NEURAL INFORMATION
PROCESSING SYSTEMS

Scan QR code for project page

Manuscript Location: https://arxiv.org/abs/2510.25769

Neural Stochastic Flows: Solver-Free Modelling and Inference for SDE Solutions

Naoki Kiyohara[1,2], Edward Johns[1], Yingzhen Li[1]

[1]Imperial College London [2]Canon Inc.

IMPERIAL Canon

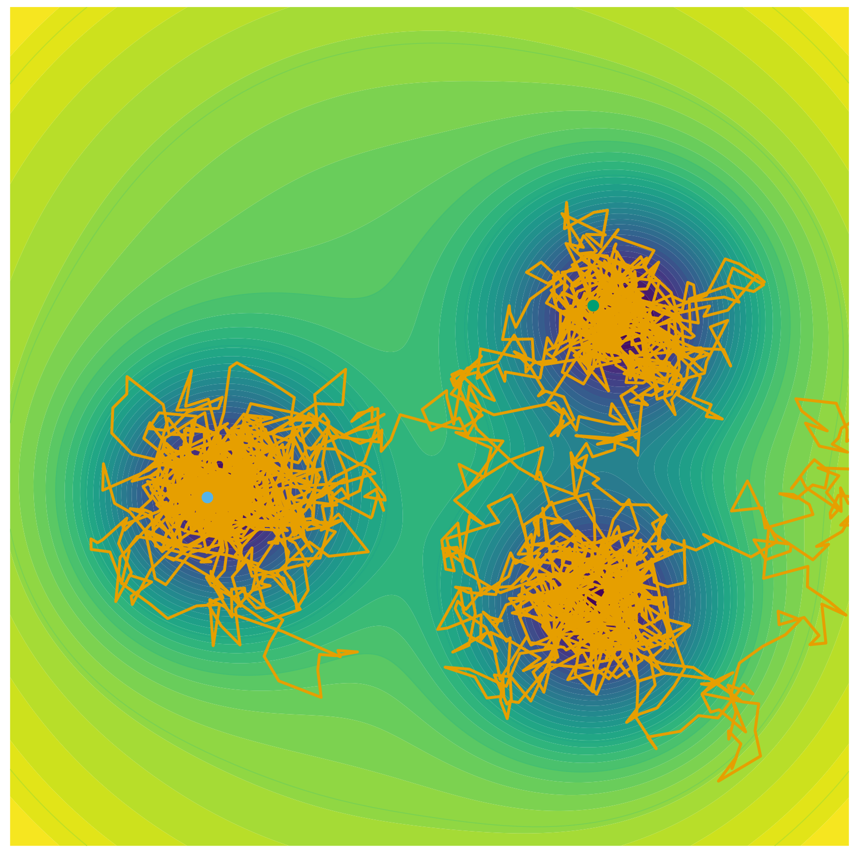
Stochastic Differential Equations (SDEs) are Everywhere

$$\text{Itô SDE:} \quad dX_t = \underbrace{\mu(X_t, t)\,dt}_{\text{Drift term}} + \underbrace{\sigma(X_t, t)\,dW_t}_{\text{Diffusion term}}, \quad X_0 \sim p_0$$

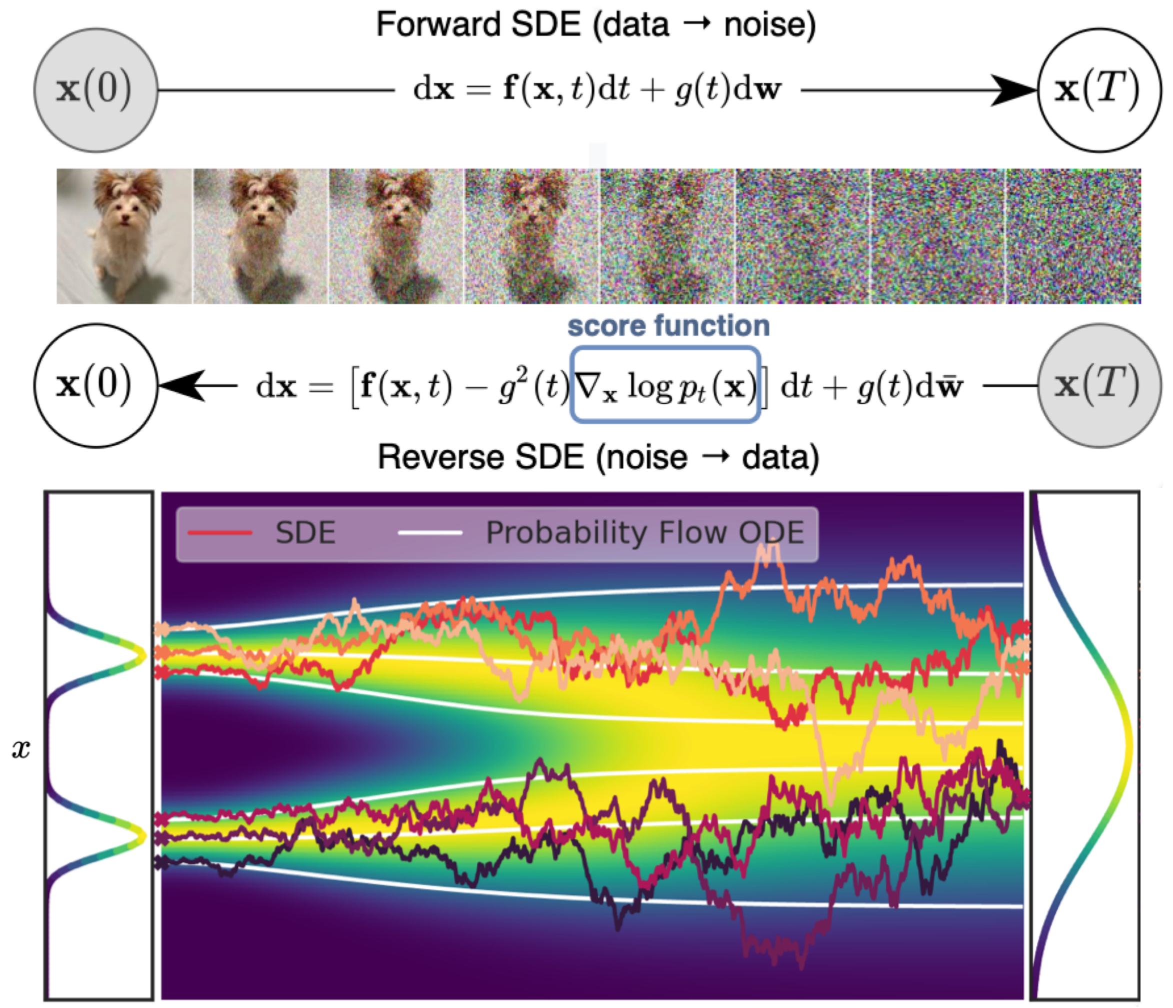
Drift term   Diffusion term

Finance

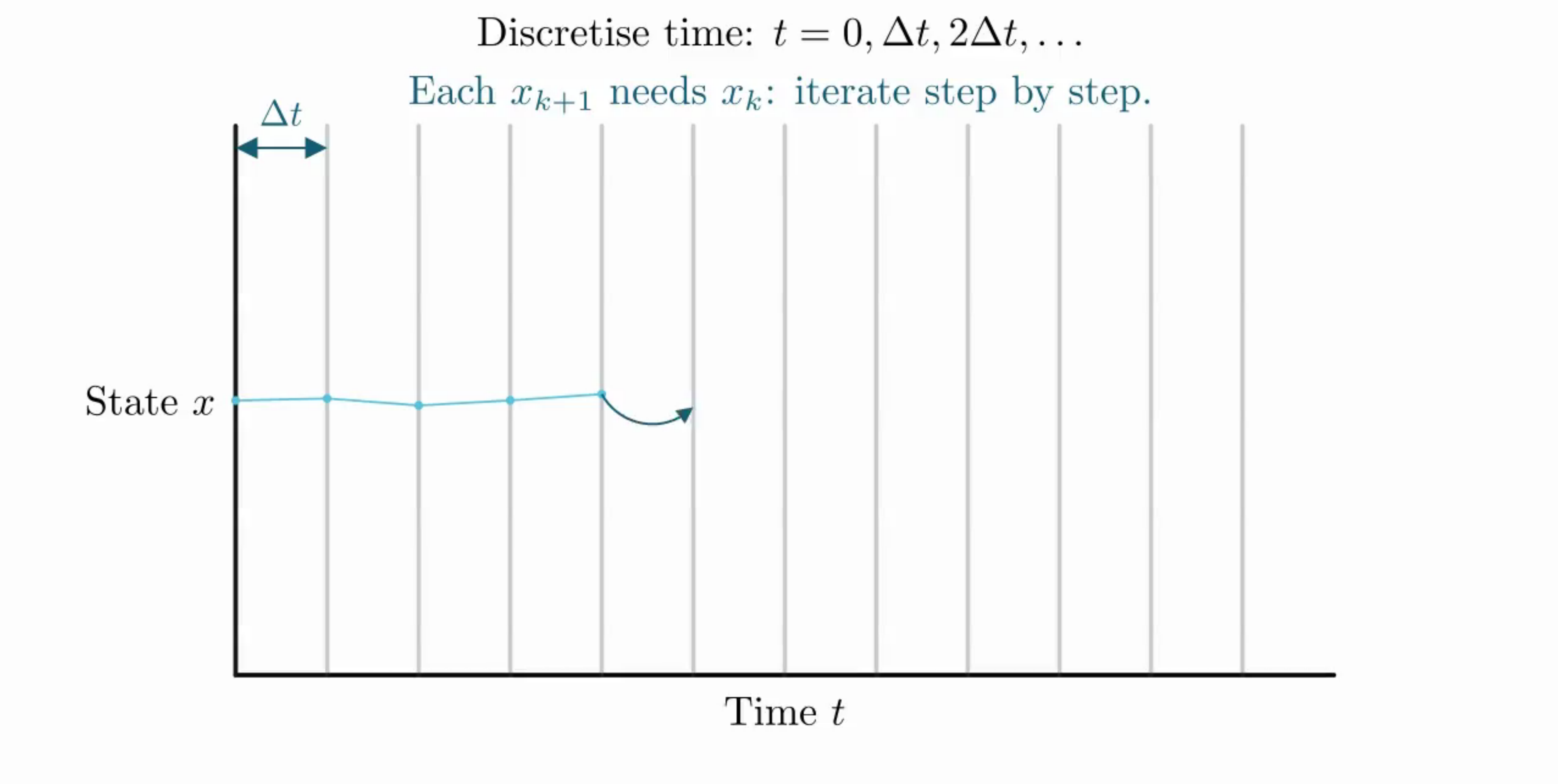
Physics



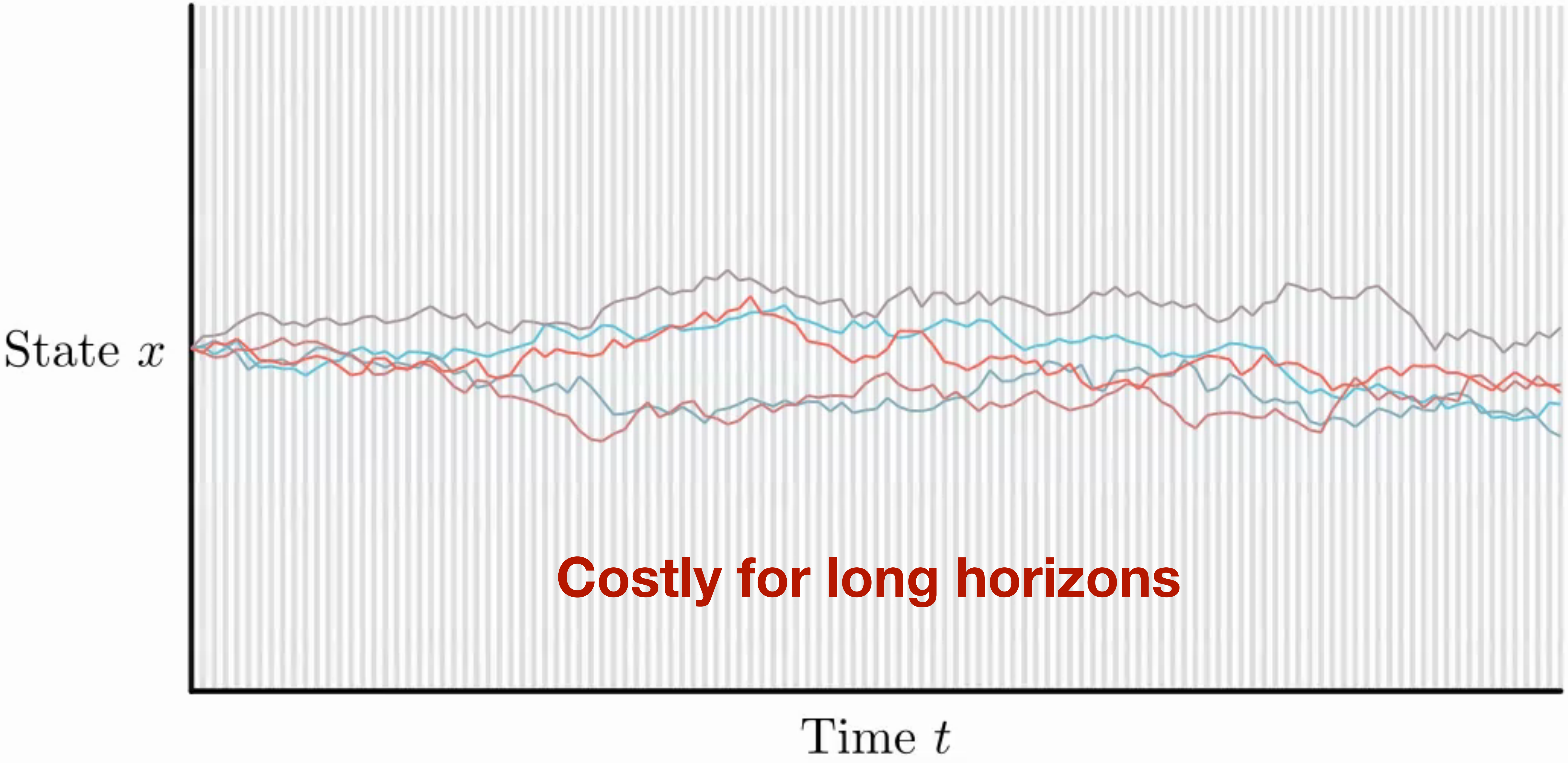
Generative Modelling



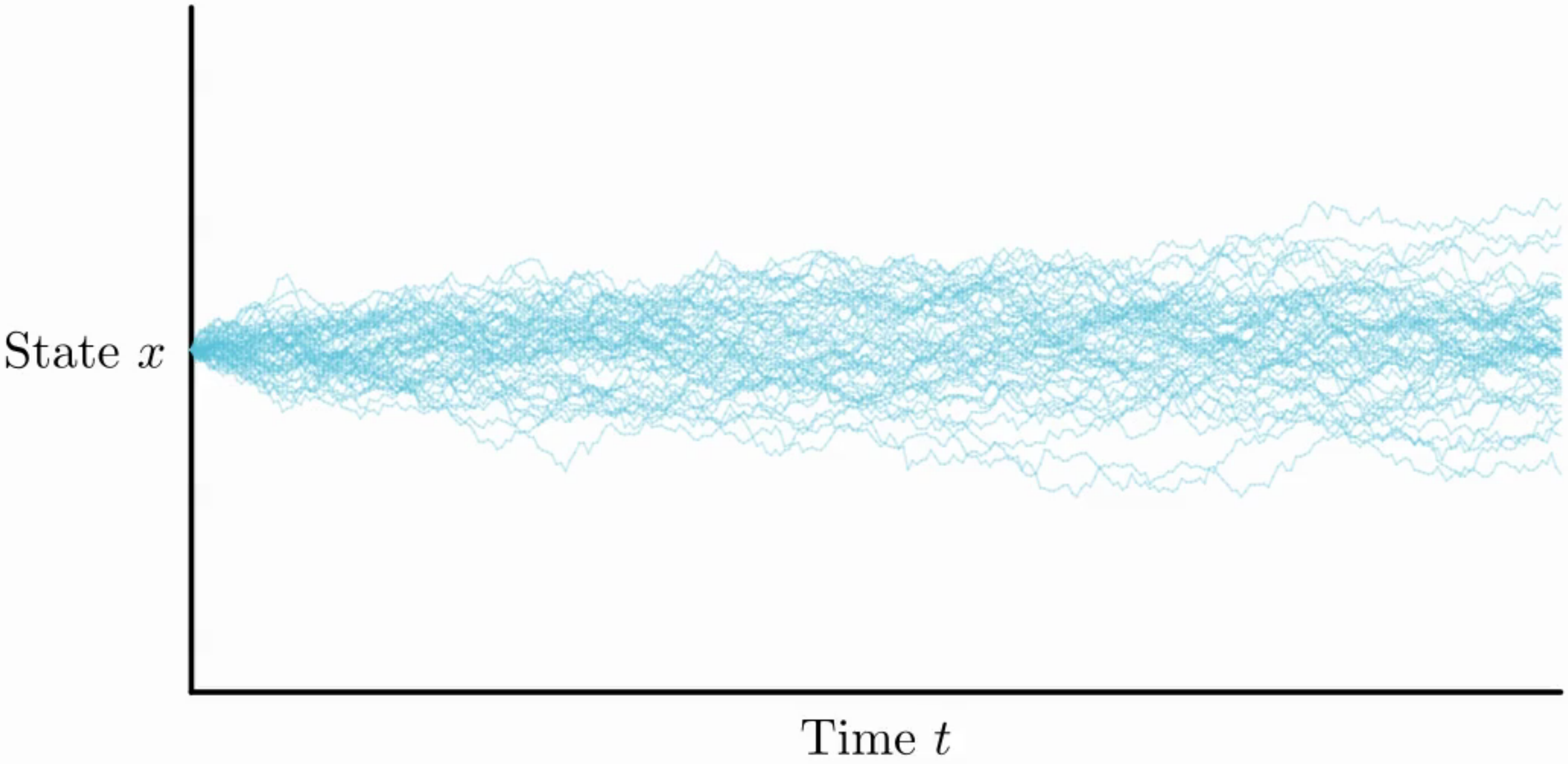
Computational Challenges of Neural SDEs



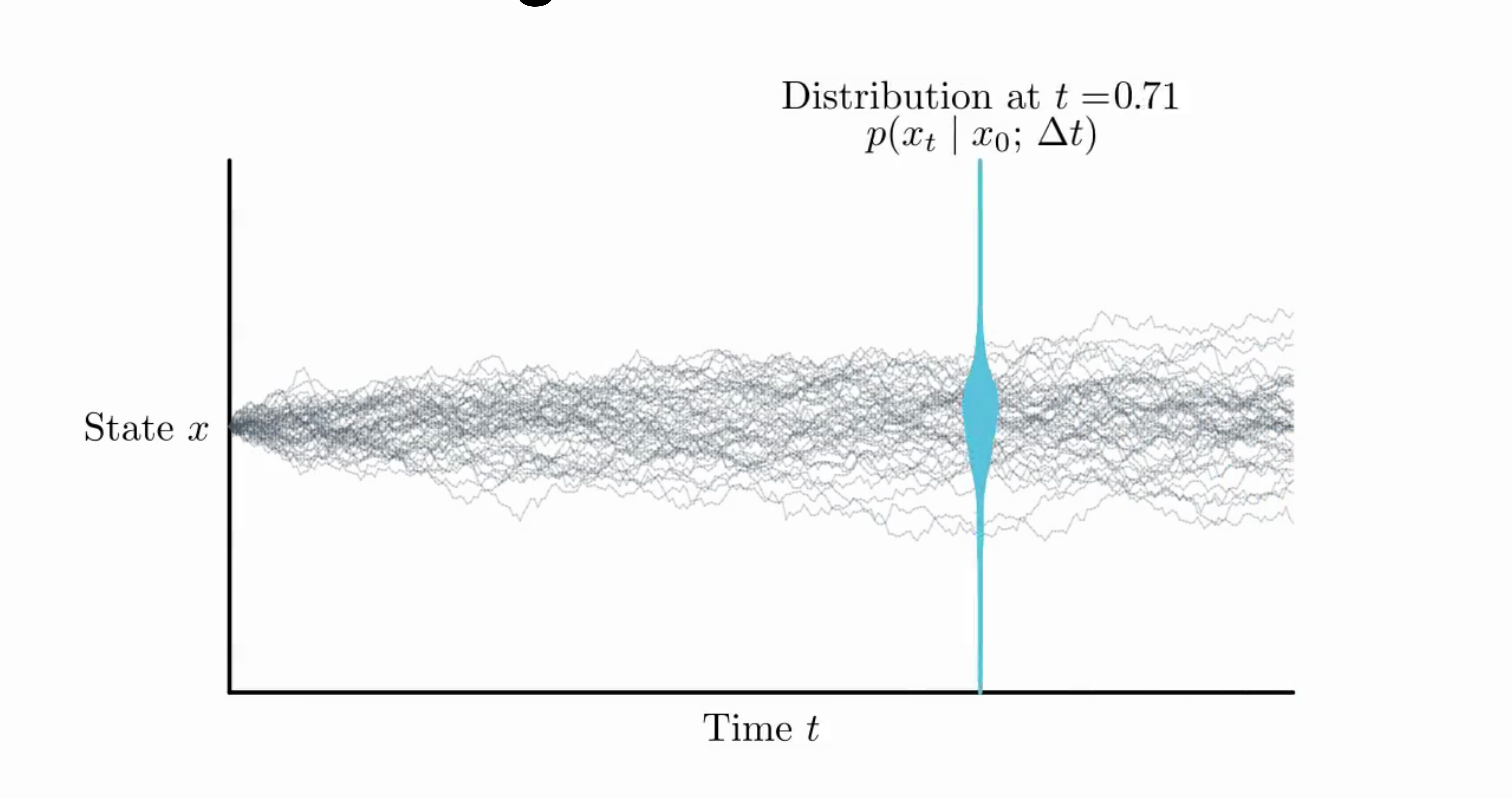
Computational Challenges of Neural SDEs



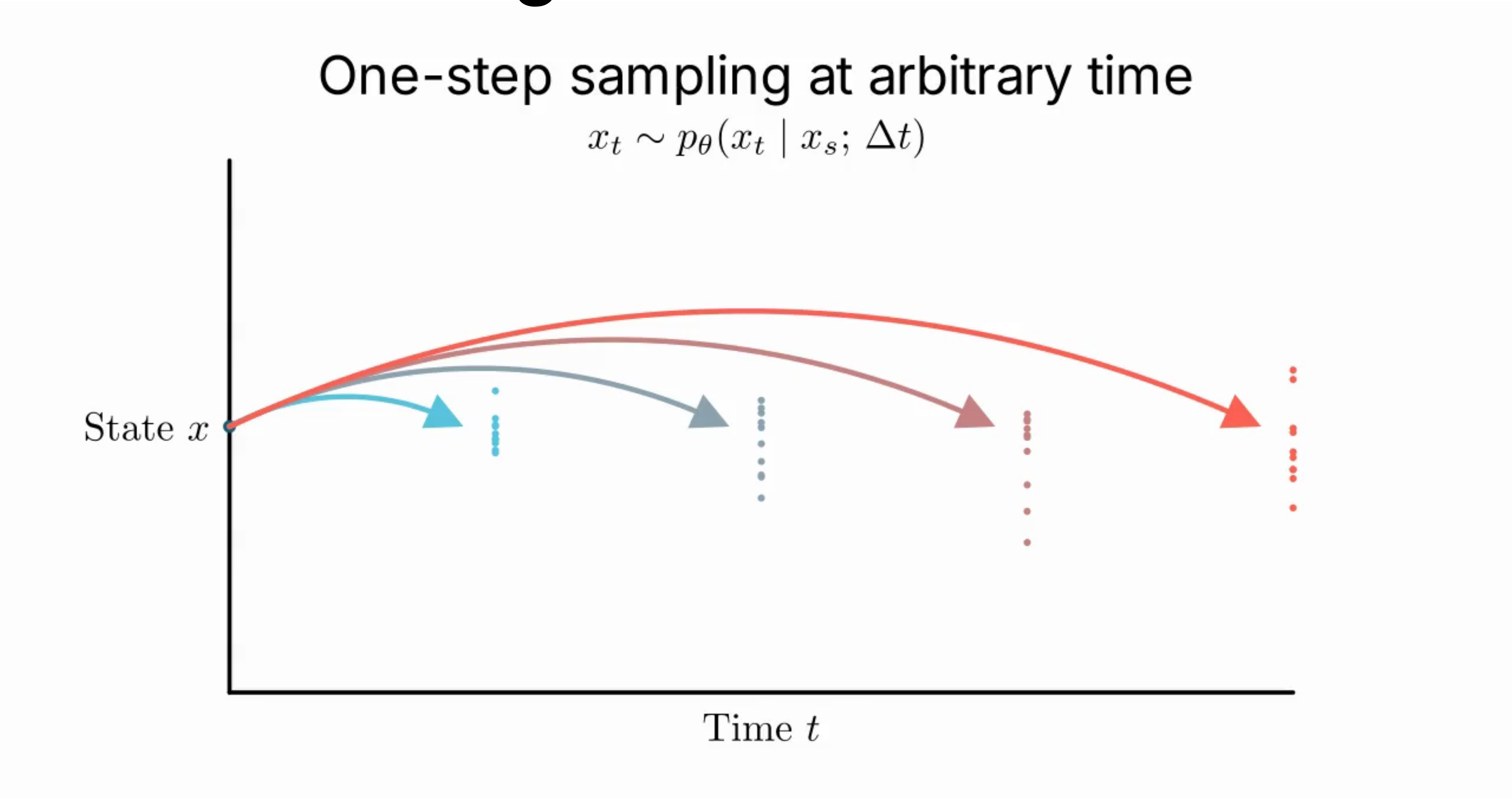
Direct Modelling of Transition Distributions



Direct Modelling of Transition Distributions



Direct Modelling of Transition Distributions



Mathematical Requirements for Weak Solutions

Properties of transition kernel $p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_s; s, \Delta t = t - s)$

1. **Independence.**

   For any $0 \leq t_1 \leq \cdots \leq t_n$, $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{k+1}} \mid \boldsymbol{x}_{t_k}; t_k, t_{k+1} - t_k)$ are independent.

2. **Identity.**

   When $t = s$, $p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_s; s, 0) = \delta(\boldsymbol{x}_t - \boldsymbol{x}_s)$

3. **Flow property.**

   For any $0 \leq t_i \leq t_j \leq t_k$, $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_k} \mid \boldsymbol{x}_{t_i}; t_i, t_k - t_i) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_k} \mid \boldsymbol{x}_{t_j}; t_j, t_k - t_j)\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_j} \mid \boldsymbol{x}_{t_i}; t_i, t_j - t_i)\, \mathrm{d}\boldsymbol{x}_{t_j}$

4. **Stationarity (for autonomous SDEs).**

   $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_j} \mid \boldsymbol{x}_s; t_i, t_j - t_i) = p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_j+r} \mid \boldsymbol{x}_s; t_i + r, t_j - t_i)$

Mathematical Requirements for Weak Solutions

Properties of transition kernel $p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_s; s, \Delta t = t - s)$

1. **Independence.** **✓ Realised by non-overlapping sampling.**

   For any $0 \leq t_1 \leq \cdots \leq t_n$, $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_{k+1}} \mid \boldsymbol{x}_{t_k}; t_k, t_{k+1} - t_k)$ are independent.

2. **Identity.** **✓ Realised by architecture.**

   When $t = s$, $p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_s; s, 0) = \delta(\boldsymbol{x}_t - \boldsymbol{x}_s)$

3. **Flow property.** **✓ Softly enforced by regularisation (flow loss).**

   For any $0 \leq t_i \leq t_j \leq t_k$, $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_k} \mid \boldsymbol{x}_{t_i}; t_i, t_k - t_i) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_k} \mid \boldsymbol{x}_{t_j}; t_j, t_k - t_j)\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_j} \mid \boldsymbol{x}_{t_i}; t_i, t_j - t_i)\, \mathrm{d}\boldsymbol{x}_{t_j}$

4. **Stationarity (for autonomous SDEs).** **✓ Realised by architecture optionally.**

   $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_j} \mid \boldsymbol{x}_s; t_i, t_j - t_i) = p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_j + r} \mid \boldsymbol{x}_s; t_i + r, t_j - t_i)$

Architectural Design

Conditional normalising flow for $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_j} \mid \boldsymbol{x}_{t_i}; t_i, \Delta t)$

Base distribution

Using conditioning $\boldsymbol{c} := (\boldsymbol{x}_{t_i}, \Delta t, t_i)$,

$$\boldsymbol{z} = \boldsymbol{x}_{t_i} + \Delta t \cdot \mathrm{MLP}_{\mu}(\boldsymbol{c}; \boldsymbol{\theta}_{\mu}) + \sqrt{\Delta t} \cdot \mathrm{MLP}_{\sigma}(\boldsymbol{c}; \boldsymbol{\theta}_{\sigma}) \odot \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

Conditional affine coupling layers [4, 13, 28]

$$\boldsymbol{x}_{t_j} = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{c}) = \boldsymbol{f}_L(\cdot; \boldsymbol{c}, \boldsymbol{\theta}_L) \circ \boldsymbol{f}_{L-1}(\cdot; \boldsymbol{c}, \boldsymbol{\theta}_{L-1}) \circ \cdots \circ \boldsymbol{f}_1(\boldsymbol{z}; \boldsymbol{c}, \boldsymbol{\theta}_1)$$

With $\boldsymbol{z}_{\mathrm{A}}, \boldsymbol{z}_{\mathrm{B}} = \mathrm{Split}(\boldsymbol{z})$, each layer is designed as:

$$\mathrm{Concat}\Big(\boldsymbol{z}_{\mathrm{A}}, \boldsymbol{z}_{\mathrm{B}} \odot \exp\big(\Delta t\, \mathrm{MLP}^{(i)}_{\mathrm{scale}}(\boldsymbol{z}_{\mathrm{A}}, \boldsymbol{c}; \boldsymbol{\theta}^{(i)}_{\mathrm{scale}})\big) + \Delta t\, \mathrm{MLP}^{(i)}_{\mathrm{shift}}(\boldsymbol{z}_{\mathrm{A}}, \boldsymbol{c}; \boldsymbol{\theta}^{(i)}_{\mathrm{shift}})\Big)$$

Architectural Design

Conditional normalising flow for $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_j} \mid \boldsymbol{x}_{t_i}; t_i, \Delta t)$

Base distribution

Using conditioning $\boldsymbol{c} := (\boldsymbol{x}_{t_i}, \Delta t, t_i)$,

$$\boldsymbol{z} = \boldsymbol{x}_{t_i} + \Delta t \cdot \mathrm{MLP}_{\mu}(\boldsymbol{c}; \boldsymbol{\theta}_{\mu}) + \sqrt{\Delta t} \cdot \mathrm{MLP}_{\sigma}(\boldsymbol{c}; \boldsymbol{\theta}_{\sigma}) \odot \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

Conditional affine coupling layers [4, 13, 28]

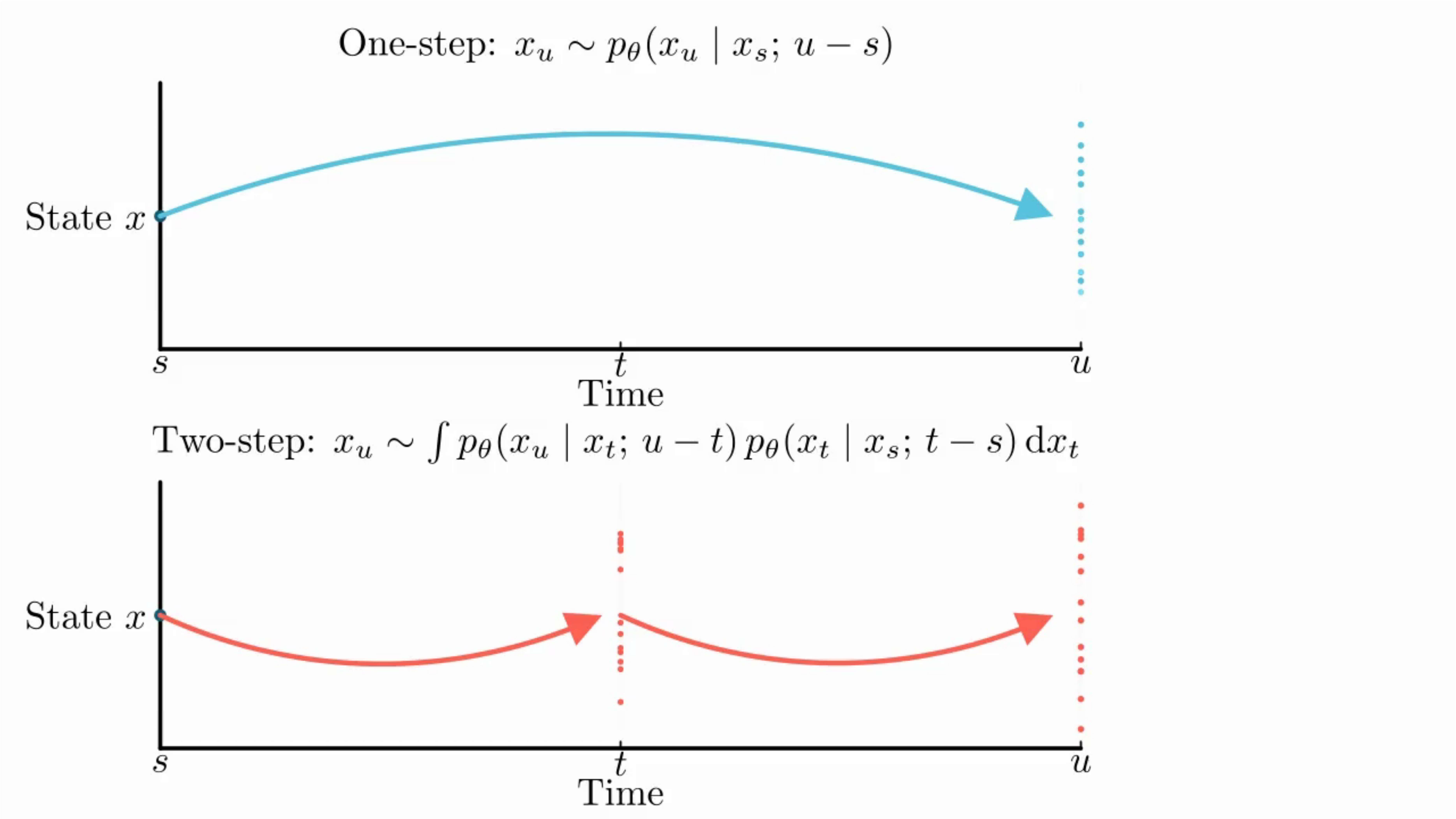
$$\boldsymbol{x}_{t_j} = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{c}) = \boldsymbol{f}_L(\cdot; \boldsymbol{c}, \boldsymbol{\theta}_L) \circ \boldsymbol{f}_{L-1}(\cdot; \boldsymbol{c}, \boldsymbol{\theta}_{L-1}) \circ \cdots \circ \boldsymbol{f}_1(\boldsymbol{z}; \boldsymbol{c}, \boldsymbol{\theta}_1)$$

With $\boldsymbol{z}_{\mathrm{A}}, \boldsymbol{z}_{\mathrm{B}} = \mathrm{Split}(\boldsymbol{z})$, each layer is designed as:

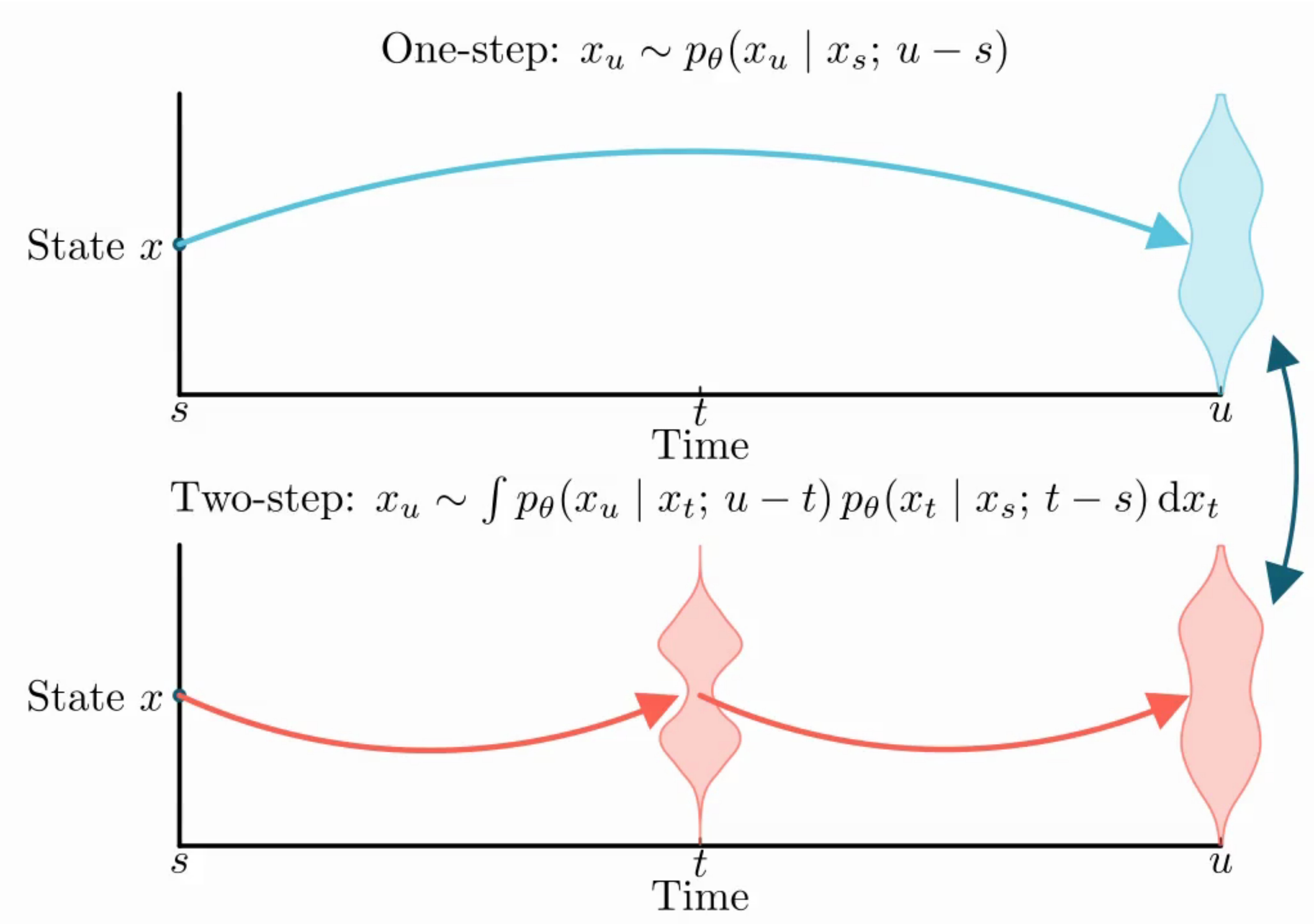
$$\mathrm{Concat}\Big(\boldsymbol{z}_{\mathrm{A}}, \boldsymbol{z}_{\mathrm{B}} \odot \exp\big(\Delta t\, \mathrm{MLP}^{(i)}_{\mathrm{scale}}(\boldsymbol{z}_{\mathrm{A}}, \boldsymbol{c}; \boldsymbol{\theta}^{(i)}_{\mathrm{scale}})\big) + \Delta t\, \mathrm{MLP}^{(i)}_{\mathrm{shift}}(\boldsymbol{z}_{\mathrm{A}}, \boldsymbol{c}; \boldsymbol{\theta}^{(i)}_{\mathrm{shift}})\Big)$$

✓ Identity when $\Delta t = 0$. **✓ Stationary when we drop t_i from c.**

Regularisation for Flow Consistency



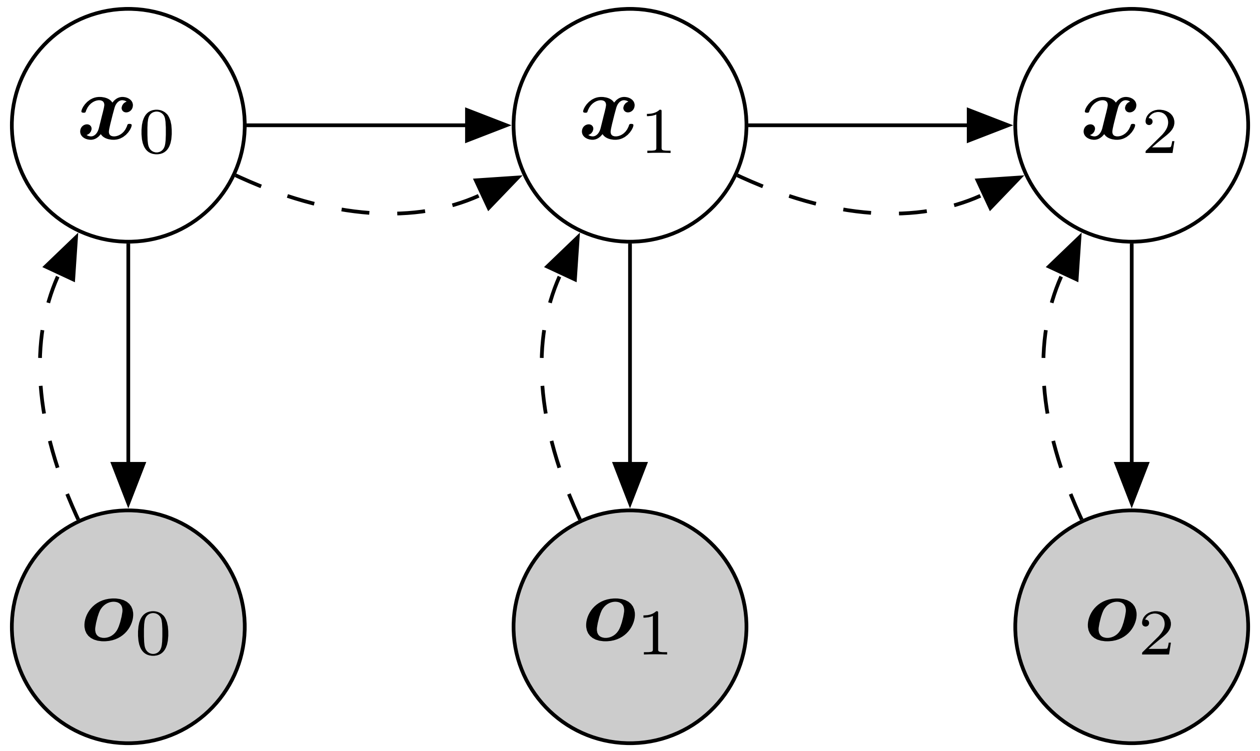
Regularisation for Flow Consistency



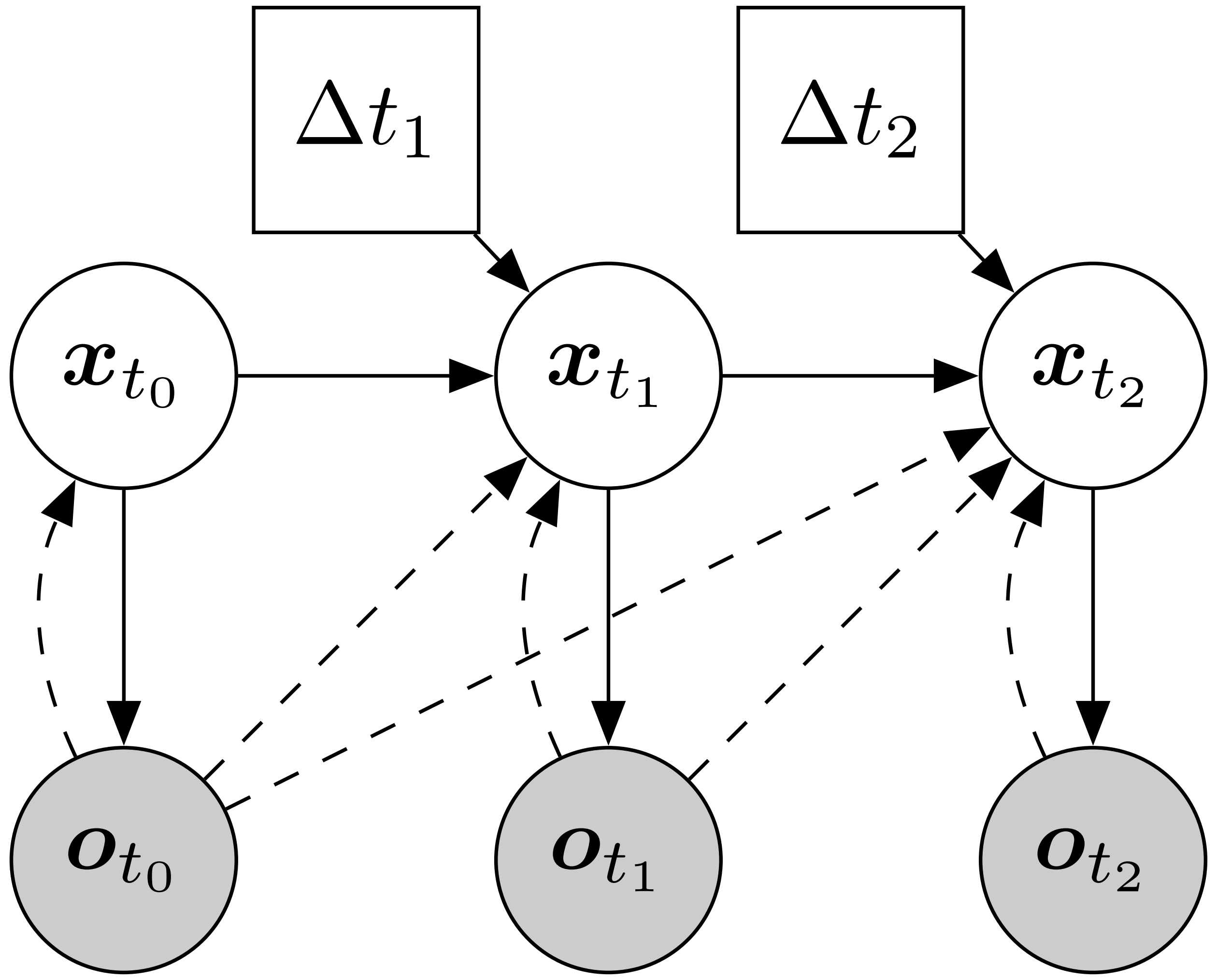
Match these distribution by upper bound of KL divergence

Latent Neural Stochastic Flows

Variational State Space Model [8, 21, 22, 31]



Latent NSF (ours)



$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\beta\text{-NELBO}} + \lambda \mathcal{L}_{\text{flow}} + \beta_{\text{skip}} \mathcal{L}_{\text{skip}},$$

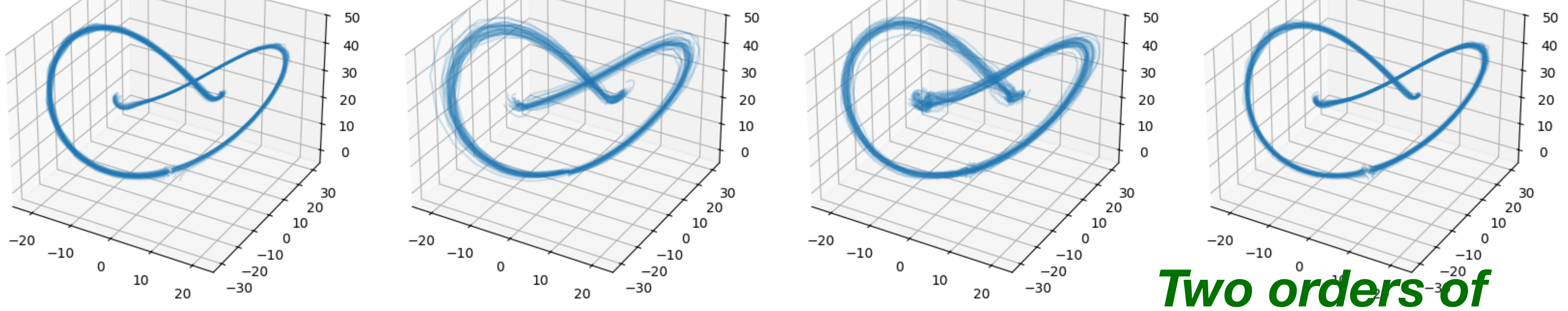
$$\mathcal{L}_{\beta\text{-NELBO}} = \sum_{i=0}^{T} \Big[- \mathop{\mathbb{E}}_{q_{\boldsymbol{\phi}}} \big[\log p_{\boldsymbol{\psi}}(\boldsymbol{o}_{t_i} \mid \boldsymbol{x}_{t_i}) \big] + \beta \mathrm{KL}\big(q_{\boldsymbol{\phi}}(\boldsymbol{x}_{t_i} \mid \boldsymbol{o}_{\le t_i}) \,\|\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_i} \mid \boldsymbol{x}_{t_{i-1}}) \big) \Big],$$

$$\mathcal{L}_{\text{skip}} = \sum_{i=0}^{T-\tau} \mathop{\mathbb{E}}_{j \sim \mathcal{U}\{i+2,\, i+\tau\}} \Big[\mathrm{KL}\big(q_{\boldsymbol{\phi}}(\boldsymbol{x}_{t_j} \mid \boldsymbol{o}_{\le t_j}) \,\|\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_j} \mid \boldsymbol{x}_{t_i}) \big) \Big].$$

Experimental Results

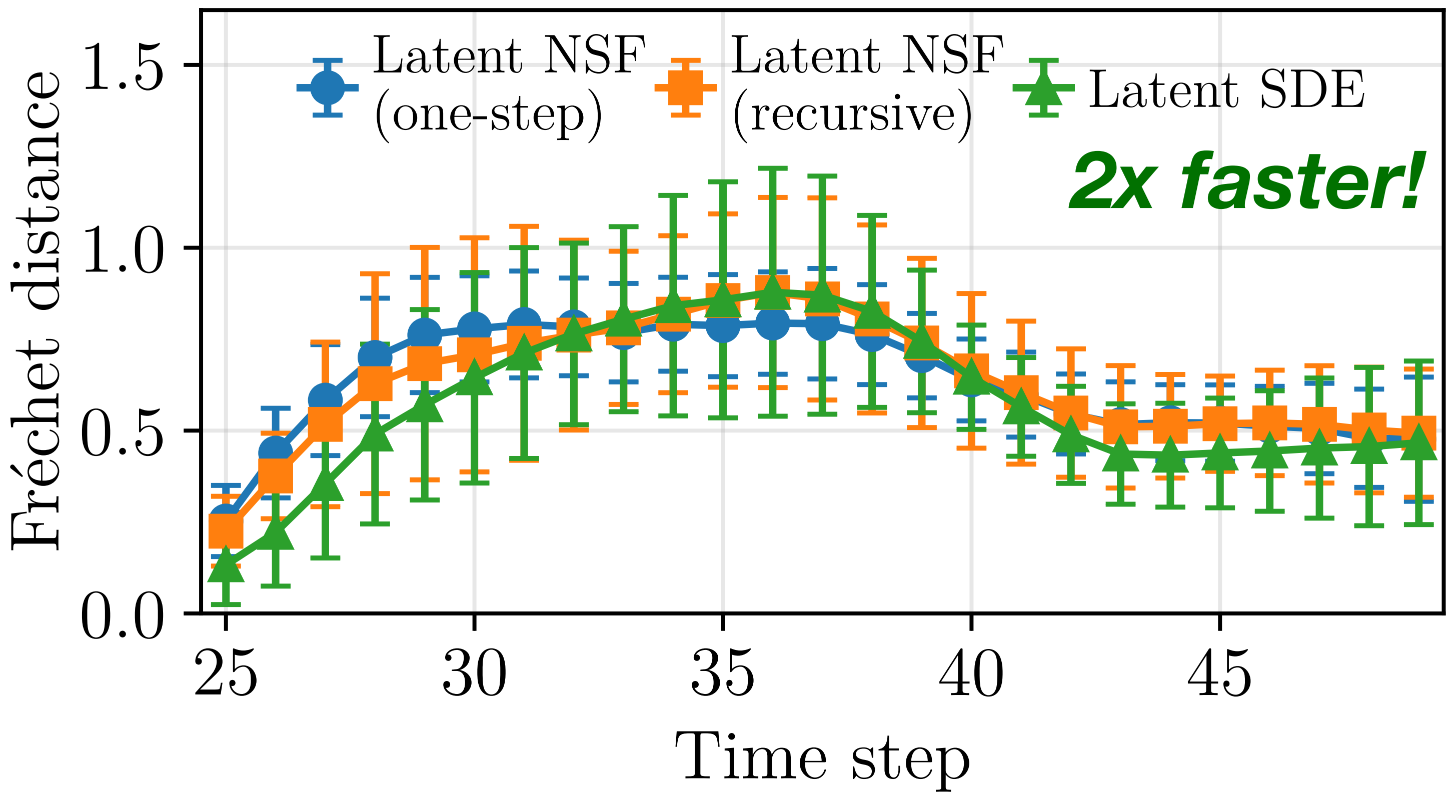
Stochastic Lorenz Attractor

Ground truth | Latent SDE [33] | SDE matching [3] | NSF (ours)



Two orders of magnitude faster!

Stochastic Moving MNIST (Latent NSF)



CMU Motion Capture (Latent NSF)

Methods	Setup 1	Setup 2
npODE [17]	$22.96^{\dagger}$	–
Neural ODE [6]	$22.49 \pm 0.88^{\dagger}$	–
ODE2VAE-KL [54]	$8.09 \pm 1.95^{\dagger}$	–
Latent ODE [43]	$5.98 \pm 0.28^{*}$	$31.62 \pm 0.05^{\S}$
Latent SDE [33]	12.91 ± 2.90^{1}	$9.52 \pm 0.21^{\S}$
Latent Approx SDE [46]	$7.55 \pm 0.05^{\S}$	10.50 ± 0.86
ARCTA [9]	$7.62 \pm 0.93^{\ddagger}$	9.92 ± 1.82
NCDSSM [2]	$5.69 \pm 0.01^{\S}$	$4.74 \pm 0.01^{\S}$
SDE matching [3]	$\mathbf{5.20 \pm 0.43}^{2}$	4.26 ± 0.35
Latent NSF (ours)	8.62 ± 0.32	$\mathbf{3.41 \pm 0.27}$

Two orders of magnitude faster!

Summary

- **We proposed Neural Stochastic Flows, which enable solver-free in both training and inference, with closed-form log-densities.**
- **Extended to Latent NSF: a variational state-space model using NSF as the latent transition for partially-observed/high-dimensional data.**

Key Benefits:

- **One-step prediction** for arbitrary time gaps.
- **Speed**: up to two orders of magnitude faster.
- **Stable training** with tractable densities.

Limitations:

- **Chapman–Kolmogorov relation is enforced approximately.**
- Affine-coupling design constrains **expressivity**.

Project page