partant donc d'un noyau k «statique», on va passer de chemins à valeurs dans $\mathcal{X}$ à des chemins à valeurs dans $\mathcal{H}$ l'espace de redescription canonique associée à k. Notre fonction de redescription ne sera rien d'autre que la signature.

L'astuce du noyau dans ce cas là se résumera ainsi de manière informelle : si l'on veut utiliser nos algorithmes favoris de ML, il faudrait être en mesure de calculer une matrice de Gram contenant tous les produits scalaire des différentes signatures (pour un produit scalaire qui reste à définir). Une analyse astucieuse nous montrera qu'il n'est pas nécessaire de calculer ni de stocker les signatures de nos échantillons : on peut construire un noyau $k^{\oplus}$ par dessus k, et son évaluation est peu coûteuse en ressources de calculs.

On propose d'abord de regarder en détail la construction de $k^{\oplus}$. Ensuite on présente des résultats issu de l'article *The signature Kernel is the solution of a Goursat PDE*, qui montre que le noyau $k^{\oplus}$ est étroitement lié à une famille d'EDP hyperboliques ouvrant la voie à l'utilisation de signatures non-tronquées ainsi qu'à d'autres méthodes de calculs.

4.1 Produit scalaire de signature et noyau pour chemins

Reprenons notre feuille de route : on part d'un noyau k sur $\mathcal{X}$ (par exemple un noyau gaussien), et on veut en tirer un noyau pour $\mathcal{P}_{\mathcal{X}}$, l'espace des chemins à valeurs dans $\mathcal{X}$.

On «passe» d'abord notre chemin dans son espace de représentation canonique.

$$\begin{array}{rcl} \mathcal{P}_{\mathcal{X}} & \to & \mathcal{P}_{\mathcal{H}} \\ t \mapsto x(t) & \mapsto & t \mapsto k_x(t) = k(x(t), \cdot) \end{array}$$

Puis on considère l'évaluation d'un certain produit scalaire lui même défini sur l'espace des signatures des chemins dans $\mathcal{P}_{\mathcal{H}}$. [1]

$$\begin{array}{rcl} \mathcal{P}_{\mathcal{H}} \times \mathcal{P}_{\mathcal{H}} & \to & \mathbb{R} \\ (k_x, k_y) & \mapsto & \langle S(k_x), S(k_y) \rangle \end{array}$$

1. On remarque que l'on n'utilise pas une définition alternative du noyau vue comme fonction définie positive car la construction explicite à partir du produit scalaire est possible. Ce n'est pas toujours le cas quand on utilise des méthodes à noyaux.

où

$$\begin{aligned}\langle S(k_x), S(k_y)\rangle &= \sum_{m\geq 0} \langle S^m(k_x), S^m(k_y)\rangle_{\mathcal{H}^{\otimes m}} \\ &= \sum_{m\geq 0} \prod_{i=0}^{m} \langle S_i^m(k_x), S_i^m(k_y)\rangle_{\mathcal{H}} \\ &\text{où } S^m(k_x) = S_1^m(k_x) \otimes \ldots \otimes S_m^m(k_x) \in \mathcal{H}^{\otimes m}\end{aligned}$$

Finalement il ne reste qu'à tout mettre bout à bout pour obtenir notre noyau à signature $k^{\oplus}$:

$$\begin{aligned}k^{\oplus} : \mathcal{P}_{\mathcal{X}} \times \mathcal{P}_{\mathcal{X}} &\to \mathbb{R} \\ (x, y) &\mapsto k^{\oplus}(x, y) = \langle S(k_x), S(k_y)\rangle\end{aligned}$$

Notons que la signature est bien définie comme limite de sommes de Riemmann-Stieltjes, ici l'intégrale prend ses valeurs dans $\mathcal{H}$, voir par exemple Differential Equations Driven by Rough Paths.

4.2 Stratégies de calcul de la matrice de Gram

On commence par montrer que le noyau $k^{\oplus}$ s'écrit sous une forme récursive qui s'avérera déterminante pour les considérations pratiques.

Théorème 3. *Pour tout x, y dans $\mathcal{P}$:*

$$k^{\oplus}(x, y) = \sum_{m\geq 0} \int_{\substack{s_1<\ldots<s_m \\ t_1<\ldots<t_m}} \prod_{i=1}^{m} d\kappa(s_i, t_i)$$

où κ est une mesure signée sur $[0, 1]^2$ définie comme :

$$\kappa([s, t] \times [u, v]) = k(x(t), y(v)) - k(x(s), y(v) - k(s(t), y(v)) + k(x(s), y(u))$$

On en déduit la forme récursive :

$$k^{\oplus}(x, y) = 1 + \int_{(s_1,t_1)\in[0,1]^2} \left(1 + \int_{(s_2,t_2)\in[0,s_1]\times[0,t_1]} (1 + \ldots) d\kappa_{x,y}(s_2, t_2)\right) d\kappa_{x,y}(s_1, t_1)$$

Preuve :

$$
\begin{aligned}
k^{\oplus}(x,y) &= \langle S(k_x), S(k_y)\rangle \\
&= \sum_{m\geq 0} \langle \int dk_x^{\otimes m}, \int dk_y^{\otimes m}\rangle_{\mathcal{H}^{\otimes m}} \\
&= \sum_{m\geq 0} \langle \int_0^1 \left(\int_0^{s_{m-1}} dk_x^{\otimes(m-1)}\right)\otimes dk_{x(s_m)}, \int_0^1 \left(\int_0^{t_{m-1}} dk_y^{\otimes(m-1)}\right)\otimes dk_{y(t_m)}\rangle \\
&= 1 + \sum_{m\geq 1} \int_{\substack{s_m\in[0,1]\\ t_m\in[0,1]}} \langle \int_0^{s_m} dk_x^{\otimes(m)}, \int_0^{t_{m-1}} dk_y^{\otimes(m-1)}\rangle \langle dk_{x(s_m)}, dk_{y(t_m)})) \\
&= \dots \\
&= 1 + \sum_{m\geq 1} \int_{\substack{s_m\in[0,1]\\ t_m\in[0,1]}} \int_{\substack{s_{m-1}\in[0,s_m]\\ t_{m-1}\in[0,t_m]}} \cdots \int_{\substack{s_1\in[0,s_2]\\ t_1\in[0,t_2]}} \langle dk_{x(s_1)}, dk_{y(t_1)}\rangle \dots \langle dk_{x(s_m)}, dk_{y(t_m)}\rangle
\end{aligned}
$$

On considère dans un premier temps les chemins linéaires par morceaux x (resp. y) avec sa subdivision de l'intervalle $s_1 \leq \ldots \leq s_k$ (resp. $t_1 \leq \ldots \leq t_l$). On a alors :

$$
\begin{aligned}
&\int_{(s_1,s_k)\times(t_1,t_l)} \langle dk_{x(s)}, dk_{y(t)}\rangle_{\mathcal{H}} \\
&= \sum_{\substack{1\leq i\leq k-1\\ 1\leq j\leq l-1}} \int_{(s_1,s_k)\times(t_1,t_l)} \langle dk_{x(s)}, dk_{y(t)}\rangle_{\mathcal{H}} \\
&= \sum_{\substack{1\leq i\leq k-1\\ 1\leq j\leq l-1}} \langle k_{x(s_{i+1})} - k_{x(s_i)}, k_{y(t_{j+1})} - k_{y(t_j)}\rangle_{\mathcal{H}} \\
&= \sum_{\substack{1\leq i\leq k-1\\ 1\leq j\leq l-1}} k(x(s_{i+1}), y(t_{j+1})) - k(x(s_i), y(t_{j+1})) - k(x(s_{i+1}, y(t_j)) + k(x(s_i), y(t_j)) \\
&= \sum_{\substack{1\leq i\leq k-1\\ 1\leq j\leq l-1}} \kappa([s_1,s_k]\times[t_1,t_l]) = \int_{(s_1,s_k)\times(t_1,t_l)} d\kappa_{x,y}(s,t)
\end{aligned}
$$

On a ainsi montré la formule dans le cas de chemins linéaires par morceaux. Le résultat suit par approximation linéaire par morceaux dans le cas continu, en choisissant une subdivision dont le pas maximal tend vers 0 et en considérant κ_{x^n,y^n} la mesure associée à l'approximation. D'une part κ_{x^n,y^n} converge clairement faiblement vers $\kappa_{x,y}$ et d'autre part $\langle S(x^n), S(y^n)\rangle$ converge vers $\langle S(x), S(y)\rangle$. Ce qui permet de conclure dans le cas général.

De part la nature séquentielle de nos données et motivée par la preuve précédente, il est naturel de se demander à quel point l'approximation linéaire par morceau est fidèle au signal original, en termes de signatures. L'article fournit des bornes qui contrôlent cette erreur d'approximation.

Théorème 4. *Quelque soit h dans $C^1([0,1],\mathcal{H})$ et π une subdivision de $[0,1]$.*

$$\|S^+(h^\pi) - S(h)\| \leq \|h\|_1 e^{\|h\|_1} \cdot \max_{i=1,\dots,l-1} \|h_{[t_i,t_{i+1}]}\|_1$$

où $S^+(h^\pi)$ est la signature de l'approximation linéaire par morceaux aux points de la subdivision π, et $\|\cdot\|_1$ est la variation totale.

Finalement on abouti à un noyau discrétisé. On doit, pour espérer une quelconque implémentation, tronquer les signatures jusqu'à un ordre m, hyper-paramètre de notre modèle que l'on a l'habitude de traiter avec la signature en ML.
On définit alors le noyau à signature discrétisé k_m^+ tronqué à l'ordre m.

Définition 11.

$$k_m^+ : \mathcal{X}^{+^2} \to \mathbb{R}, k_m^+ = \langle S^+(k_x), S^+(k_y)\rangle_m$$

Où $\mathcal{X}^+ = \cup_{l\geq 0}\mathcal{X}^l$ est l'ensemble des séquences de points de $\mathcal{X}$. Une forme récursive est déduite des théorèmes précédents :

$$\begin{aligned} k_m^+(x,y) &= \sum_{d=1}^{m} \sum_{\substack{1\leq i_1<\ldots<i_d<|x| \\ 1\leq j_1<\ldots<j_d<|y|}} \prod_{r=1}^{d} \nabla_{i_r,j_r} k(x,y) \\ &= 1 + \sum_{\substack{|x|>i_1\geq 1 \\ |y|>j_1\geq 1}} \nabla_{i_1,j_1} k(x,y) \left(1 + \ldots \left(1 + \sum_{\substack{|x|>i_m\geq i_{m-1} \\ |y|>j_m\geq j_{m-1}}} \nabla_{i_m,j_m} k(x,y)\right)\right) \end{aligned}$$

avec

$$\nabla_{i,j} k(x,y) := k(x_{i+1}, yj+1) + k(x_i, y_j) - k(x_i, y_{j+1}) - k(x_{i+1}, y_j)$$

Si l'on implémente directement cette forme récursive en tirant parti des opérations vectorisées de bases fournies par *NumPy* par exemple, on se retrouve à calculer directement la matrix de Gram avec une complexité en temps en $O(n^2 l^2 m)$ et en mémoire en $O(l^2)$. L'astuce du noyau réside ici en l'absence de dépendance de la dimension de $\mathcal{H}$.

4.3 Méthodes tirant parti de l'approximation de matrice à rang faible

L'article propose deux niveaux d'approximations. L'une portant sur la matrice de Gram $(k_m^+(x_i, x_j))_{1\leq i,j\leq n}$ (sequence-vs-sequence), et l'autre sur les matrices de similarités entre les séquences $(\nabla_{a,b} k(x,y))_{\substack{1\leq a\leq |x|, \\ 1\leq b\leq |y|}}$ à l'aide d'approximations de faible rang, afin de permettre un passage à l'échelle pour les grands jeux de données et/ou des séries plus longues.

4.3.1 Approximation de la matrice de Gram par la méthode de Nyström

Dans le but de réduire la complexité du calcul de la matrice de Gram, on introduit une méthode classique d'approximation de la matrice de Gram.[?]

5 Le noyau à signature est solution d'une EDP hyperbolique.

On présente ici des travaux récents qui vont dans le prolongement de l'article précédent. Dans *The Signature Kernel is the solution of a Goursat PDE* [?], il est montré que le noyau non tronqué est solution d'une EDP hyperbolique, l'EDP de Goursat. Un schéma numérique de résolution

est proposé et sa complexité est comparable à ce qui a été exposé avant, profitant de la possible parallélisation dans le GPU.

On énonce le résultat central de cet article.

Théorème 5. *Soit $I = [u, u^{'}]$ et $J = [v, v^{'}]$ deux intervalles compacts et soit $x \in C^1(I, V)$ où V est un espace de Banach. Le noyau signature $k^{\oplus}$ est solution d'une EDP. Plus exactement on pose :*

$$\forall (s,t) \in I \times J, \quad k_{x,y}(s,t) = \langle S(x)_{[u,s]}, S(y)_{[v,t]} \rangle$$

On regarde donc désormais le produit scalaire de deux signatures sur des intervalles que l'on paramètre par la droite. Cette fonction est solution de l'EDP suivante.

$$\frac{\partial^2 k_{x,y}}{\partial s \partial t} = \langle \dot{x_s}, \dot{y_t} \rangle_V k_{x,y}, \quad k_{x,y}(u, \cdot) = k_{x,y}(\cdot, v) = 1$$

Preuve : Les conditions initiales sont clairement vérifiées