Random Signature Fourier Features

Csaba Toth
Work in progress with Harald Oberhauser and Zoltan Szabo
Mathematical Institute, University of Oxford

New interfaces of Stochastic Analysis and Rough Paths,
Banff International Research Station, September 4 - 9, 2022

Introduction

Motivation

- Supervised learning is concerned with finding representations of the data $\mathbf{x} \mapsto \Phi(\mathbf{x})$ over which linear functions $\langle \ell, \Phi(\mathbf{x}) \rangle \approx \mathbf{y}$ can approximate input-output maps $\mathbf{x} \mapsto \mathbf{y}$
- Signatures are interesting both theoretically and empirically for modelling functions of sequences
- Scalability, on the other hand, is a well-known bottleneck for structured data types, such as sequences, trees, graphs, etc. . .
- The signature is no exception, where the bottleneck is either in

$$\underbrace{\Phi_m(\mathbf{x}) \in \left(\mathbb{R}^d\right)^{\otimes m}}_{O(d^m) \text{ coordinates for level-m}} \quad \text{or} \quad \underbrace{\left(\langle \mathbf{x}_k, \mathbf{y}_l \rangle\right)_{k,l=1}^{L} \in \mathbb{R}^{L \times L}}_{O(L^2) \text{ pairwise comparisons}}$$

- ... this work: try to get the best of both worlds!

Recap on signatures

Recap on signatures

Sequential data

Assume our data observations lie in some Banach space $\mathcal{B}$. Computationally, we deal with sequential data in discrete chunks,

$$\mathrm{Seq}_V(\mathcal{B}) := \{\mathbf{x}_{1:L_\mathbf{x}} = (\mathbf{x}_1, \dots, \mathbf{x}_{L_\mathbf{x}}) \mid \mathbf{x}_{t_i} \in V, L_\mathbf{x} \in \mathbb{N}, \|\mathbf{x}_{1:L_\mathbf{x}}\|_{\text{1-var}} \leq V\}.$$

But theoretically, we may prefer to think about it in continuous time,

$$\mathrm{Paths}_V(\mathcal{B}) = \left\{(\mathbf{x}_t)_{t\in[0,T_\mathbf{x}]} \in C([0, T_\mathbf{x}], V) \mid \mathbf{x}_0 = 0,\, T_\mathbf{x} \in \mathbb{R}_+,\, \|\mathbf{x}\|_{\text{1-var}} \leq V\right\},$$

where the bounded variation norm of a sequence $\mathbf{x}_{1:L_\mathbf{x}} \in \mathrm{Seq}_V(\mathcal{B})$ is

$$\|\mathbf{x}_{1:L_\mathbf{x}}\|_{\text{1-var}} := \sum_{l=1}^{L_\mathbf{x}-1} \|\mathbf{x}_{l+1} - \mathbf{x}_l\|_\mathcal{B} \leq V$$

and of a path $(\mathbf{x}_t)_{t\in[0,T_\mathbf{x}]} \in \mathrm{Paths}_V(\mathcal{B})$ is

$$\|\mathbf{x}\|_{\text{1-var}} = \sup_{0=t_1<\cdots<t_n=T_\mathbf{x}} \sum_{i=1}^{n-1} \|\mathbf{x}_{t_{i+1}} - \mathbf{x}_{t_i}\|_\mathcal{B} \leq V$$

Recap on signatures

Signature features and tensor algebras

For a given integer $m \in \mathbb{N}$, the collection of level-m iterated integrals of a path $\mathbf{x} \in \text{Paths}(V)$ is defined as the tensor of order-m

$$S_m(\mathbf{x}) = \int_{0<t_1<\cdots<t_m<T_{\mathbf{x}}} d\mathbf{x}_{t_1} \otimes \cdots \otimes d\mathbf{x}_{t_m} \in \mathcal{B}^{\otimes m}.$$

The collection of iterated integrals for all levels $m \in \mathbb{N}$ is the signature of the path living in a feature space known as the tensor algebra over $\mathcal{B}$,

$$S(\mathbf{x}) = (1, S_1(\mathbf{x}), S_2(\mathbf{x}), \ldots, S_m(\mathbf{x}), \ldots) \in \prod_{m=1}^{\infty} \mathcal{B}^{\otimes m}.$$

The truncated signature of level-M is given by ignoring tensors of order larger than M, and it lives in the truncated tensor algebra,

$$S_{\leq M}(\mathbf{x}) = (1, S_1(\mathbf{x}), S_2(\mathbf{x}), \ldots, S_m(\mathbf{x})) \in \prod_{m=1}^{M} \mathcal{B}^{\otimes m}.$$

.

Recap on signatures

Signature kernels and the kernel trick

Kiraly and Oberhauser (2019) have introduced a kernel trick for the inner product of level-m iterated integrals of paths $\mathbf{x}, \mathbf{y} \in \text{Paths}(\mathcal{H})$,

$$\langle S_m(\mathbf{x}), S_m(\mathbf{y}) \rangle = \int \int_{\substack{0<s_1<\cdots<s_m<T_{\mathbf{x}} \\ 0<t_1<\cdots<t_m<T_{\mathbf{y}}}} \langle d\mathbf{x}_{s_1}, d\mathbf{y}_{t_1} \rangle \cdots \langle d\mathbf{x}_{s_m}, d\mathbf{y}_{t_m} \rangle.$$

This allowed them to compute truncated signature kernels by simply inner product evaluation on $\mathcal{H}$, leading to 1) circumventing the curse of dimensionality incurred by high order tensors, 2) compute signature kernels over by paths evolving in an infinite-dimensional RKHS,

$$\mathrm{k}_{S_M}(\mathbf{x}, \mathbf{y}) = \sum_{m=0}^{M} \langle S_m(\mathbf{x}), S_m(\mathbf{y}) \rangle$$

Recent work by Salvi, Cass, Foster, Lyons and Yang (2021) devised a PDE-approach to approximating the full signature kernel,

$$\mathrm{k}_{S_{PDE}}(\mathbf{x}, \mathbf{y}) \approx \sum_{m=0}^{\infty} \langle S_m(\mathbf{x}), S_m(\mathbf{y}) \rangle$$

Discretized signature features (1/2)

- To go from continuous-time to discrete-time, the canonical choice is to lift signatures to paths by e.g. linear interpolation
- Other possible choices introduced in e.g. Kiraly and Oberhauser (2019) also introduced order-p discretizations of signatures:

$$\Phi_m^{(p)}(\mathbf{x}) = \sum_{1 \leq i_1 \leq \cdots \leq i_m < L_{\mathbf{x}}} \frac{1}{\#^{(p)}(\mathbf{i})!} \nabla \mathbf{x}_{i_1} \otimes \cdots \otimes \nabla \mathbf{x}_{i_m},$$

  where $\nabla \mathbf{x}_{i_k} = \mathbf{x}_{i_k+1} - \mathbf{x}_{i_k}$ is sequence differencing, and $\#$ is a bin count operator such that $\#(i_1, \ldots, i_m) = (c_1, \ldots, c_d)$ and c_j counts the number of occurrences of the jth unique element in $\mathbf{i}$. Then,

$$\#^{(p)}(i_1, \ldots, i_m) = \begin{cases} c_1! \cdots c_d! & \text{if } c_1, \ldots, c_d \leq p \\ \infty & \text{else} \end{cases}$$

- For $p = m$, it is the level-m signature of a piecewise linear path

Discretized signature features (2/2)

- Although the order-p discretized signature does not correspond to the signature of a path for $p < m$, a there is a concatenation property and for e.g.$p = 1$ a quasi-half shuffle product, which allows to get universality when precomposed with a universal "static" feature map.
- For example, the order-p kernel $\mathrm{k}^{(p)}(\mathbf{x}, \mathbf{y}) = \sum_{m \geq 0} \mathrm{k}_m^{(p)}(\mathbf{x}, \mathbf{y})$ where

$$\mathrm{k}_m^{(p)}(\mathbf{x}_{1:K}, \mathbf{y}_{1:L}) = \sum_{\substack{1 \leq i_1 \leq \cdots \leq i_m < K \\ 1 \leq j_1 \leq \cdots \leq j_m < L}} \frac{1}{\#^{(p)}(\mathbf{i})! \#^{(p)}(\mathbf{j})!} \nabla\kappa(\mathbf{x}_{i_1}, \mathbf{y}_{j_1}) \cdots \nabla\kappa(\mathbf{x}_{i_m}, \mathbf{y}_{j_m})$$

  is universal for all $p \in \mathbb{N}$ if $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a universal kernel
- Following, we focus on the case $p = 1$ for simplicity of calculations:

$$\mathrm{k}_m^{(1)}(\mathbf{x}_{1:K}, \mathbf{y}_{1:L}) = \sum_{\substack{1 \leq i_1 < \cdots < i_m < K \\ 1 \leq j_1 \leq \cdots \leq j_m < L}} \nabla\kappa(\mathbf{x}_{i_1}, \mathbf{y}_{j_1}) \cdots \nabla\kappa(\mathbf{x}_{i_m}, \mathbf{y}_{j_m})$$

Random Fourier Features

Main ideas

- We are given a static kernel $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that $\mathcal{H}_\kappa = \overline{\mathrm{span}}\{\kappa(\mathbf{x}, \cdot) \,|\, \mathbf{x} \in \mathbb{R}^d\}$ is infinite-dimensional
- It is well-known that there exist many $\phi : \mathbb{R}^d \to \mathcal{H}$ into some Hilbert space $\mathcal{H}$ such that

$$\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle = \kappa(\mathbf{x}, \mathbf{y}) \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$$

- E.g. we may take for $\phi(\mathbf{x}) = \kappa(\mathbf{x}, \cdot)$, or $\phi(\mathbf{x}) = (\sqrt{\lambda_i}\mathbf{e}_i(\mathbf{x}))_{i \geq 0}$, where $\mathbf{e}_i$ are the eigenfunctions from Mercer's theorem
- We want a finite-dimensional (!) $\tilde{\phi} : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ such that

$$\left\langle \tilde{\phi}(\mathbf{x}), \tilde{\phi}(\mathbf{y}) \right\rangle \approx \kappa(\mathbf{x}, \mathbf{y}) \text{ for } \mathbf{x}, \mathbf{y} \in \mathcal{M} \subset \mathbb{R}^d,$$

  such that we approximately capture the inner product, i.e. $\kappa(\mathbf{x}, \mathbf{y})$, but we don't care about capturing any feature map $\phi(\mathbf{x})$ itself!

Random Fourier Features

Spectral representation of stationary kernels

- Question: Given an infinite-dimensional kernel κ over $\mathbb{R}^d$, how do we find an approximate feature map $\tilde{\phi} : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$

$$\kappa(\mathbf{x}, \mathbf{y}) = \left\langle \tilde{\phi}(\mathbf{x}), \tilde{\phi}(\mathbf{x}) \right\rangle \text{ for } \mathbf{x}, \mathbf{y} \in \mathcal{M} \subset \mathbb{R}^d$$

- Rahimi and Recht (2007) proposed an approach for approximating stationary kernels, i.e. $\kappa(\mathbf{x}, \mathbf{y}) = \kappa(\mathbf{x} + \mathbf{z}, \mathbf{y} + \mathbf{z})$ for all $\mathbf{x}, \mathbf{y}, \mathbf{z}$ in $\mathbb{R}^d$
- Bochner's theorem: if κ is a continuous, bounded stationary kernel, it can be written as the Fourier transform of a non-negative finite measure,

$$\kappa(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} \exp(i\mathbf{w}^\top(\mathbf{x} - \mathbf{y}))d\Lambda,$$

  where Λ is the spectral measure of κ (wlog assume $\Lambda(\mathbb{R}^d) = 1$)
- This represents the kernel as an inner product in an L_2 space of RVs

$$\kappa(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\mathbf{w} \sim \Lambda} \left[e^{i\mathbf{w}^\top \mathbf{x}} e^{i\mathbf{w}^\top \mathbf{y}^\star} \right]$$

Random Fourier Features

- Random Fourier features are defined for $W = [\mathbf{w}_i]_{i=1}^{\tilde{d}}$, $\mathbf{w}_i \sim \Lambda$

$$\tilde{\phi}_W(\mathbf{x}) := \frac{1}{\sqrt{\tilde{d}}} \left(\sin(W^\top \mathbf{x}), \cos(W^\top \mathbf{x}) \right) \in \mathbb{R}^{2\tilde{d}}$$

- First and foremost, RFFs approximate the true kernel κ since

$$\tilde{\kappa}(\mathbf{x}, \mathbf{y}) = \left\langle \tilde{\phi}_W(\mathbf{x}), \tilde{\phi}_W(\mathbf{x}) \right\rangle = \frac{1}{\tilde{d}} \sum_{i=1}^{\tilde{d}} \cos(\mathbf{w}_i^\top (\mathbf{x} - \mathbf{y})),$$

and hence

$$\mathbb{E}\left[\left\langle \tilde{\phi}_W(\mathbf{x}), \tilde{\phi}_W(\mathbf{y}) \right\rangle \right] = \frac{1}{\tilde{d}} \sum_{i=1}^{\tilde{d}} \mathbb{E}\left[\cos(\mathbf{w}_i^\top (\mathbf{x} - \mathbf{y}) \right] = \kappa(\mathbf{x}, \mathbf{y})$$

- Analogously by DCT, we also have for the derivatives of RFFs

$$\partial^{\mathbf{p},\mathbf{q}} \tilde{\kappa}(\mathbf{x}, \mathbf{y}) = \left\langle \partial^{\mathbf{p}} \tilde{\phi}_(\mathbf{x}), \partial^{\mathbf{q}} \tilde{\phi}(\mathbf{y}) \right\rangle \approx \partial^{\mathbf{p},\mathbf{q}} \kappa(\mathbf{x}, \mathbf{y})$$

Theoretical guarantees for RFFs

- Rahimi and Recht (2007) showed the following uniform bound

$$\mathbb{P}\left[\sup_{\mathbf{x},\mathbf{y}\in\mathcal{M}} |\tilde{\kappa}(\mathbf{x},\mathbf{y}) - \kappa(\mathbf{x},\mathbf{y})|\right] \leq C\left(\frac{\sigma_\Lambda^2 \operatorname{diam}(\mathcal{M})}{\epsilon}\right)^2 \exp\left(\frac{-\tilde{d}\epsilon^2}{d(\epsilon+2)}\right)$$

  over a convex, compact domain $\mathcal{M} \subset \mathbb{R}^d$, where $\sigma^2 = \mathbb{E}_{\mathbf{w}\sim\Lambda} \|\mathbf{w}\|^2$ is the trace of the second moment of Λ.
- In particular, this bound shows that the error converges uniformly to 0 at a rate $O_p\left(|\mathcal{M}| \sqrt{\tilde{d}^{-1} \log \tilde{d}}\right)$
- This bound was later tightened by Sriperumbudur and Szabo (2015), where a rate of $O_{a.s.}\left(\sqrt{\log |\mathcal{M}|\, \tilde{d}^{-1}}\right)$, and analogous results extended to the derivatives of RFFs in Szabo and Sriperumbudur (2019), and generalized to Orlicz Spaces in Chamakh, Gobet and Szabo (2020)

Random Signature Fourier features

Approximating signature kernels

- We apply the same idea, namely that we do not approximate the signature of a path
- Instead, we focus on approximately capturing the kernel itself, that is, we find features $\tilde{\Phi}_m : \mathrm{Seq}(\mathbb{R}^d) \to \tilde{\mathcal{H}}$ such that

$$\left\langle \tilde{\Phi}_m(\mathbf{x}), \tilde{\Phi}_m(\mathbf{x}) \right\rangle \approx \langle \Phi_m(\mathbf{x}), \Phi_m(\mathbf{y}) \rangle \quad \text{for} \;\; \mathbf{x}, \mathbf{y} \in \mathrm{Seq}(\mathbb{R}^d),$$

  where $\Phi_m(\mathbf{x}) \in \mathcal{H}^{\otimes m}$ and $\tilde{\Phi}_m(\mathbf{x}) \in \tilde{\mathcal{H}}$, hence the features themselves need not even live in the same space
- Naive approach: take signature features over the sequence lifted using a given copy of the random Fourier feature mapping, i.e.

$$\tilde{\Phi}_m(\mathbf{x}) = \sum_{1 \le i_1 < \cdots < i_m < L_\mathbf{x}} \nabla \tilde{\phi}(\mathbf{x}_{i_1}) \otimes \cdots \otimes \nabla \tilde{\phi}(\mathbf{x}_{i_m})$$

Random Signature Fourier features

Tensorizing Fourier features

- Naively, one might think to extend RFFs to signature kernels by replacing κ with $\tilde{\kappa}$. However, this would not work in general since

$$\mathbb{E}\left[\tilde{\kappa}(\mathbf{x}_1, \mathbf{y}_1) \cdots \tilde{\kappa}(\mathbf{x}_m, \mathbf{y}_m)\right] \neq \kappa(\mathbf{x}_1, \mathbf{y}_1) \cdots \kappa(\mathbf{x}_m, \mathbf{y}_m)$$

- Workaround: sample $W_i = [\mathbf{w}_{i,j}]_{j=1}^{\tilde{d}}$, $\mathbf{w}_{i,j} \sim \Lambda$, and denoting $\tilde{\kappa}_i(\mathbf{x}, \mathbf{y}) \equiv \tilde{\kappa}_{W_i}(\mathbf{x}, \mathbf{y})$, use independent features across the product

$$\mathbb{E}\left[\tilde{\kappa}_1(\mathbf{x}_1, \mathbf{y}_1) \cdots \tilde{\kappa}_m(\mathbf{x}_m, \mathbf{y}_m)\right] = \kappa(\mathbf{x}_1, \mathbf{y}_1) \cdots \kappa(\mathbf{x}_m, \mathbf{y}_m)$$

- Specifically, for the level-m order-1 signature kernel

$$\mathrm{k}_m(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{i},\mathbf{j}} \nabla\kappa(\mathbf{x}_{i_1}, \mathbf{y}_{j_1}) \cdots \nabla\kappa(\mathbf{x}_{i_m}, \mathbf{y}_{j_m})$$

- We can define the random, unbiased approximation

$$\tilde{\mathrm{k}}_m(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{i},\mathbf{j}} \nabla\tilde{\kappa}_1(\mathbf{x}_{i_1}, \mathbf{y}_{j_1}) \cdots \nabla\tilde{\kappa}_m(\mathbf{x}_{i_m}, \mathbf{y}_{j_m})$$

Theoretical guarantees of RSFF (1/2)

- With the previous definition, we have for any $\mathbf{x}, \mathbf{y} \in \mathrm{Seq}_V(\mathbb{R}^d)$

$$\mathbb{E}\left[\tilde{\mathrm{k}}_m(\mathbf{x}, \mathbf{y})\right] = \mathrm{k}_m(\mathbf{x}, \mathbf{y})$$

- Under the assumption that κ is C-Lipschitz for some $C > 0$ such that $\|\kappa_{\mathbf{x}} - \kappa_{\mathbf{y}}\|_{\mathcal{K}} \leq C\,\|\mathbf{x} - \mathbf{y}\|$, we have the uniform control over BV sequences

$$\sup_{\mathbf{x},\mathbf{y} \in \mathrm{Seq}_V(\mathcal{M})} \left|\tilde{\mathrm{k}}_m(\mathbf{x}, \mathbf{y}) - \mathrm{k}_m(\mathbf{x}, \mathbf{y})\right|$$

$$\leq V^{2m} \sum_{k=1}^{m} \frac{C^{2(m-k)}\,\|W_1\|_2^2 \cdots \|W_{k-1}\|_2^2}{\tilde{d}^{k-1}((k-1)!)^2} \sup_{\mathbf{x},\mathbf{y} \in \mathcal{M}} \|\partial_{1,2}\tilde{\kappa}(\mathbf{x}, , \mathbf{y}) - \partial_{1,2}\kappa(\mathbf{x}, \mathbf{y})\|_2\,,$$

where $\partial_{1,2}\kappa(\mathbf{x}, \mathbf{y}) := [\partial^2\kappa(\mathbf{x}, \mathbf{y})/\partial x_i \partial y_j]_{i,j=1}^d$ and $V > 0$ is the maximal 1-var.

Theoretical guarantees of RSFF (2/2)

- Assuming the following additional conditions
  1. κ is three-times differentiable ($\mathcal{C}^3$)
  2. $\mathbb{E}_{\mathbf{w}\sim\Lambda}[w_i w_j \, \|\mathbf{w}\|_2] < \infty$ for all $i, j \in [d]$.
  3. $\mathbb{E}_{\mathbf{w}\sim\Lambda}[|w_i|^k \, |w_j|^k] \leq \frac{k!\sigma^2 R^{m-2}}{2}$ for $i, j, \in [d]$,
- it holds that

$$\mathbb{P}\left[\sup_{\mathbf{x},\mathbf{y}\in \mathsf{Seq}_V(\mathcal{M})} \left|\mathrm{k}_m(\mathbf{x},\mathbf{y}) - \tilde{\mathrm{k}}_m(\mathbf{x},\mathbf{y})\right| \geq \epsilon\right]$$

$$\leq M \cdot \begin{cases} \left(16\,|\mathcal{M}|\,(\bar{D}_{\mathcal{M}} + \bar{E})\left(\frac{\beta_{d,V,M}}{\epsilon}\right) + d\right)\exp\left(-\frac{\tilde{d}}{2(d+1)}\frac{\left(\frac{\epsilon}{\beta_{d,V,M}}\right)^2}{\sigma^2 + R\left(\frac{\epsilon}{\beta_{d,V,M}}\right)}\right) \\ \left(16\,|\mathcal{M}|\,(\bar{D}_{\mathcal{M}} + \bar{E})\left(\frac{\beta_{d,V,M}}{\epsilon}\right)^{1/M} + d\right)\exp\left(-\frac{\tilde{d}}{2(d+1)}\frac{\left(\frac{\epsilon}{\beta_{d,V,M}}\right)^{2/M}}{\sigma^2 + R\left(\frac{\epsilon}{\beta_{d,V,M}}\right)^{1/M}}\right) \end{cases}$$

where $\beta_{d,V,M} := \left(2V^2 \max(C,1)^2 \max(\sigma_\Lambda^2, d)\right)^M$.

Experiments

Time Series Classification

- 2 tested variations of the previous idea: 1) RSFF_{H}, 2) RSFF_{TRP}
- On some simple TSC experiments, we compared against other variants of sequence/signature kernels with the random signatures
- We observe that the performance is close across different variations of signatures, but the RSFFs perform competitively

Datasets/Kernels	GAK	Sig(n)	Sig-PDE	K_{Sig}	RSFF_{H}	RSFF_{TRP}
ArticularyWordRecognition	98	92.3	98.3	**99.0**	98.2	98.5
BasicMotions	97.5	97.5	**100**	**100**	**100**	**100**
Cricket	97.2	86.1	97.2	**98.6**	94.6	95.8
ERing	**93.7**	84.1	93.3	84.1	90.7	90.1
Libras	79	81.7	81.7	91.6	**92.2**	91.2
NATOPS	90.6	88.3	93.3	**93.9**	93.7	92.3
RacketSports	84.2	80.2	84.9	**86.2**	80.0	79.9
FingerMovements	**61**	51	58	**63**	59.5	59.4
Heartbeat	70.2	72.2	73.6	71.2	**74.8**	74.4
SelfRegulationSCP1	**92.4**	75.4	88.7	89.8	89.4	89.6
UWaveGestureLibrary	**87.5**	83.4	87	**87.5**	84.3	84.3

Next steps

Next steps

- Empirical evaluation on large-scale datasets
- Tighten the bounds using the optimal RFF rates from Szabo and Sriperumbudur (2019), and Chamakh, Gobet and Szabo (2020)
- Generalize the technique of Rudi and Rosasco (2017) for downstream learning performance estimates
- Extend the bounds to "order-p" signature features

Thanks for your attention!