Spectral Diffusion Processes

Angus Phillips*†, Thomas Seror *‡, Michael Hutchinson†, Valentin De Bortoli‡, Arnaud Doucet†, Émile Mathieu§

*Equal contribution. †Dept. of Statistics, University of Oxford, UK. ‡Dept. of Computer Science ENS, CNRS, PSL University Paris, France. §Dept. of Engineering, University of Cambridge, UK.

UNIVERSITY OF OXFORD

Motivation and overview

- **Score-based generative models (SGMs)** [5, 6, 4] have proven highly effective for modelling densities on **finite dimensional space**.
- We often wish to model distributions over **functional spaces**.
- We represent **functional data in spectral space** to dissociate the **stochastic and space-time components**. Using dimensionality reduction techniques we then **sample from their stochastic component using finite dimensional SGMs**.

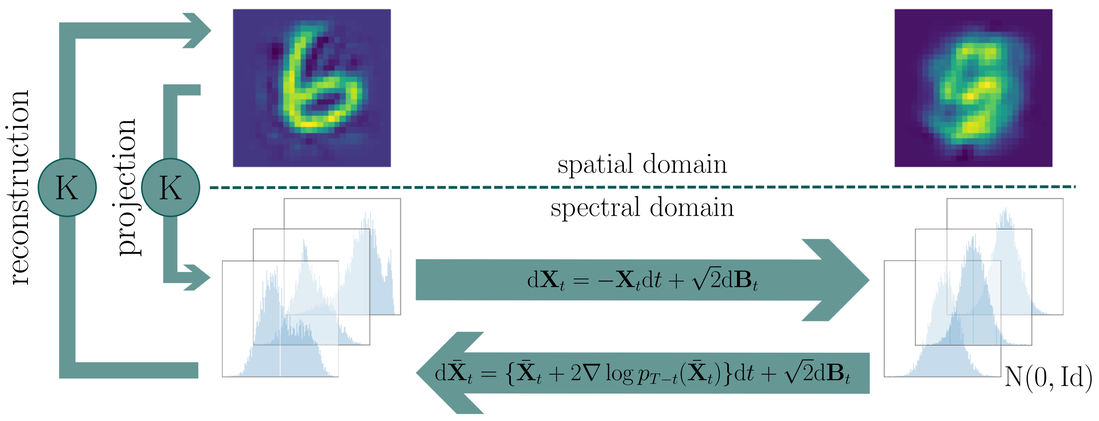

Figure 1: Illustration of our methodology. SGM is performed in a spectral space.

Stochastic processes

A stochastic process in $\mathbb{R}^d$ is a collection of $\mathbb{R}^d$-valued random variables $(\boldsymbol{Y}_x)_{x\in\mathcal{X}}$, with $\mathcal{X}$ a compact input space.

- Existing generative modelling approaches parametrise the finite dimensional marginals $\{\boldsymbol{Y}_{x_i} : i \in \{1,\dots,n\}, x_i \in \mathcal{X}\}$ for every $n \in \mathbb{N}$.
- The **Kolmogorov extension theorem** [2, Thm15.26] requires **exchangeability** and **consistency** of the marginals to define a valid probability distribution on the function space.

In order to derive a consistent method, we dissociate the **stochastic** part of the process from its **space-time** part using the **Karhunen-Loéve theorem**.

Theorem 1: Karhunen-Loéve Theorem

Let $(\boldsymbol{Y}_x)_{x\in\mathcal{X}}$ be a continuous stochastic process with continuous covariance function $K_Y(x_1, x_2) = \mathbb{E}[\langle \boldsymbol{Y}_{x_1}, \boldsymbol{Y}_{x_2}\rangle]$. Let $(e_m)_{m\in\mathbb{N}}$ be the orthonormal basis of $L^2(\mathcal{X})$ formed by the eigenfunctions of the linear operator $T_{K_Y} : f \mapsto \int_{\mathcal{X}} K_Y(x,\cdot)f(x)dx$ with eigenvalues $(\lambda_m)_{m\in\mathbb{N}}$. Then we have:

$$\lim_{M\to+\infty} \mathbb{E}\Big[\sup_{x\in\mathcal{X}} \big\| \boldsymbol{Y}_x - \sum_{m=0}^{M} \lambda_m^{\frac{1}{2}} Z_m e_m(x)\big\|^2\Big] \tag{1}$$

where for any $m \in \mathbb{N}$, $Z_m = \lambda_m^{-\frac{1}{2}} \int_{\mathcal{X}} \langle \boldsymbol{Y}_x, e_m(x)\rangle dx$.

- A natural and principled approximation of the process $(\boldsymbol{Y}_x)_{x\in\mathcal{X}}$ is given by $F_M \triangleq \sum_{m=0}^{M} \lambda_m{}^{1/2} Z_m e_m$,
- Finite marginals induced by $(F_M(x))_{x\in\mathcal{X}}$ are **exchangeable** and **consistent**.
- In order to sample from $(F_M(x))_{x\in\mathcal{X}}$ one only needs to sample $\{Z_m\}_{m=0}^M$, which is a finite dimensional problem **well suited to SGMs**.
- If $\{Z_m\}_{m=1}^M \sim \mathrm{N}(0, \mathrm{Id})$ then F_M is a **Gaussian process**.
- By taking a Gaussian reference measure in the SGM, we induce a **Gaussian process reference measure** in the original space. This induced Gaussian process is **closest to the target distribution** in the following sense:

Proposition 1

Let $\{\bar{Z}_{m,0}\}_{m=0}^M \sim \mathrm{N}(0, \mathrm{Id})$ and let π^0 be the distribution of $\sum_{m=0}^{+\infty} \lambda_m^{1/2} \bar{Z}_{m,0} e_m$ and π the target distribution. Let $\mathrm{GP}(\mathcal{X})$ be the space of Gaussian processes on $\mathcal{X}$ and assume K is the covariance kernel. Then, $\pi^0 \in \arg\min_{\pi_{\mathrm{GP}}\in\mathrm{GP}(\mathcal{X})} \mathrm{KL}\left(\pi|\pi_{\mathrm{GP}}\right)$.

- We can consider **arbitrary kernels** due to Mercer's theorem [3, Thm 1.1].

References

[1] C. T. H. Baker. Numerical Integration in the Treatment of Integral Equations. In G. Hämmerlin, editor, *Numerische Integration: Tagung Im Mathematischen Forschungsinstitut Oberwolfach Vom 1. Bis 7. Oktober 1978*, pages 44–53. Birkhäuser Basel, Basel, 1979.

[2] D. Charalambos and B. Aliprantis. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer-Verlag Berlin and Heidelberg GmbH & Company KG, 2013.

[3] J. Ferreira and V. Menegatto. Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral Equations and Operator Theory*, 64(1):61–81, 2009.

[4] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

[5] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019.

[6] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[7] G. Wynne and A. B. Duncan. A kernel two-sample test for functional data. *Journal of Machine Learning Research*, 23(73):1–51, 2022.

SP-SGM Algorithm

Algorithm 1 Spectral Process Score-Based Generative Model (SP-SGM)

Require: $T, \mathcal{D}, \theta_0, N_{\mathrm{iter}}, \varepsilon, K$

/// TRAINING ///
Get $\mathcal{D}^M, \{(\lambda_m, e_m)\}_{m=0}^M$ from $\mathcal{D}$ using Algorithm 2 ▷ Dataset projection
for $n \in \{0, \dots, N_{\mathrm{iter}} - 1\}$ **do**
  Get $\{\mathbf{Y}_{m,0}\}_{m=0}^M$ mini-batch from $\mathcal{D}^M$
  $t \sim U([\varepsilon, T])$ ▷ Uniform sampling between ε and T
  $\mathbf{Y}_{m,t} = \mathrm{e}^{-t}\mathbf{Y}_{m,0} + (1 - \mathrm{e}^{-2t})^{1/2} G, \quad G \sim \mathrm{N}(0, \mathrm{Id})$ ▷ Diffuse
  Get DSM loss $\ell(\theta_n)$ ▷ Compute score matching loss
  $\theta_{n+1} = \mathtt{optimiser\_update}(\theta_n, \ell(\theta_n))$ ▷ ADAM optimiser step
$\theta^\star = \theta_{N_{\mathrm{iter}}}$
/// SAMPLING ///
$\{\bar{Y}_{m,0}\}_{m=0}^M \sim \mathrm{N}(0, \mathrm{Id})$ ▷ Sample from Gaussian distribution
$b_\theta^\star(t, y) = \mathbf{s}_{\theta^\star}(T - t, \bar{y})$ for any $t \in [0, T]$, $\bar{y} \in \mathbb{R}^{M+1}$ ▷ Reverse process drift
$\{\bar{Y}_{m,n}\}_{m=0,n=0}^{M,N}$ Euler-Maruyama with drift $b_\theta^\star$ ▷ Approximate reverse diffusion
return $\theta^\star$, $x \mapsto \sum_{m=0}^M \lambda_m^{1/2} \bar{Y}_{m,N} e_m(x)$

Algorithm 2 Spectral dataset projection

Require: $\mathcal{D} = \{\{\mathbf{Y}^i_{x_{i,n}}\}_{n=1}^{N_i}\}_{i=1}^L, K, M$

Gram matrix $\mathrm{K}_{i,j} = K(x_i, x_j)$ for $x_i, x_j \in \{x_s\}_{s=1}^S = \{x_{i,n} : i \in \{1, ..., L\}, n \in \{1, ..., N_i\}\}$ and $S = \sum_{i=1}^L N_i$
Solve eigensystem $\frac{1}{S}\mathrm{K}u_m = \lambda_m^{\mathrm{mat}} u_m$
$\hat{e}_m(x) \approx (\sqrt{S}\lambda_m^{\mathrm{mat}})^{-1} \sum_{s=1}^S K(x, x_s) u_m(x_s)$ ▷ [1, Theorem 3.4]
$Z_m^i \approx \frac{1}{N_i} \sum_{n=1}^{N_i} (\mathbf{Y}_{x_n} - \mu(x_n))(\lambda_m^{\mathrm{mat}})^{-1/2} \hat{e}_m(x_n)$ where $\mu(x) = \mathbb{E}[\mathbf{Y}_x]$
return $\{Z_m^i\}_{i=1,\dots,L,\ m=0,\dots,M}, \{(\lambda_m, e_m)\}_{m=0}^M$

1D datasets

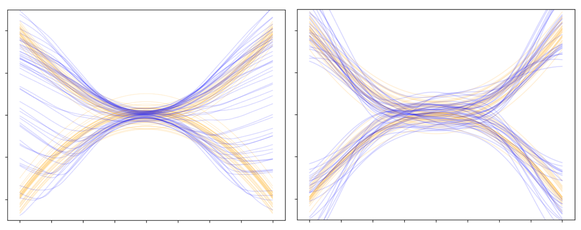

Figure 2: Samples from the Quadratic dataset (orange), and from a trained NP [Left, blue] and a trained SP-SGM [Right, blue].

	SP-SGM	NP	GP
Quadratic	**5.4**$_{\pm 0.7}$	8.6$_{\pm 1.5}$	100.0$_{\pm 0.0}$
Melbourne	**5.3**$_{\pm 0.7}$	10.1$_{\pm 1.9}$	20.1$_{\pm 4.0}$
Gridwatch	**4.7**$_{\pm 0.5}$	51.8$_{\pm 15.1}$	29.2$_{\pm 5.5}$

Table 1: Power (percent) of a kernel two-sample hypothesis test on 1D datasets. Lower is better. Statistically significant best result is in bold.

- We are able to capture bi-modality while Neural processes (NPs) are not.
- Quantitatively we outperform Gaussian processes (GPs) and NPs based on the power of a two-sample hypothesis test [7].

MNIST dataset

- By performing the diffusion in the spectral space, we capture and incorporate spatial correlations while an ordinary diffusion process does not.

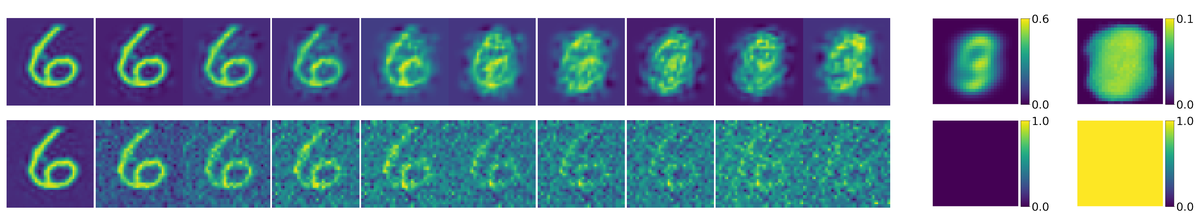

Figure 3: Forward process in SP-SGM [Top] vs standard SGM [Bottom] on MNIST digits. Pixel-wise mean and standard deviation of reference measure in rightmost columns respectively.

We also performed ablation studies on truncation order and kernel choice:

- Best performance (in terms of functional MMD) was obtained for intermediate numbers of spectral components M.
- Best performance was obtained with the covariance kernel. Smoothness imposed by the RBF kernel tends to produce blurry samples.

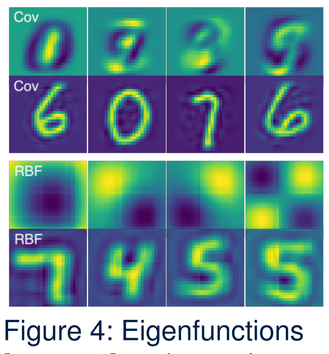

Figure 4: Eigenfunctions [row 1, 3] and samples from SP-SGM [row 2, 4] for covariance and RBF kernels.

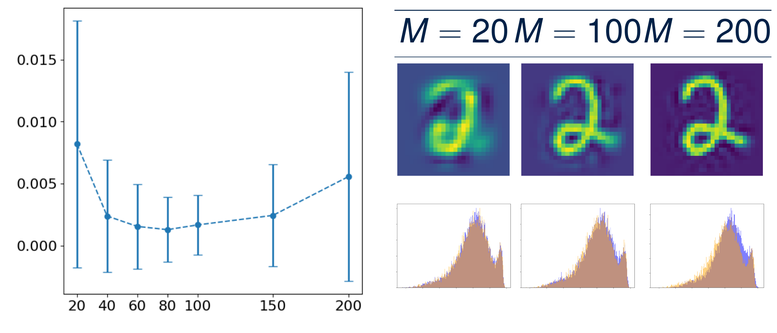

Figure 5: On the left, functional MMD vs M. Lower is better. On the right, Karhunen-Loève recompositions of MNIST samples [top row] and distribution of first spectral component from the dataset [orange] and from the SGM in spectral space [blue].