Geometric Neural Diffusion Processes

Émile Mathieu

May 29th, 2023

ICTP
Youth in High-Dimensions:
Recent Progress in Machine Learning, High-Dimensional Statistics and Inference

UNIVERSITY OF CAMBRIDGE

Outline

1. Very brief intro to generative modelling
2. Background on continuous diffusion models
3. Extension to function space: neural diffusion processes
4. Modelling feature fields: geometric neural diffusion processes

Émile
Mathieu

Vincent
Dutordoir

Michael
Hutchinson

Valentin
De Bortoli

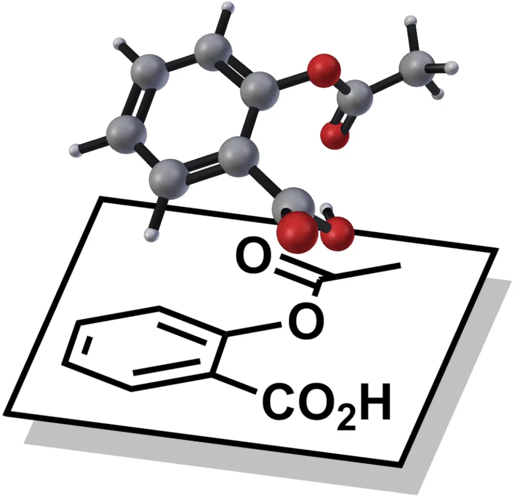
Yee Whye
Teh

Richard E.
Turner

Deep generative modelling

Motivating examples

Molecular conformation generation (Xu et al., 2022)

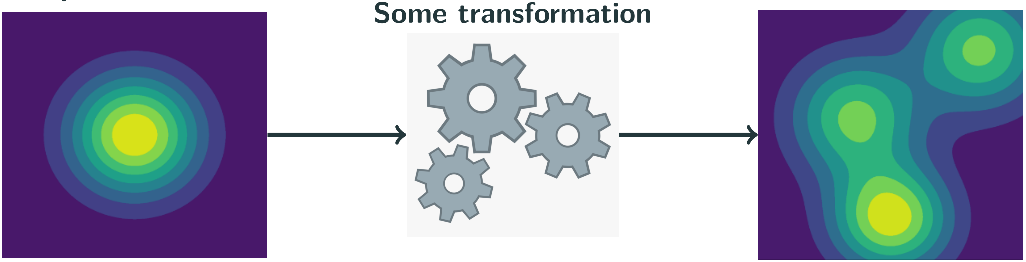


Motivating examples (Cont'd)

Meteorology (nowcasting) (Ravuri et al., 2021)

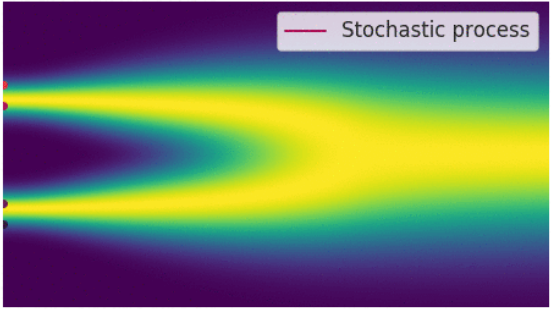
What is generative modelling?

Generically, it is the parametrisation of a density. We assume access to samples for training purposes (vs. unnormalised density for *sampling*). We might want to draw more samples, or evaluate the likelihood.



Continuous diffusion models

Principles of continuous diffusion models



- Idea: Destruct data with *continuous* series of noise.
- Do this by constructing an **SDE** forward noising process $(\mathbf{Y}_t)_{t\in[0,T]}$.
- Have this noising converge to a **known distribution**.
- **Invert** this SDE noising process to get $(\bar{\mathbf{Y}}_t)_{t\in[0,T]} = (\mathbf{Y}_{T-t})_{t\in[0,T]}$.

Continuous noising processes

A wide class of stochastic differential equations (SDEs) can be written as:

$$\mathrm{d}\mathbf{Y}_t = b(t, \mathbf{Y}_t)\,\mathrm{d}t + \sigma(t, \mathbf{Y}_t)\,\mathrm{d}\mathbf{B}_t \tag{1}$$

$$= -\tfrac{1}{2}\nabla_{\mathbf{Y}_t} U(\bar{\mathbf{Y}}_t)\beta_t\,\mathrm{d}t + \sqrt{\beta_t}\,\mathrm{d}\mathbf{B}_t \tag{2}$$

admits **invariant** measure: $\mu_{\mathrm{inv}} \propto \mathrm{e}^{-U(y)}\mathrm{Leb}$ (Durmus, 2016, Section 2.4).

Continuous noising processes

A wide class of stochastic differential equations (SDEs) can be written as:

$$\mathrm{d}\mathbf{Y}_t = b(t, \mathbf{Y}_t)\,\mathrm{d}t + \sigma(t, \mathbf{Y}_t)\,\mathrm{d}\mathbf{B}_t \tag{1}$$

$$= -\tfrac{1}{2}\nabla_{\mathbf{Y}_t} U(\bar{\mathbf{Y}}_t)\beta_t\,\mathrm{d}t + \sqrt{\beta_t}\,\mathrm{d}\mathbf{B}_t \tag{2}$$

admits **invariant** measure: $\mu_{\mathrm{inv}} \propto \mathrm{e}^{-U(y)}\mathrm{Leb}$ (Durmus, 2016, Section 2.4).

A couple of common examples:

	Brownian Motion	Ornstein-Uhlenbeck process
$b(t, \mathbf{Y}_t)$	0	$-\frac{1}{2}\mathbf{Y}_t$
$\sigma(t, \mathbf{Y}_t)$	1	1
Invariant measure	Lebesgue	$\mathrm{N}(0, 1)$
Conditional $\mathbf{Y}_t \vert \mathbf{Y}_0$	$\mathbf{Y}_0 + \mathbf{B}_t$	$e^{-t}\mathbf{Y}_0 + \mathbf{B}_{1-e^{-2t}}$

Continuous score-based models: Time reversal process

Theorem 1: (Cattiaux et al., 2021; Haussmann and Pardoux, 1986)

Under mild conditions on p_0, the time-reversed process $(\bar{\mathbf{Y}}_t)_{t \geq 0} = (\mathbf{Y}_{T-t})_{t \in [0,T]}$, with forward process $\mathrm{d}\mathbf{Y}_t = b(t, \mathbf{Y}_t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}\mathbf{B}_t$, also satisfies an SDE given by

$$\mathrm{d}\bar{\mathbf{Y}}_t = \left[-b(T-t, \bar{\mathbf{Y}}_t) + \sigma(T-t)^2 \nabla \log p_{T-t}(\bar{\mathbf{Y}}_t)\right] \mathrm{d}t + \sigma(T-t)\,\mathrm{d}\mathbf{B}_t,$$

assuming $\bar{\mathbf{Y}}_0$ is distributed the same as $\mathbf{Y}_T$.

Continuous score-based models: Time reversal process

Theorem 2: (Cattiaux et al., 2021; Haussmann and Pardoux, 1986)

Under mild conditions on p_0, the time-reversed process $(\bar{\mathbf{Y}}_t)_{t \geq 0} = (\mathbf{Y}_{T-t})_{t \in [0,T]}$, with forward process $\mathrm{d}\mathbf{Y}_t = b(t, \mathbf{Y}_t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}\mathbf{B}_t$, also satisfies an SDE given by

$$\mathrm{d}\bar{\mathbf{Y}}_t = \left[-b(T-t, \bar{\mathbf{Y}}_t) + \sigma(T-t)^2\, \nabla \log p_{T-t}(\bar{\mathbf{Y}}_t) \right] \mathrm{d}t + \sigma(T-t)\,\mathrm{d}\mathbf{B}_t,$$

assuming $\bar{\mathbf{Y}}_0$ is distributed the same as $\mathbf{Y}_T$.

Problems:

1. We do not have access to $\mathbf{Y}_T \Rightarrow$ Approximate as $\mathbf{Y}_T \approx \mathbf{Y}_\infty$!
2. The Stein score $\nabla \log p_t$ is intractable (requires solving Fokker-Planck...) $\Rightarrow$ learn it!
3. Cannot solve the SDE exactly $\Rightarrow$ discretise!

Score approximation

- One make use of the following denoising score matching identity

$$\nabla_{y_t} \log p_t(y_t) = \int_{\mathbb{R}^d} \nabla \log p_{t|0}(y_t|y_0)\ p_{0|t}(y_0|y_t)\mathrm{d}y_0, \tag{3}$$

  where $p_{t|0}(y_t|y_0)$ is the transition density of the forward noising process.
- It follows directly that $\nabla \log p_t$ is the minimiser of $\mathcal{L}(\mathbf{s}) = \mathbb{E}\left[\| \mathbf{s}(t, \mathbf{Y}_t) - \nabla_{y_t} \log p_{t|0}(\mathbf{Y}_t|\mathbf{Y}_0)\|^2\right]$ over functions $\mathbf{s} : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$.
- This readily gives a loss over the finite parameters $\theta \in \Theta$ of the neural network $\mathbf{s}_\theta : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$:

$$\mathcal{L}(\theta; \lambda(t)) = \mathbb{E}[\lambda(t)\| \mathrm{s}_\theta(t, \mathbf{Y}_t) - \nabla \log p_t(\mathbf{Y}_t|\mathbf{Y}_0)\|^2] \tag{4}$$

  where the expectation is taken over the joint $(t, \mathbf{Y}_0, \mathbf{Y}_t)$.

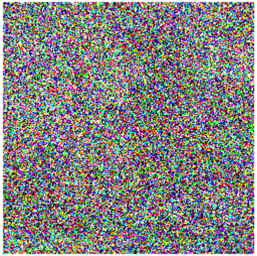
Sampling from SDEs

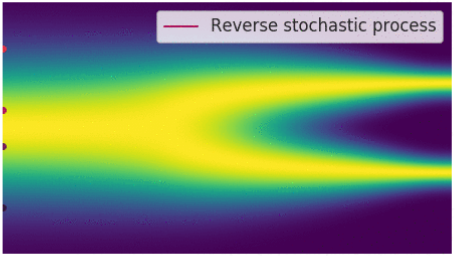
Euler–Maruyama discretisation of backward SDE

$$\mathrm{d}\bar{\mathbf{Y}}_t = \left[-b(T-t,\bar{\mathbf{Y}}_t) + \sigma(T-t)^2 \nabla \log p_{T-t}(\bar{\mathbf{Y}}_t)\right]\mathrm{d}t + \sigma(T-t)\,\mathrm{d}\mathbf{B}_t,$$

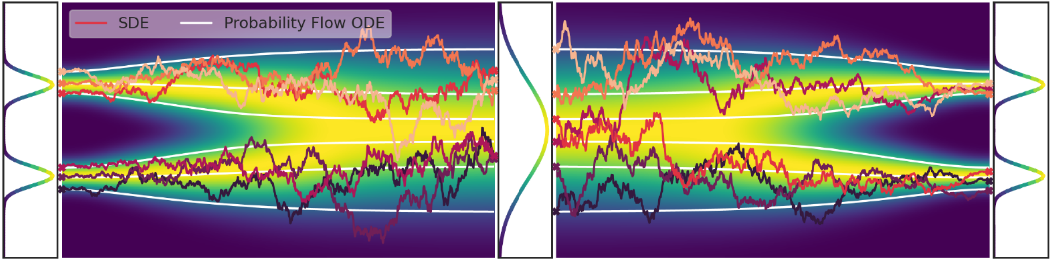
with some time step γ_t gives an approximate trajectory of the SDE

$$\bar{\mathbf{Y}}_{t+1} \approx \bar{\mathbf{Y}}_t + \gamma_t[-b(T-t,\bar{\mathbf{Y}}_t) + \sigma(T-t)^2 \underbrace{\nabla \log p_{T-t}(\bar{\mathbf{Y}}_t)]}_{\approx \mathbf{s}_\theta(T-t)} + \sqrt{\gamma_t}\,\sigma(T-t,\bar{\mathbf{Y}}_t)\,\mathbf{Z}_t.$$





Recap: Continuous diffusion models



- Continuously **noise** data samples with forward SDE
- Aim: time-reversal of this process $\Rightarrow$ **denoising** process

Geometric neural diffusion processes

Overview

- Aim
  - Construct probabilistic model over feature/tensor fields.
  - Enforce invariance w.r.t. group transformations.
- Achieve this by
  - Constructing diffusion model over function spaces by correlating marginals.
  - Enforce invariance using a kernel and score network which are group equivariant.

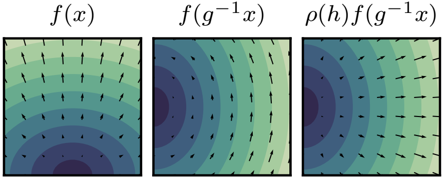
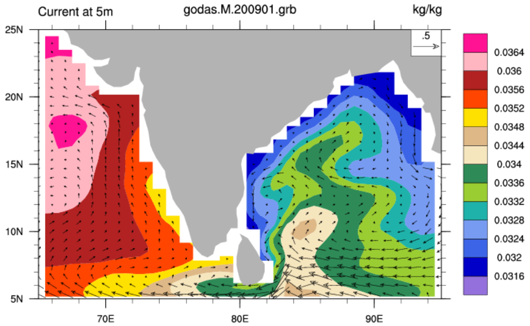
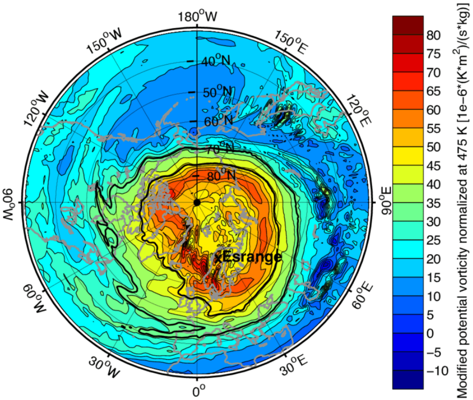


Figure 1: Transformation of feature field.

Feature fields: e.g. salinity, current and vorticity



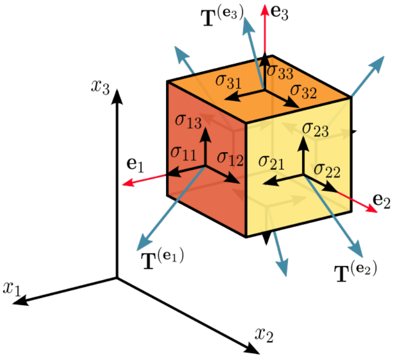
(a) Salinity scalar field and current vector field.



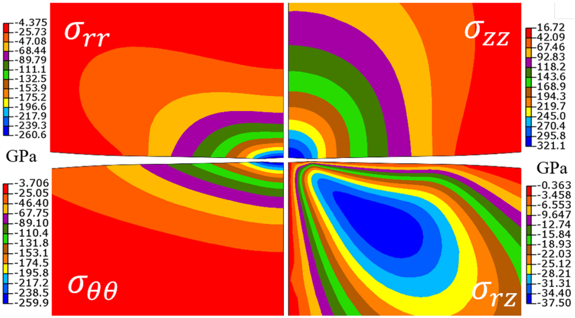
(b) Vorticity pseudo-vector field.

Figure 2: Credits Gregory et al. 2020.

Feature fields: e.g. Cauchy stress



(a) Stress tensor diagram.



(b) Components of the three-dimensional stress tensor in diamond cell.

Figure 3: Credits to wikipedia and Gregory et al. 2020.

Steerable feature fields

A **feature field** is a tuple (f, ρ) with $f : \mathcal{X} \to \mathbb{R}^d$ a mapping between $x \in \mathcal{X}$ to some feature $f(x)$ with representation $\rho : G \to \mathrm{GL}(\mathbb{R}^d)$ (Scott and Serre, 1996).

Steerable feature fields

A **feature field** is a tuple (f, ρ) with $f : \mathcal{X} \to \mathbb{R}^d$ a mapping between $x \in \mathcal{X}$ to some feature $f(x)$ with representation $\rho : G \to \mathrm{GL}(\mathbb{R}^d)$ (Scott and Serre, 1996).
The action of $G = \mathrm{E}(d) = \mathrm{T}(d) \rtimes \mathrm{O}(d)$ on the feature field f given by

$$g \cdot f(x) = (uh) \cdot f(x) \triangleq \rho(h) f\left(h^{-1}(x - u)\right) \tag{5}$$

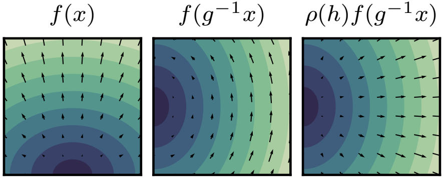
Steerable feature fields

A **feature field** is a tuple (f, ρ) with $f : \mathcal{X} \to \mathbb{R}^d$ a mapping between $x \in \mathcal{X}$ to some feature $f(x)$ with representation $\rho : G \to \mathrm{GL}(\mathbb{R}^d)$ (Scott and Serre, 1996).
The action of $G = \mathrm{E}(d) = \mathrm{T}(d) \rtimes \mathrm{O}(d)$ on the feature field f given by

$$g \cdot f(x) = (uh) \cdot f(x) \triangleq \rho(h)\, f\left(h^{-1}(x - u)\right) \tag{5}$$

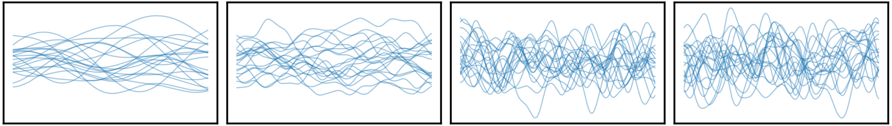
Typical examples of feature fields include:

- **Scalar fields** $\rho_{\text{triv}}(h) \triangleq 1$ e.g. temperature or potential fields.
- **Vectors fields** $\rho_{\text{Id}}(h) \triangleq h$ e.g. wind or force fields.
- **Rank-2 tensors** $\rho(h) = h \otimes h$, $g \cdot f(x) = (h \otimes h) \operatorname{vec}\left(f\left(g^{-1}x\right)\right) = h\, f\left(g^{-1}x\right) h^\top$.



Continuous diffusion on function space

Continuous noising process

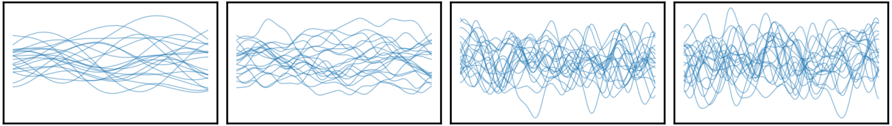


We construct the forward **noising process** $(\mathbf{Y}_t(x))_{t \geq 0} \triangleq (\mathbf{Y}_t(x^1), \ldots, \mathbf{Y}_t(x^n))_{t \geq 0}$ defined by the multivariate SDE (multivariate Ornstein-Uhlenbeck process)

$$\mathrm{d}\mathbf{Y}_t(x) = \frac{1}{2}\left\{m(x) - \mathbf{Y}_t(x)\right\}\beta_t \,\mathrm{d}t + \beta_t^{1/2}\mathrm{K}(x,x)^{1/2}\,\mathrm{d}\mathbf{B}_t, \tag{6}$$

where $\mathrm{K}(x,x)_{i,j} = k(x^i, x^j)$ with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a kernel and $m : \mathcal{X} \to \mathcal{Y}$.

Continuous noising process



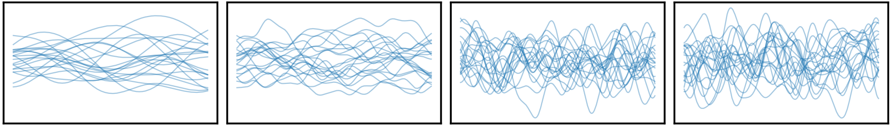
We construct the forward **noising process** $(\mathbf{Y}_t(x))_{t\geq 0} \triangleq (\mathbf{Y}_t(x^1), \ldots, \mathbf{Y}_t(x^n))_{t\geq 0}$ defined by the multivariate SDE (multivariate Ornstein-Uhlenbeck process)

$$\mathrm{d}\mathbf{Y}_t(x) = \frac{1}{2}\left\{m(x) - \mathbf{Y}_t(x)\right\}\beta_t \,\mathrm{d}t + \beta_t^{1/2}\mathrm{K}(x,x)^{1/2}\,\mathrm{d}\mathbf{B}_t, \tag{6}$$

where $\mathrm{K}(x,x)_{i,j} = k(x^i, x^j)$ with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a kernel and $m : \mathcal{X} \to \mathcal{Y}$.

- $\mathbf{Y}_t(x) \to \mathrm{N}(m(x), \mathrm{K}(x,x))$ with geometric rate, for any $x \in \mathcal{X}^n$.
- $\mathbf{Y}_t \to \mathrm{GP}(m,k) \triangleq \mathbf{Y}_\infty$ (Phillips et al., 2022).

Continuous noising process



We construct the forward **noising process** $(\mathbf{Y}_t(x))_{t\geq 0} \triangleq (\mathbf{Y}_t(x^1), \ldots, \mathbf{Y}_t(x^n))_{t\geq 0}$ defined by the multivariate SDE (multivariate Ornstein-Uhlenbeck process)

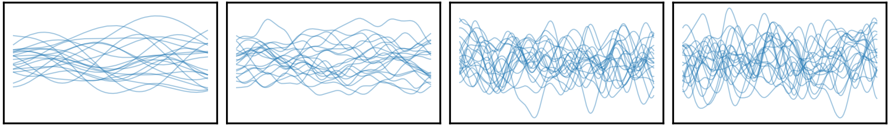
$$\mathrm{d}\mathbf{Y}_t(x) = \tfrac{1}{2}\left\{m(x) - \mathbf{Y}_t(x)\right\}\beta_t \, \mathrm{d}t + \beta_t^{1/2}\mathrm{K}(x,x)^{1/2}\,\mathrm{d}\mathbf{B}_t, \tag{6}$$

where $\mathrm{K}(x,x)_{i,j} = k(x^i, x^j)$ with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a kernel and $m : \mathcal{X} \to \mathcal{Y}$.

- $\mathbf{Y}_t(x) \to \mathrm{N}(m(x), \mathrm{K}(x,x))$ with geometric rate, for any $x \in \mathcal{X}^n$.
- $\mathbf{Y}_t \to \mathrm{GP}(m,k) \triangleq \mathbf{Y}_\infty$ (Phillips et al., 2022).
- $\mathbf{Y}_t$ interpolates between $\mathbf{Y}_0$ and $\mathbf{Y}_\infty$.
- $\mathbf{Y}_t(x)|\boldsymbol{y}_0 = \mathrm{N}(m_t(x;\boldsymbol{y}_0), \mathrm{K}_t(x,x;\boldsymbol{y}_0))$ for any $x \in \mathcal{X}^n$.

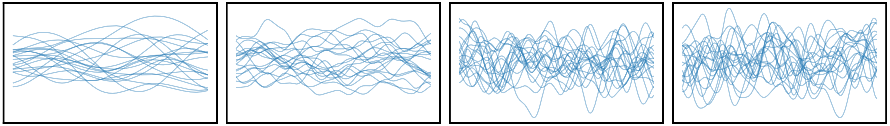
Continuous noising process

- $k(x, x') = k_{\mathrm{rbf}}(x, x') = \sigma^2 \exp\left(\frac{\|x-x'\|^2}{2l^2}\right)$, with $l = 1$.

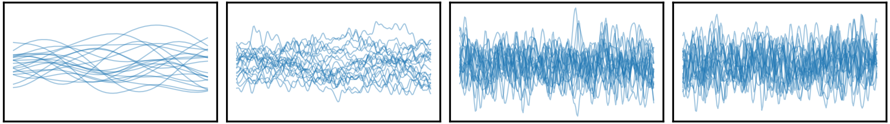


Continuous noising process

- $k(x, x') = k_{\mathrm{rbf}}(x, x') = \sigma^2 \exp\left(\frac{\|x - x'\|^2}{2l^2}\right)$, with $l = 1$.

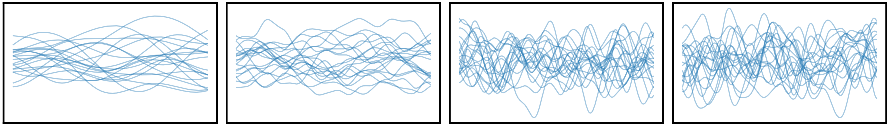


- $k(x, x') = k_{\mathrm{rbf}}(x, x')$, with $l = 0.2$.

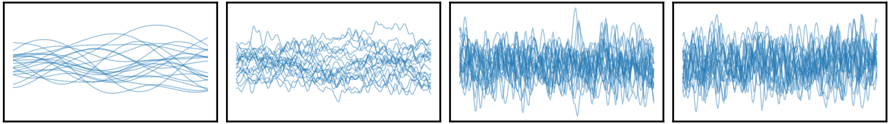


Continuous noising process

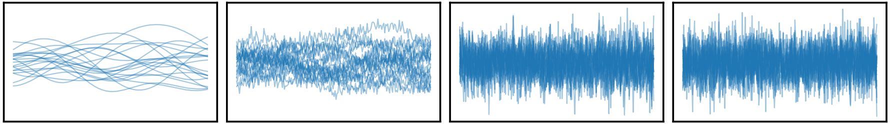
- $k(x,x') = k_{\mathrm{rbf}}(x,x') = \sigma^2 \exp\left(\frac{\|x-x'\|^2}{2l^2}\right)$, with $l = 1$.



- $k(x,x') = k_{\mathrm{rbf}}(x,x')$, with $l = 0.2$.



- $k(x,x') = \delta_x(x')$.



Denoising process

Proposition 1: Time-reversal.

Under mild conditions on the distribution of $\mathbf{Y}_0(x)$, the **time-reversal process** $(\bar{\mathbf{Y}}_t(x))_{t \geq 0}$ also satisfies an SDE (Cattiaux et al., 2021; Haussmann and Pardoux, 1986) given by

$$\mathrm{d}\bar{\mathbf{Y}}_t(x) = \{-\tfrac{1}{2}(m(x) - \bar{\mathbf{Y}}_t(x)) + \mathrm{K}(x,x)\nabla \log p_{T-t}(\bar{\mathbf{Y}}_t(x))\}\beta_{T-t}\,\mathrm{d}t + \beta_{T-t}^{1/2}\mathrm{K}(x,x)^{1/2}\,\mathrm{d}\mathbf{B}_t, \tag{7}$$

with $\bar{\mathbf{Y}}_0 \sim \mathrm{GP}(m,k)$ and p_t the density of $\mathbf{Y}_t(x)$ w.r.t. the Lebesgue measure.

Score approximation

- We directly approximate $\mathrm{K}(x,x)\nabla \log p_t$ with a neural network $(\mathrm{Ks})_\theta : [0,T] \times \mathcal{X}^n \times \mathcal{Y}^n \to \mathrm{T}\mathcal{Y}^n$, where $\mathrm{T}\mathcal{Y}$ is the tangent bundle of $\mathcal{Y}$. The **conditional score** of the noising process (6) is given by

$$\nabla_{\mathbf{Y}_t} \log p_t(\mathbf{Y}_t(x)|\mathbf{Y}_0(x)) = -\Sigma_{t|0}^{-1}(\mathbf{Y}_t(x) - m_{t|0}) = -\sigma_{t|0}^{-1}\mathrm{K}(x,x)^{-1/2}\varepsilon\,, \tag{8}$$

since $\mathbf{Y}_t = m_{t|0} + \Sigma_{t|0}^{1/2}\varepsilon$ with $\varepsilon \sim \mathrm{N}(0, \mathrm{Id})$, and $\Sigma_{t|0} = \sigma_{t|0}^2\mathrm{K}$.

Score approximation

- We directly approximate $\mathrm{K}(x,x)\nabla \log p_t$ with a neural network $(\mathrm{Ks})_\theta : [0,T] \times \mathcal{X}^n \times \mathcal{Y}^n \to \mathrm{T}\mathcal{Y}^n$, where $\mathrm{T}\mathcal{Y}$ is the tangent bundle of $\mathcal{Y}$. The **conditional score** of the noising process (6) is given by

$$\nabla_{\mathbf{Y}_t} \log p_t(\mathbf{Y}_t(x)|\mathbf{Y}_0(x)) = -\Sigma_{t|0}^{-1}(\mathbf{Y}_t(x) - m_{t|0}) = -\sigma_{t|0}^{-1}\mathrm{K}(x,x)^{-1/2}\varepsilon, \qquad (8)$$

since $\mathbf{Y}_t = m_{t|0} + \Sigma_{t|0}^{1/2}\varepsilon$ with $\varepsilon \sim \mathrm{N}(0, \mathrm{Id})$, and $\Sigma_{t|0} = \sigma_{t|0}^2\mathrm{K}$.

- We learn the **preconditioned score** $(\mathrm{Ks})_\theta$ by minimising the following denoising score matching (DSM) loss (Vincent et al., 2010) weighted by $\Lambda(t) = \sigma_{t|0}^2 \, \mathrm{K}^\top \mathrm{K}$

$$\mathcal{L}(\theta; \Lambda(t)) = \mathbb{E}[\|\mathrm{s}_\theta(t, \mathbf{Y}_t) - \nabla \log p_t(\mathbf{Y}_t|\mathbf{Y}_0)\|_{\Lambda(t)}^2] = \mathbb{E}[\|\sigma_{t|0} \cdot (\mathrm{K}s)_\theta(t, \mathbf{Y}_t) + \mathrm{K}^{1/2}\varepsilon\|_2^2],$$

where $\|x\|_\Lambda^2 = x^\top \Lambda x$.

Invariant neural diffusion processes

Invariant stochastic processes

- A stochastic process $f \sim \mu$ is said to be $G-$invariant if $\mu(g \cdot \mathsf{A}) = \mu(\mathsf{A})$ for any $g \in G$, with $\mu \in \mathcal{P}(\mathrm{C}(\mathcal{X}, \mathcal{Y}))$, and $\mathsf{A} \subset \mathrm{C}(\mathcal{X}, \mathcal{Y})$ measurable.

Invariant stochastic processes

- A stochastic process $f \sim \mu$ is said to be $G-$invariant if $\mu(g \cdot \mathsf{A}) = \mu(\mathsf{A})$ for any $g \in G$, with $\mu \in \mathcal{P}(\mathrm{C}(\mathcal{X}, \mathcal{Y}))$, and $\mathsf{A} \subset \mathrm{C}(\mathcal{X}, \mathcal{Y})$ measurable.
- From a sample perspective, this means that with input-output pairs $\mathcal{C} = \{(x^c, f(x^c))\}_{c \in C}$, and denoting the action of G on this set as $g \cdot \mathcal{C} \triangleq \{(g^{-1} \cdot x^c, \rho(g) f(x^c))\}_{c \in C}$, $f \sim \mu$ is $G-$invariant if and only if $g \cdot \mathcal{C}$ has the same distribution as $\mathcal{C}$.

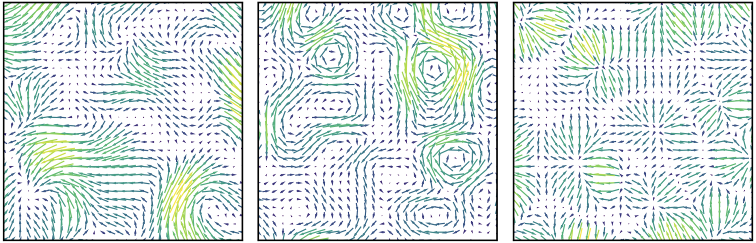
Invariant Gaussian processes

Proposition 2: Invariant (stationary) Gaussian process (Holderrieth et al., 2021).

We have that a Gaussian process $\mathrm{GP}(m, k)$ is G-invariant if and only if its mean m and covariance k are G-equivariant—that is, for all $x, x' \in \mathcal{X}, g \in G$

$$m(g \cdot x) = \rho(g)m(x) \quad \text{and} \quad k(g \cdot x, g \cdot x') = \rho(g)k(x, x')\rho(g)^\top. \tag{9}$$

$\mathrm{E}(d)$-invariant Gaussian processes



- $\mathrm{E}(d)$-equivariant kernels $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ include
  - Diagonal kernels $k = k_0 \, \mathrm{Id}$ with k_0 invariant (Holderrieth et al., 2021).
  - $k_{\mathrm{curl}} = k_0 A$ with $A(x, x') = \mathrm{Id} - \frac{(x-x')(x-x')^\top}{l^2}$ (Macêdo and Castro, 2010).
  - $k_{\mathrm{div}} = k_0 B$ with $B(x, x') = \frac{(x-x')(x-x')^\top}{l^2} + \left(n - 1 - \frac{\|x-x'\|^2}{l^2}\right) \mathrm{Id}$.

Invariant neural diffusion processes

Proposition 3: Invariant neural diffusion process (Yim et al., 2023).

We denote by $(\bar{\mathbf{Y}}_t^\theta(x))_{x \in \mathcal{X}, t \in [0,T]}$ the process induced by

$$\mathrm{d}\bar{\mathbf{Y}}_t^\theta(x) = \{-\tfrac{1}{2}(m(x) - \bar{\mathbf{Y}}_t(x)) + \mathbf{K}\mathbf{s}_\theta(T-t, x, \bar{\mathbf{Y}}_t(x))\}\beta_{T-t}\,\mathrm{d}t + \beta_{T-t}^{1/2}\mathrm{K}(x,x)^{1/2}\,\mathrm{d}\mathbf{B}_t, \tag{10}$$

and with initial (limiting) process is given by $\mathcal{L}(\bar{\mathbf{Y}}_0) = \mathrm{GP}(m, k)$.
Assuming m and k are respectively G-equivariant, if we additionally have that the score network is G-equivariant vector field, i.e. $\mathbf{s}_\theta(t, g \cdot x, \rho(g)y) = \rho(g)\mathbf{s}_\theta(t, x, y)$ for all $x \in \mathcal{X}, g \in G$, then for any $t \in [0, T]$ the process $(\bar{\mathbf{Y}}_t^\theta(x))_{x \in \mathcal{X}}$ is G-invariant.

Invariant neural diffusion processes (Cont'd)

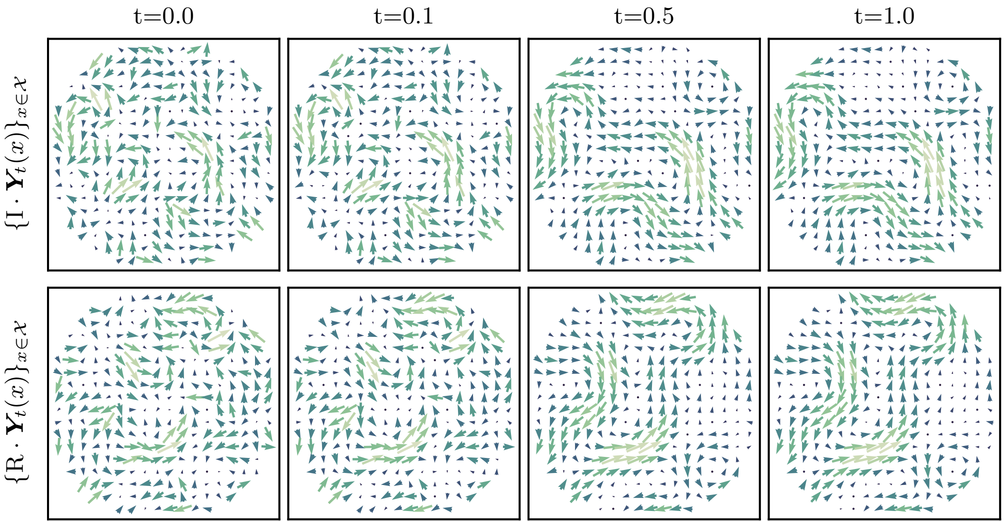


Figure 4: $(g\cdot\mathbf{Y}_t(x))_{x\in\mathcal{X}}$

Equivariant conditional processes

A stochastic process with distribution μ given a context $\mathcal{C}$ is said to be conditionally $G-$equivariant if the conditional satisfies $\mu(\mathsf{A}|g \cdot \mathcal{C}) = \mu(g \cdot \mathsf{A}|\mathcal{C})$, or equivalently $\mu(g^{-1} \cdot \mathsf{A}|g \cdot \mathcal{C}) = \mu(\mathsf{A}|\mathcal{C})$, for any $g \in G$ and $\mathsf{A} \in \mathrm{C}(\mathcal{X}, \mathcal{Y})$ measurable.

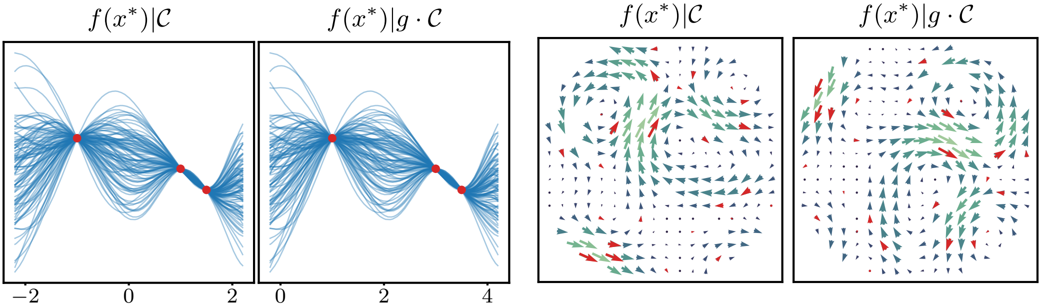


Figure 5: Samples from equivariant neural diffusion processes conditioned on context set $\mathcal{C}$ (in red) for scalar (*Left*) and 2D vector (*Right*) fields. Same model is then conditioned on transformed context $g \cdot \mathcal{C}$.

Equivariant conditional processes

Proposition 4: Equivariant conditional process.

Assume a stochastic process $f \sim \mu$ is $G-$invariant. Then the conditional process $f|\mathcal{C}$ given a set of observations $\mathcal{C}$ is G-equivariant.

Conditional neural diffusion processes: Predictor

- **Aim**: Sample $y^* \sim p(\cdot|x^*, \mathcal{C})$ given a set of observations context $\mathcal{C} = \{(x^c, y^c)\}_{c \in C}$.
- **Predictor**: Single-step backward: $\mathbf{Y}^*_{t-\gamma}, \mathbf{Y}^c_{t-\gamma} \leftarrow \mathbf{Y}^*_t, \mathbf{Y}^c_t$
  - **Noise context**: $\mathbf{Y}^c_t | \mathbf{Y}^c_0 \sim p_{t|0}$
  - **Denoise joint**: with $x \triangleq [x^*, x^c]$
    $\left[\_, \tilde{\mathbf{Y}}^*_{t-\gamma}\right] =$
    $[\mathbf{Y}^c_t, \mathbf{Y}^*_t] + \gamma \left\{-\frac{1}{2}\left(m(x) - [\mathbf{Y}^c_t, \mathbf{Y}^*_t]\right) + \mathbf{K}\mathbf{s}_\theta(t, x, \left[\mathbf{Y}^c_t, \tilde{\mathbf{Y}}^*_t\right])\right\} + \sqrt{\gamma}\mathrm{K}(x, x)^{1/2} Z$
- **Exact** as $\gamma \to 0$, or can correct with SMC (Trippe et al., 2022).
- **Problem**: In practice even with tiny γ tend to dismiss context $\mathcal{C}$!

Conditional neural diffusion processes: Corrector

- **Corrector**: (Multi-steps to) Target $p_t(y_t^*|y_t^c)$
  - RePaint (Lugmayr et al., 2022): denoise and re-noise $[\mathbf{Y}_t^*, \mathbf{Y}_t^c]$ to increase correlation.
  - Note that $\nabla_{y_t^*} \log p(y_t^*|y_t^c) = \nabla_{y_t^*} \log p(y_t^*, y_t^c) - \nabla_{y_t^*} \log p(y_t^c) = \nabla_{y_t^*} \log p(y_t^*, y_t^c)$
  - Langevin dynamics $\mathrm{d}\mathbf{Y}_s^* = \frac{1}{2}\mathrm{K}\nabla_{\mathbf{Y}_s^*} \log p_{T-t}(\mathbf{Y}_s^*, \mathbf{Y}_s^c)\mathrm{d}s + \sqrt{\mathrm{K}}\mathrm{d}\mathbf{B}_s$

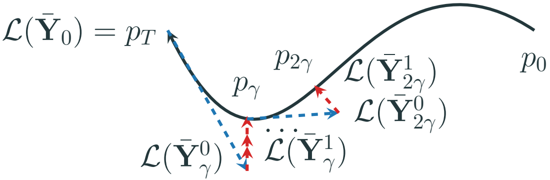


Figure 6: Illustration of Langevin corrected conditional sampling. The black line represents the noising process dynamics $(p_t)_{t\in[0,T]}$. The time reversal (i.e. predictor) step, is combined with a Langevin corrector step projecting back onto the dynamics.

Conditional neural diffusion processses: Langevin corrector

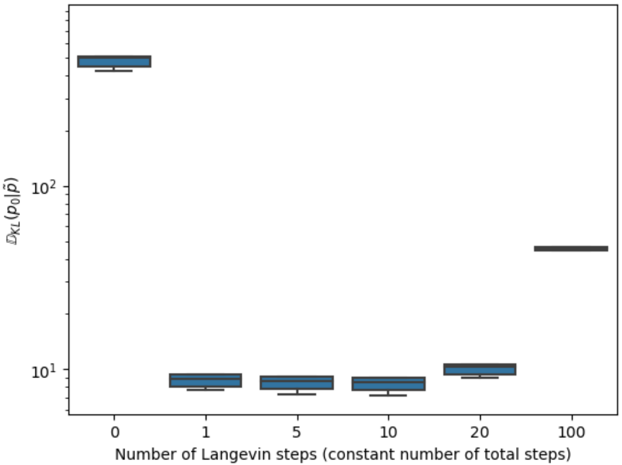
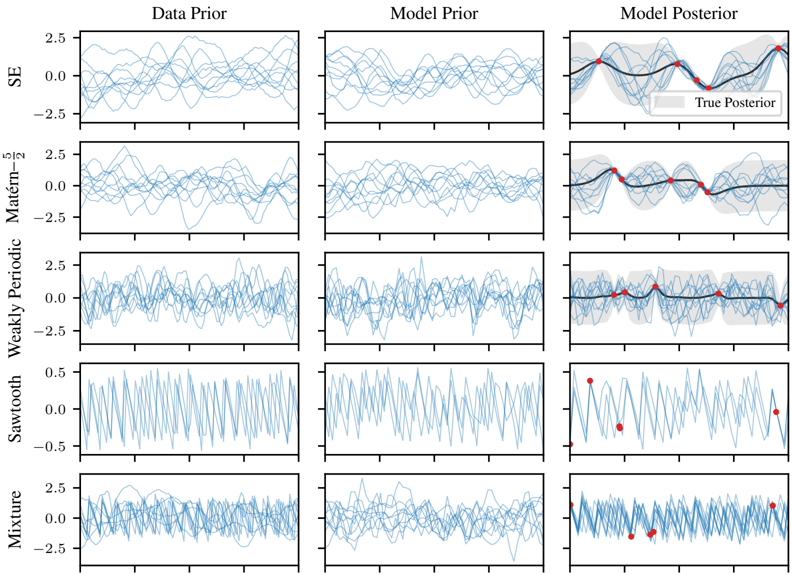


Figure 7: Ablation of number of corrector steps for conditional sampling.

Experimental results

1D regression: Datasets

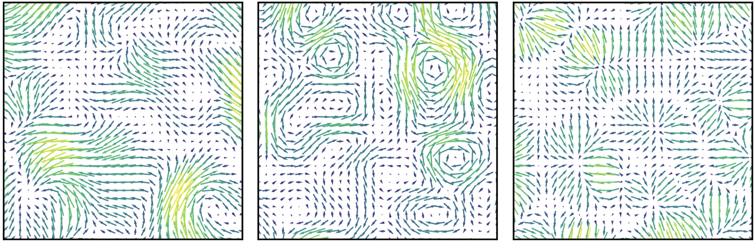


1D regression: Predictive log-likelihood (Cont'd)

Table 1: Mean test log-likelihood (TLL) (↑) ± 1 standard error estimated over 4096 test samples are reported. NP baselines from (Bruinsma et al., 2020).

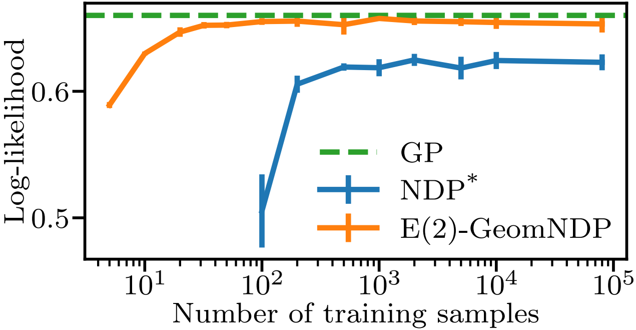
		SE	Matérn($\frac{5}{2}$)	Weakly Per.	Sawtooth	Mixture
Interpolat.	GP (optimum)	0.70±0.00	0.31±0.00	−0.32±0.00	-	-
	T(1)−GeomNDP	**0.72**±0.03	**0.32**±0.03	**−0.38**±0.03	**3.39**±0.04	**0.64**±0.08
	NDP*	**0.71**±0.03	**0.30**±0.03	**−0.37**±0.03	**3.39**±0.04	**0.64**±0.08
	GNP	**0.70**±0.01	**0.30**±0.01	−0.47±0.01	0.42±0.01	0.10±0.02
	ConvNP	−0.46±0.01	−0.67±0.01	−1.02±0.01	1.20±0.01	−0.50±0.02
Generalisat.	GP (optimum)	0.70±0.00	0.31±0.00	−0.32±0.00	-	-
	T(1)−GeomNDP	**0.70**±0.02	**0.31**±0.02	**−0.38**±0.03	**3.39**±0.03	**0.62**±0.02
	NDP*	*	*	*	*	*
	GNP	**0.69**±0.01	**0.30**±0.01	−0.47±0.01	0.42±0.01	0.10±0.02
	ConvNP	−0.46±0.01	−0.67±0.01	−1.02±0.01	1.19±0.01	−0.53±0.02

2D invariant Gaussian vector fields



Model	SE	Curl-free	Div-free
GP	$0.56_{\pm 0.00}$	$0.66_{\pm 0.00}$	$0.66_{\pm 0.00}$
NDP*	$0.55_{\pm 0.00}$	$0.62_{\pm 0.01}$	$0.62_{\pm 0.01}$
E(2)-GeomNDP	$\mathbf{0.56_{\pm 0.01}}$	$\mathbf{0.65_{\pm 0.01}}$	$\mathbf{0.66_{\pm 0.01}}$
GP (diag.)	$-1.56_{\pm 0.00}$	$-1.47_{\pm 0.00}$	$-1.47_{\pm 0.00}$
T(2)-ConvCNP	$-1.71_{\pm 0.01}$	$-1.77_{\pm 0.01}$	$-1.76_{\pm 0.00}$
E(2)-SteerCNP	$-1.61_{\pm 0.00}$	$-1.57_{\pm 0.00}$	$-1.57_{\pm 0.01}$

2D invariant Gaussian vector fields (Cont'd)



Global tropical cyclone trajectory prediction

- $f : \mathbb{R} \to \mathcal{S}^2$ with data from International Best Track Archive for Climate Stewardship (IBTrACS) (Knapp et al., 2018).
- $\mathbf{Y}_t(x) = (\mathbf{Y}_t(x_1), \cdots, \mathbf{Y}_t(x_n)) \in \mathcal{M}^n$ for any $(x_1, \ldots, x_n) \in \mathcal{X}^n$
- $\mathrm{d}\mathbf{Y}_t(x_k) = \cancelto{0}{-\frac{1}{2} \nabla U(\mathbf{Y}_t(x_k))} \; \beta_t \,\mathrm{d}t + \sqrt{\beta_t}\, \mathrm{d}\mathbf{B}_t^{\mathcal{M}} \;\; \forall k = 1, \ldots, n$ (Bortoli et al., 2022)
- $\mathcal{L}(\mathbf{Y}_t(x)) \underset{t \to \infty}{\longrightarrow} \mathrm{U}(\mathcal{S}^2)^{\otimes n}$.

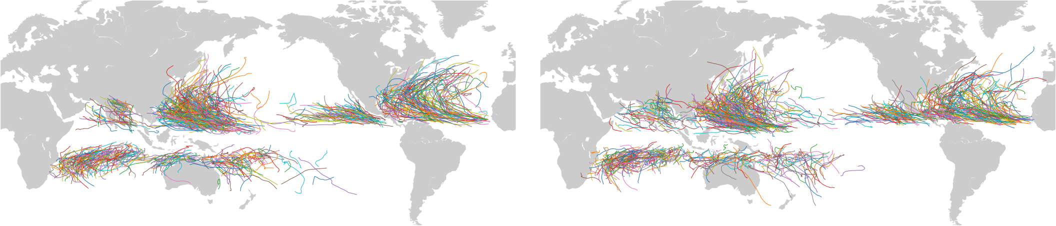
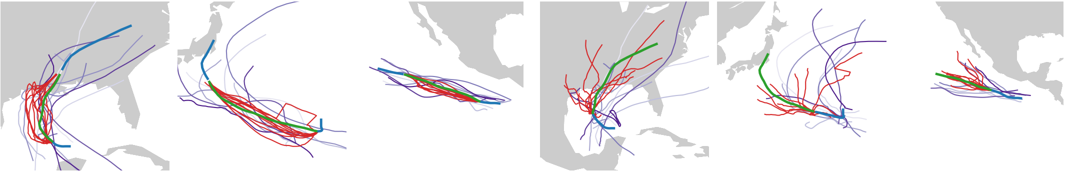


Figure 8: *Left:* 1000 samples from the training data. *Right:* 1000 samples from trained model.

Global tropical cyclone trajectory prediction (Cont'd)



(a) Interpolation **(b)** Extrapolation

Model	TEST DATA	INTERPOLATION		EXTRAPOLATION	
	Likelihood	Likelihood	MSE (km)	Likelihood	MSE (km)
GEOMNDP$(\mathbb{R} \to \mathcal{S}^2)$	$\mathbf{802}_{\pm 5}$	$\mathbf{535}_{\pm 4}$	$\mathbf{162}_{\pm 6}$	$\mathbf{536}_{\pm 4}$	$\mathbf{496}_{\pm 14}$
STEREO GP $(\mathbb{R} \to \mathbb{R}^2/\{0\})$	$393_{\pm 3}$	$266_{\pm 3}$	$2619_{\pm 13}$	$245_{\pm 2}$	$6587_{\pm 55}$
NDP $(\mathbb{R} \to \mathbb{R}^2)$	-	-	$166_{\pm 22}$	-	$769_{\pm 48}$
GP $(\mathbb{R} \to \mathbb{R}^2)$	-	-	$6852_{\pm 41}$	-	$8138_{\pm 87}$

Recap: Geometric diffusion neural processes

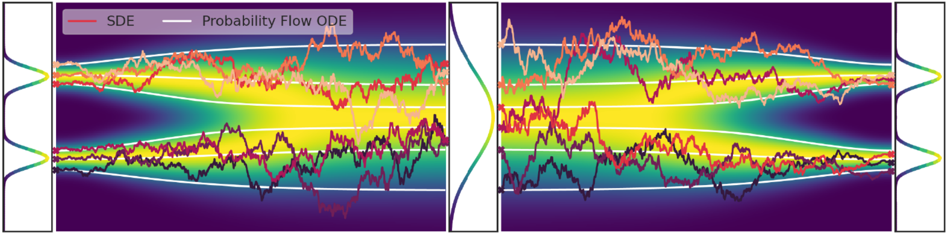
- Constructed diffusion models over function space by correlating finite marginals
- Incorporating group invariance by
  - targetting invariant Gaussian processes and
  - parameterising the score with an equivariant neural network
- Sampling from the conditional process with Langevin corrector
- Empirically demonstrated modelling capacity on scalar and vector fields, with Euclidean and spherical output space

Thanks for your attention!

Questions? Remarks?

Appendix

1D regression: Predictive log-likelihood



We can derive a deterministic process which has the same marginal density as the noising SDE (6), which given by the following ODE

$$\mathrm{d}\tilde{\mathbf{Y}}_t(x) = \{\tfrac{1}{2}(m(x) - \bar{\mathbf{Y}}_t(x)) - \tfrac{1}{2}\mathrm{K}(x,x)\nabla \log p_{T-t}(\bar{\mathbf{Y}}_t(x))\}\beta_t \,\mathrm{d}t \triangleq f_{\mathrm{ODE}}(t, x, \mathbf{Y}_t(x))\,\mathrm{d}t$$

$$\mathrm{d}\begin{pmatrix} \tilde{\mathbf{Y}}_t(x) \\ \log p_t(\tilde{\mathbf{Y}}_t(x)) \end{pmatrix} = \begin{pmatrix} f_{\mathrm{ODE}}(t, x, \mathbf{Y}_t(x)) \\ -\tfrac{1}{2}\mathrm{div}\; f_{\mathrm{ODE}}(t, x, \mathbf{Y}_t(x)) \end{pmatrix} \mathrm{d}t. \tag{11}$$

References

V. D. Bortoli, E. Mathieu, M. J. Hutchinson, J. Thornton, Y. W. Teh, and A. Doucet. Riemannian Score-Based Generative Modelling. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022. Cited on page 50.

W. Bruinsma, J. Requeima, A. Y. K. Foong, J. Gordon, and R. E. Turner. Gaussian Neural Processes. In *3rd Symposium on Advances in Approximate Bayesian Inference*, Feb. 2020. Cited on page 47.

P. Cattiaux, G. Conforti, I. Gentil, and C. Léonard. Time reversal of diffusion processes under a finite entropy condition. *arXiv preprint arXiv:2104.07708*, 2021. Cited on pages 11, 12, 30.

A. Durmus. *High Dimensional Markov Chain Monte Carlo Methods: Theory, Methods and Application*. PhD thesis, Paris-Sud XI, 2016. Cited on pages 9, 10.

U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(4):1188–1205, 1986. Cited on pages 11, 12, 30.

P. Holderrieth, M. J. Hutchinson, and Y. W. Teh. Equivariant Learning of Stochastic Fields: Gaussian Processes and Steerable Conditional Neural Processes. In *International Conference on Machine Learning*, 2021. Cited on pages 36, 37.

H. J. Knapp Kenneth R. Diamond, J. P. Kossin, M. C. Kruk, and C. J. I. Schreck. International Best Track Archive for Climate Stewardship (IBTrACS) Project, Version 4. Technical report, NOAA National Centers for Environmental Information, 2018. DOI: https://doi.org/10.25921/82ty-9e16. Cited on page 50.

A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11461–11471, 2022. Cited on page 43.

I. Macêdo and R. Castro. *Learning Divergence-Free and Curl-Free Vector Fields with Matrix-Valued Kernels*. Pré-Publicações / A.: Pré-publicações. IMPA, 2010. Cited on page 37.

A. Phillips, T. Seror, M. Hutchinson, V. De Bortoli, A. Doucet, and E. Mathieu. Spectral Diffusion Processes. Nov. 2022. URL: http://arxiv.org/abs/2209.14125. Cited on pages 24–26.

S. Ravuri, K. Lenc, M. Willson, D. Kangin, R. Lam, P. Mirowski, M. Fitzsimons, M. Athanassiadou, S. Kashem, S. Madge, R. Prudden, A. Mandhane, A. Clark, A. Brock, K. Simonyan, R. Hadsell, N. Robinson, E. Clancy, A. Arribas, and S. Mohamed. Skilful Precipitation Nowcasting Using Deep Generative Models of

Radar. *Nature*, 597(7878):672–677, Sept. 2021. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-021-03854-z. Cited on page 5.

L. Scott and J. Serre. *Linear Representations of Finite Groups*. Graduate Texts in Mathematics. Springer New York, 1996. ISBN: 978-0-387-90190-9. URL: https://books.google.co.uk/books?id=NCfZgr54TJ4C. Cited on pages 20–22.

B. L. Trippe, J. Yim, D. Tischer, T. Broderick, D. Baker, R. Barzilay, and T. Jaakkola. Diffusion Probabilistic Modeling of Protein Backbones in 3D for the Motif-Scaffolding Problem. June 2022. DOI: 10.48550/arXiv.2206.04119. Cited on page 42.

T. Vincent, L. Risser, and P. Ciuciu. Spatially Adaptive Mixture Modeling for Analysis of fMRI Time Series. *IEEE Trans. Med. Imag.*, 29(4):1059–1074, Apr. 2010. ISSN: 0278-0062. DOI: 10.1109/TMI.2010.2042064. Cited on pages 31, 32.

M. Xu, L. Yu, Y. Song, C. Shi, S. Ermon, and J. Tang. GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation. In *International Conference on Learning Representations*, 2022. URL: https://openreview.net/forum?id=PzcvxEMzvQC. Cited on page 4.

J. Yim, B. L. Trippe, V. De Bortoli, E. Mathieu, A. Doucet, R. Barzilay, and T. Jaakkola. SE (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023. Cited on page 38.