Maud Lemercier, Cristopher Salvi, Theodoros Damoulas, Edwin V. Bonilla, Terry Lyons

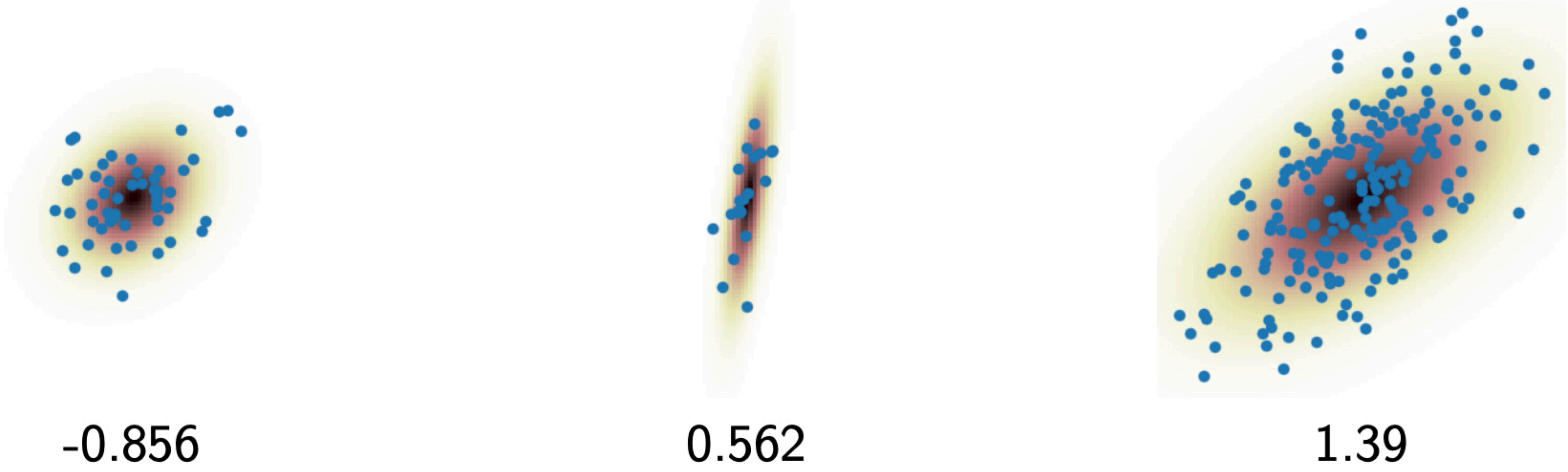
Distribution Regression for Sequential Data

DataSig

A rough path between mathematics and data science

University of Oxford
The Alan Turing Institute
Imperial College London
UCL
UKRI Engineering and Physical Sciences Research Council

Problem definition

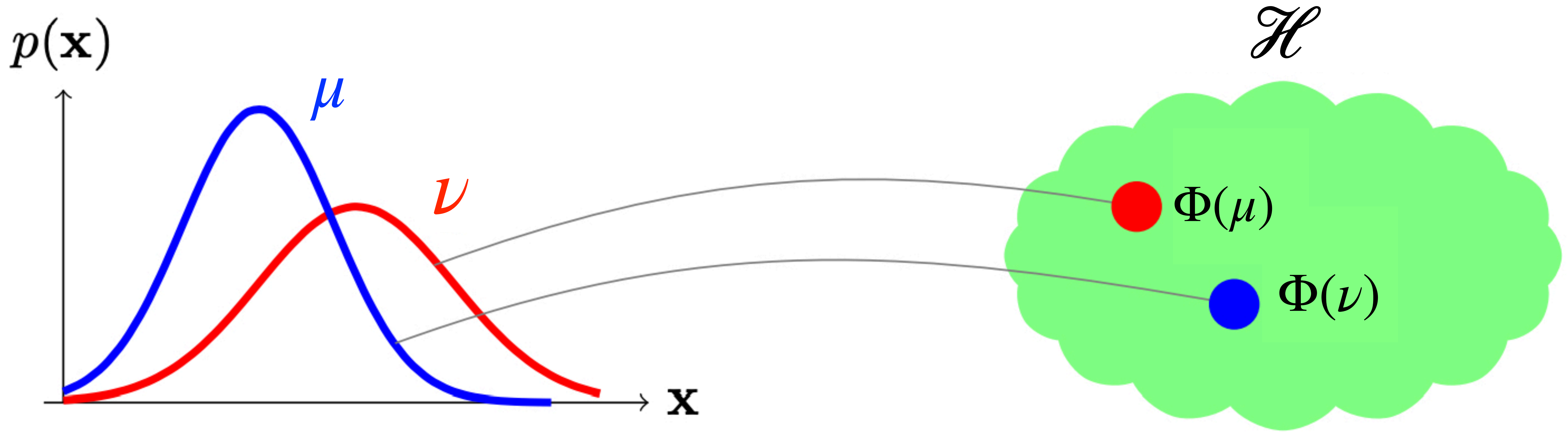


Input-output pairs $\left(\{x^{1,j}\}_{j=1}^{N_1}, y^1\right), \cdots, \left(\{x^{M,j}\}_{j=1}^{N_M}, y^M\right)$

- **input:** $\{x^{i,j}\}_{j=1}^{N_i}$ a collection of N_i points $x^{i,j}$
- **output:** $y^i \in \mathbb{R}$
- **goal:** predict $y*$ for a new $\{x^{*,j}\}_{j=1}^{N_*}$

Expected Feature Map

- Let $\mathscr{P}(\mathscr{X})$ be the set of probability measures on $\mathscr{X}$
- Let $\varphi : \mathscr{X} \to \mathscr{H}$ be a feature map from $\mathscr{X}$ to the Hilbert space $\mathscr{H}$ associated to the kernel $k^{\varphi} : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$
- Expected feature map $\Phi : \mu \mapsto \mathbb{E}_{X \sim \mu}[\varphi(X)] \in \mathscr{H}$



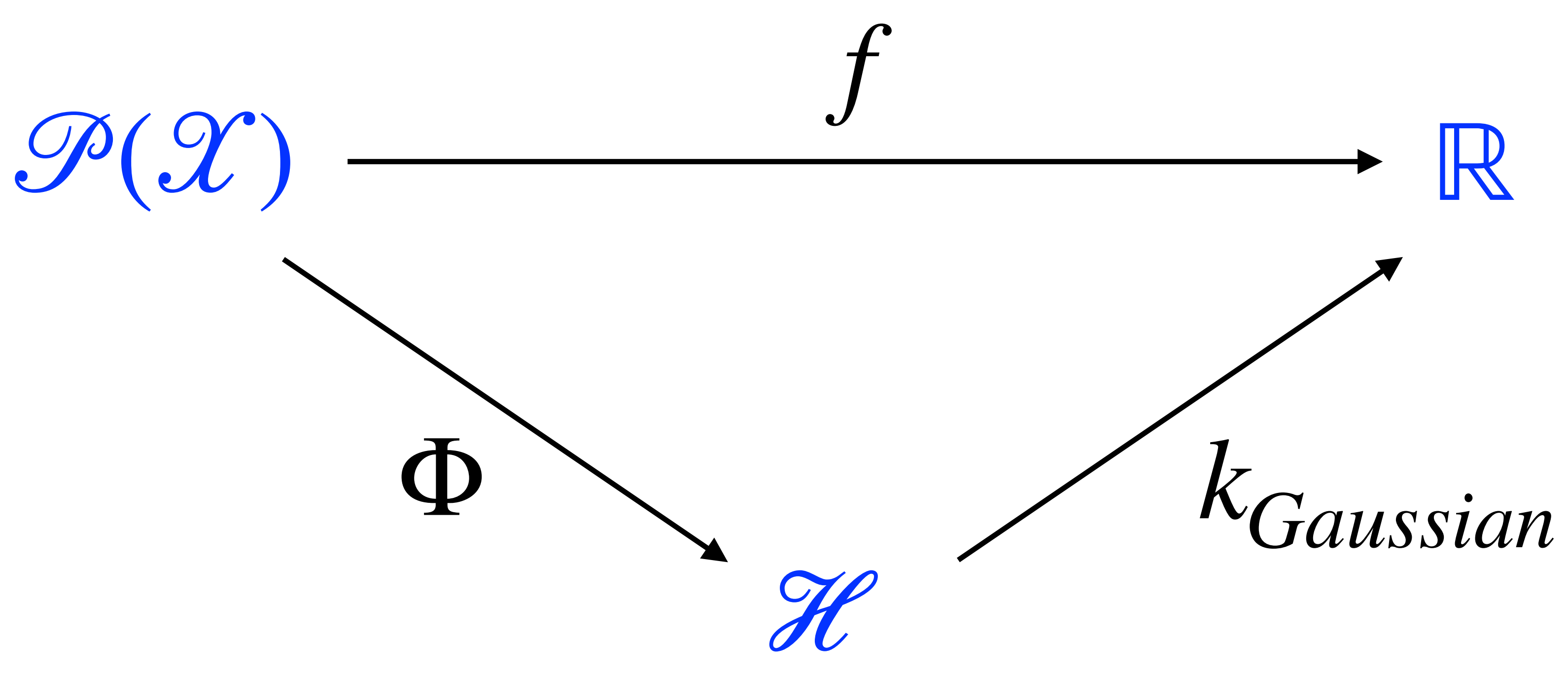
Existing result

Theorem (Christmann and Steinwart 2010)

Let $\mathscr{X}$ is a compact set. If $\Phi : \mathscr{P}(\mathscr{X}) \to \mathscr{H}$ is an injective and continuous map and $\mathscr{H}$ is a separable Hilbert space, then the kernel $k : \mathscr{P}(\mathscr{X}) \times \mathscr{P}(\mathscr{X}) \to \mathbb{R}$ defined by

$$k(\mu, \nu) = \exp\Big(-\sigma^2 \|\Phi(\mu) - \Phi(\nu)\|_{\mathscr{H}}^2 \Big), \qquad \sigma > 0,$$

is universal, i.e. the associated RKHS is dense in $\mathscr{C}(\mathscr{P}(\mathscr{X}), \mathbb{R})$.



Distribution Regression for Sequential Data

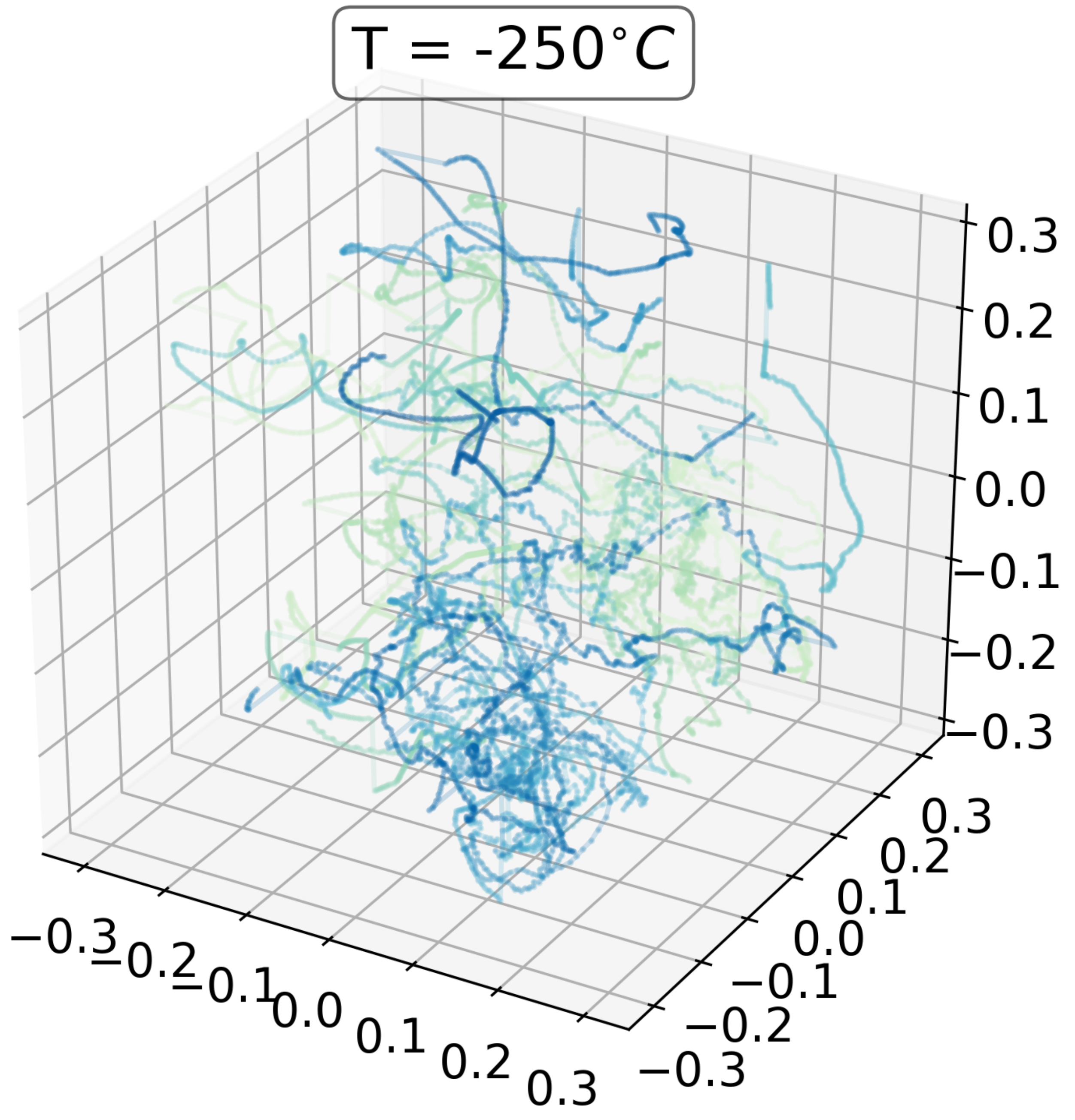


Figure 1. Simulated gases (30 particles) under different thermodynamic conditions. Higher temperatures increase the number of collisions resulting in different large-scale dynamics.

The Expected Signature

Definition

The expected signature $\mathbb{E}S : \mathscr{P}(\mathscr{X}) \to \mathscr{T}(\mathbb{R}^d)$ is the map $\mathbb{E}S : \mu \mapsto \mathbb{E}_{X\sim\mu}[S(X)]$ defined element-wise for any $k \geq 0$ and any $(i_1, \ldots, i_k) \in \{1, \ldots, d\}^k$ as

$$\mathbb{E}_{X\sim\mu}[S(X)]^{(i_1,\ldots,i_k)} = \int_{x\in\mathscr{X}} S(x)^{(i_1,\ldots,i_k)}\mu(dx) \in \mathbb{R}\,.$$

Definition

The kernel $k^{sig} : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ associated to the feature map S is defined as

$$k^{sig}(x, y) = \langle S(x), S(y)\rangle_{\mathscr{T}(\mathbb{R})^d}$$

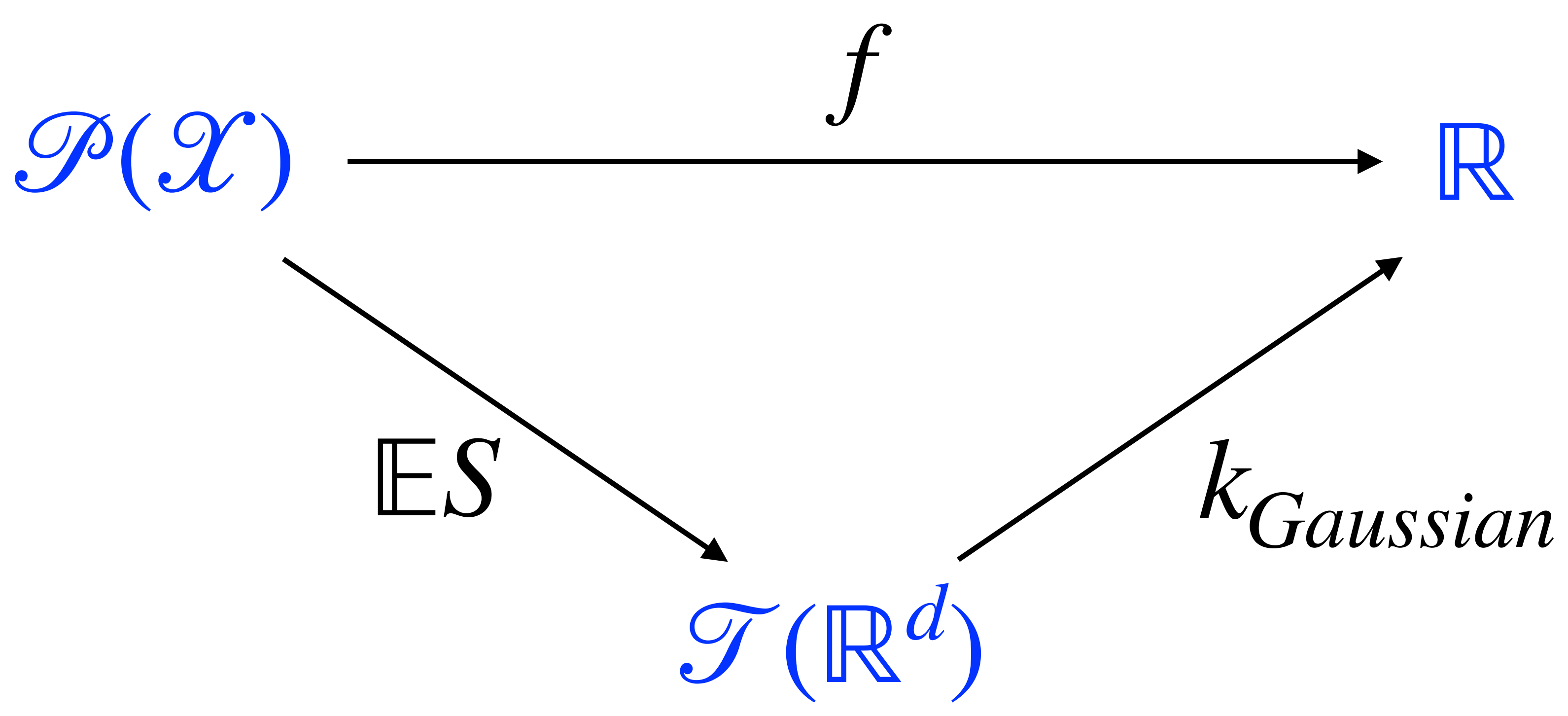
A kernel-based approach to DR

Theorem

If $\mathscr{X} \subset \mathscr{C}_{Lip}(I, \mathbb{R}^d)$ is a compact subset of paths, then the kernel $k : \mathscr{P}(\mathscr{X}) \times \mathscr{P}(\mathscr{X}) \to \mathbb{R}$ defined by

$$k(\mu, \nu) = \exp\Big(- \sigma^2 \| \mathbb{E}S(\mu) - \mathbb{E}S(\nu) \|^2_{\mathscr{T}(\mathbb{R}^d)} \Big), \qquad \sigma > 0,$$

is universal, i.e. the associated RKHS is dense in $\mathscr{C}(\mathscr{P}(\mathscr{X}), \mathbb{R})$.



A kernel-based approach to DR

Theorem

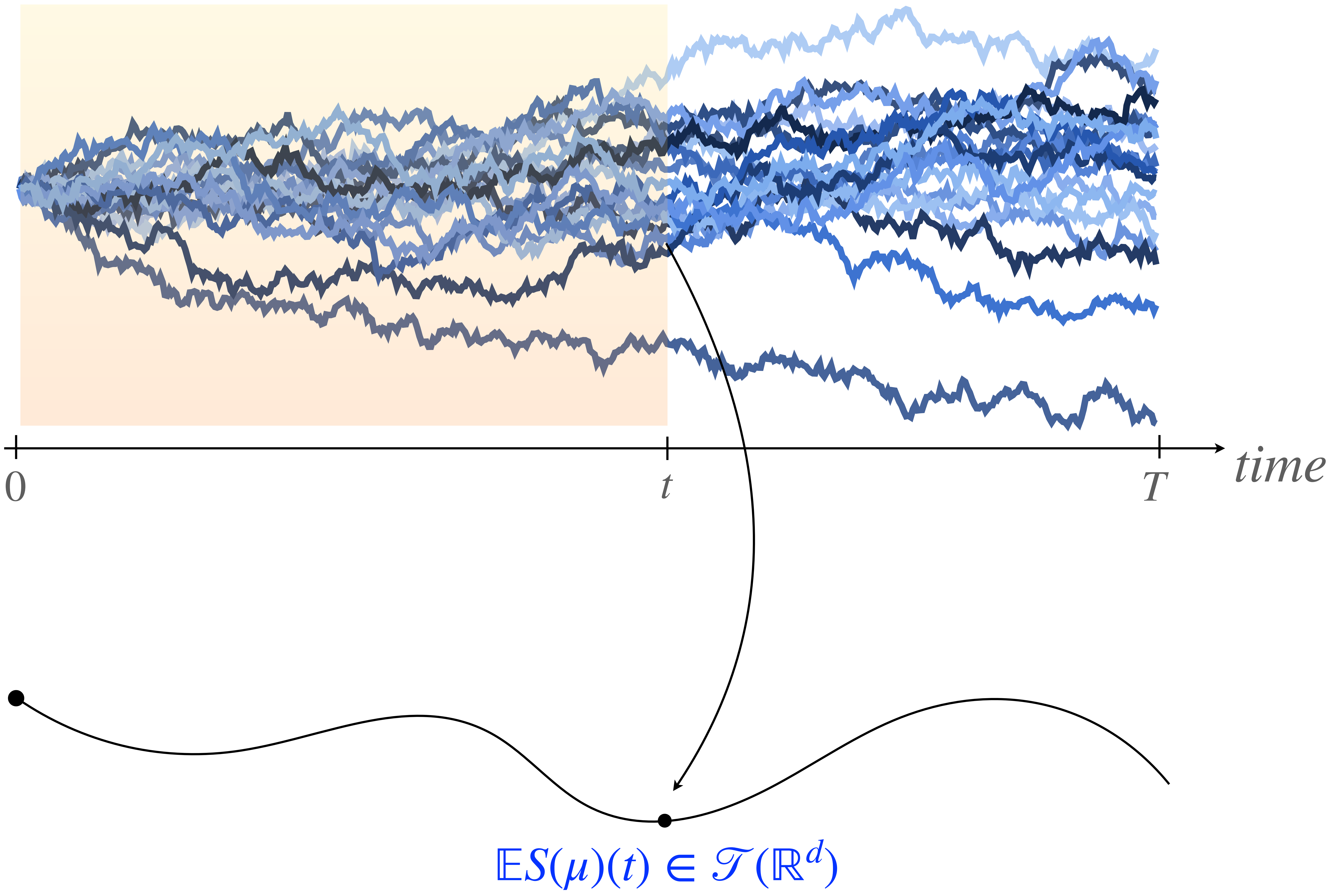
If $\mathscr{X} \subset \mathscr{C}_{Lip}(I, \mathbb{R}^d)$ is a compact subset of paths, then the kernel $k : \mathscr{P}(\mathscr{X}) \times \mathscr{P}(\mathscr{X}) \to \mathbb{R}$ defined by

$$k(\mu, \nu) = \exp\Big(- \sigma^2 \|\mathbb{E}S(\mu) - \mathbb{E}S(\nu)\|^2_{\mathscr{T}(\mathbb{R}^d)}\Big), \qquad \sigma > 0,$$

is universal, i.e. the associated RKHS is dense in $\mathscr{C}(\mathscr{P}(\mathscr{X}), \mathbb{R})$.

$$||\mathbb{E}S(\mu) - \mathbb{E}S(\nu)||^2_{\mathscr{T}(\mathbb{R}^d)} = \mathbb{E}_{X,X' \sim \mu}[k^{sig}(X, X')] - 2\mathbb{E}_{X \sim \mu, Y \sim \nu}[k^{sig}(X, Y)] + \mathbb{E}_{Y,Y' \sim \nu}[k^{sig}(Y, Y')]$$

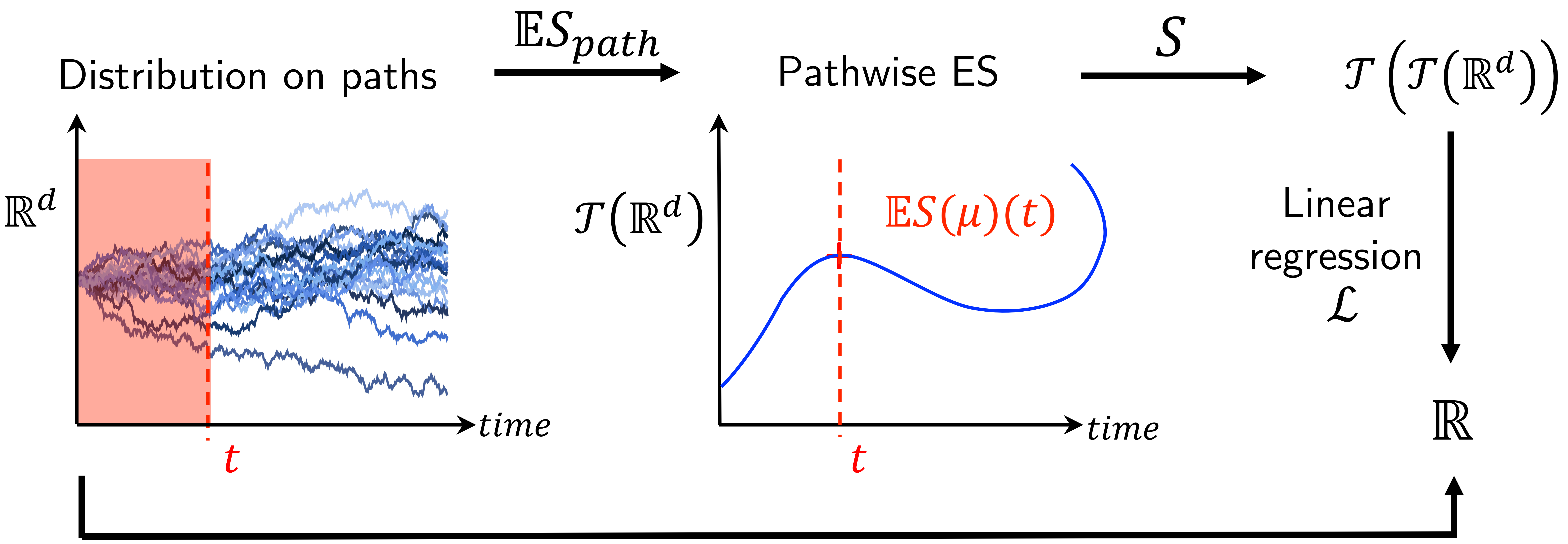
A pathwise approach to DR



A pathwise approach to DR

The Stone-Weierstrass Theorem for paths

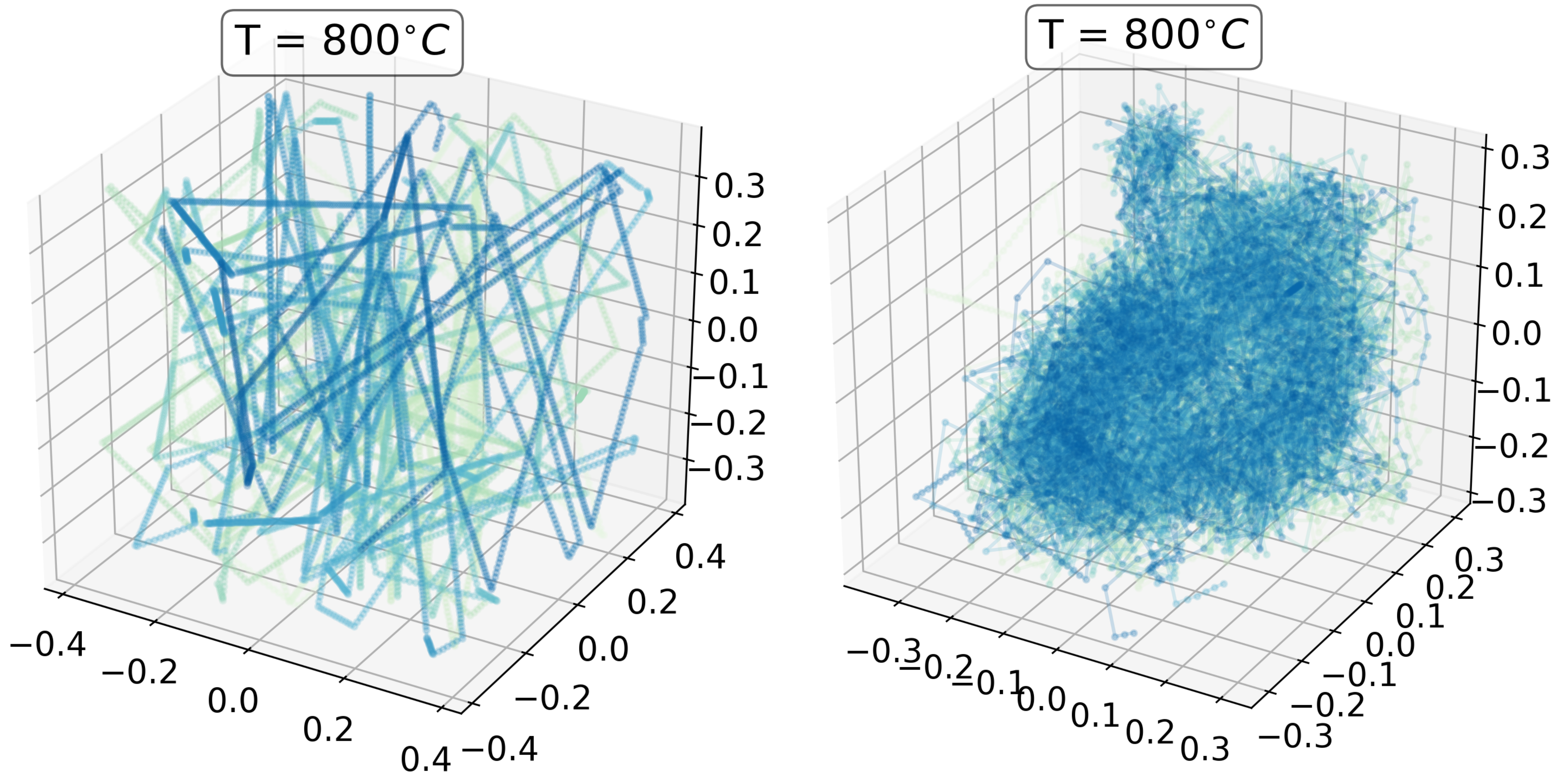
Any continuous function on a compact set of paths is well approximated by a finite linear combination of the terms of the Signature



$$f \approx \mathcal{L} \circ S \circ \mathbb{E}S_{path}$$

Experiments

Inferring the temperature of a gas



few collisions

many collisions

Model	Predictive MSE	
	few collisions	many collisions
DeepSets	$(1.4 \pm 0.7) \cdot 10^{-1}$	$(1.3 \pm 0.9) \cdot 10^{-1}$
RBF-RBF	$(2.1 \pm 0.9) \cdot 10^{-2}$	$(1.2 \pm 0.5) \cdot 10^{-1}$
RBF-GA	$(2.8 \pm 1.1) \cdot 10^{-2}$	$(4.9 \pm 5.2) \cdot 10^{-2}$
kerES	$(2.5 \pm 1.6) \cdot 10^{-2}$	$(8.7 \pm 4.6) \cdot 10^{-3}$
linSES	$(1.2 \pm 1.1) \cdot 10^{-2}$	$(7.6 \pm 5.8) \cdot 10^{-3}$

Table 3. We go from the "few collisions" settings to the "many collisions" by augmenting the radius of the particles.

Experiments

Inferring the crop yield of a region

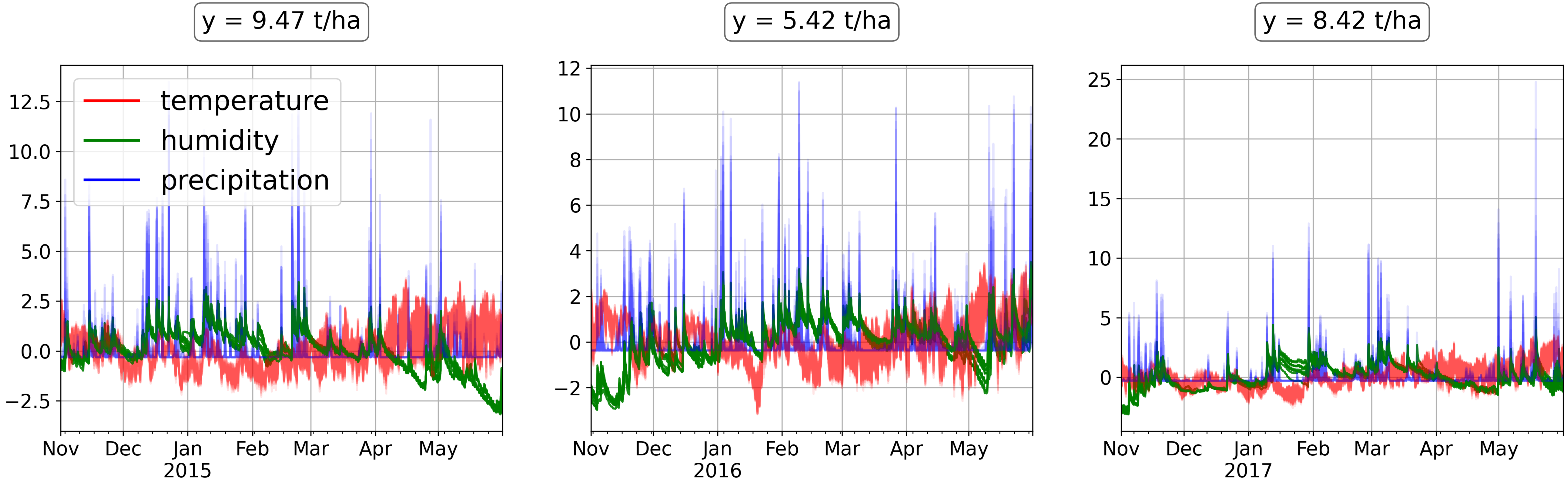


Figure 3. Normalised time-series of temperature, humidity and precipitation, measured over 10 different locations across a region within a year.

Experiments

Inferring the crop yield of a region

Model	Predictive MSE	
	0% dropped [running time]	30% dropped [running time]
RBF-RBF	0.671 ± 0.147 $[2h17m]$	0.819 ± 0.243 $[1h19m]$
kerES	0.563 ± 0.277 $[0m26s]$	0.583 ± 0.146 $[0m21s]$
linSES	0.616 ± 0.158 $[1m33s]$	0.646 ± 0.150 $[1m07s]$

Table 2. The fully observed time-series are of length 1,696 and the yield is in tons per hectar.

References

- Maud Lemercier, Cristopher Salvi, Theodoros Damoulas, Edwin V. Bonilla, and Terry Lyons. Distribution Regression for Continuous-Time Processes via the Expected Signature. *arXiv preprint arXiv:2006.05805,* 2020.
- Andreas Christmann and Ingo Steinwart. Universal kernels on non-standard input spaces. In *Advances in neural information processing systems*, pages 406-414, 2010.
- Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. In *Advances in neural information processing systems*, pages 10–18, 2012.
- Adeline Fermanian. Embedding and learning with signatures. *arXiv preprint arXiv:1911.13211*,2019
- Ilya Chevyrev and Harald Oberhauser. Signature moments to characterize laws of stochastic processes. *arXiv preprint arXiv:1810.10971*, 2018.