A dynamical systems perspective on measure transport and generative modeling

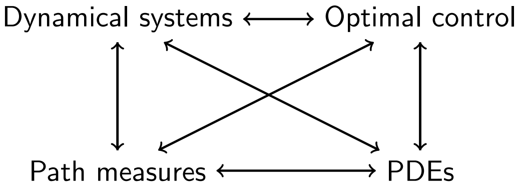
Lorenz Richter

Machine-Learning-Based Sampling in Lattice Field Theory and Quantum Chemistry
Bethe Center for Theoretical Physics, Bonn

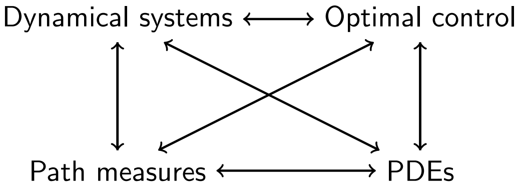
October, 2024

Overview

- Sampling via measure transport can be seen from different perspectives:



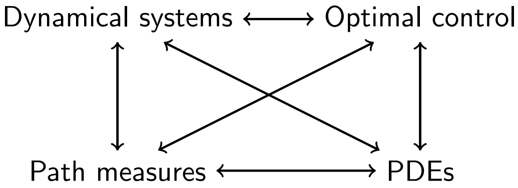
- Sampling via measure transport can be seen from different perspectives:



- The different perspectives will eventually allow us to propose new numerical algorithms.

Overview

- Sampling via measure transport can be seen from different perspectives:



- The different perspectives will eventually allow us to propose new numerical algorithms.
- This is joint work with Julius Berner (Caltech), Jingtong Sun (Caltech), Denis Blessing (KIT) and Nikolas Nüsken (King's College).

Generative modeling

Task

Sample from a complex (high-dimensional, multimodal) distribution $\mathcal{D}$.

Generative modeling

Task

Sample from a complex (high-dimensional, multimodal) distribution $\mathcal{D}$.

$\mathcal{D}$ can be given in the form of:

1. **samples** $X^{(i)} \sim \mathcal{D}$ (images, text, audio, ...).

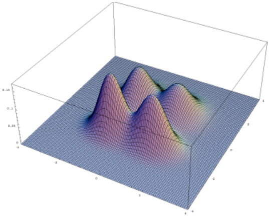
Task

Sample from a complex (high-dimensional, multimodal) distribution $\mathcal{D}$.

$\mathcal{D}$ can be given in the form of:

1. **samples** $X^{(i)} \sim \mathcal{D}$ (images, text, audio, ...).

2. an (unnormalized) **density** (e.g., in Bayesian statistics, computational physics and chemistry).



Generative modeling

- Impressive results for the first case:

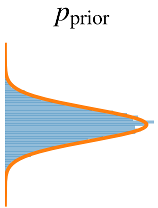
- Impressive results for the first case:

- The second case is a focus of (our) current research.

Sampling via dynamical systems

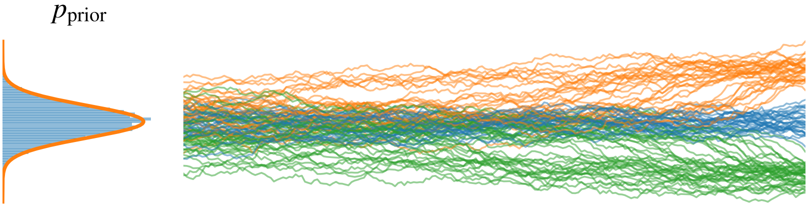
Goal: We want to sample from distribution $p_{\mathrm{target}} = \rho / \mathcal{Z}$.

Goal: We want to sample from distribution $p_{\text{target}} = \rho/\mathcal{Z}$.

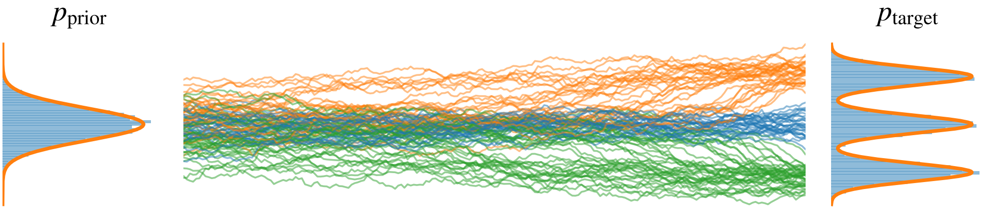


Sampling via dynamical systems

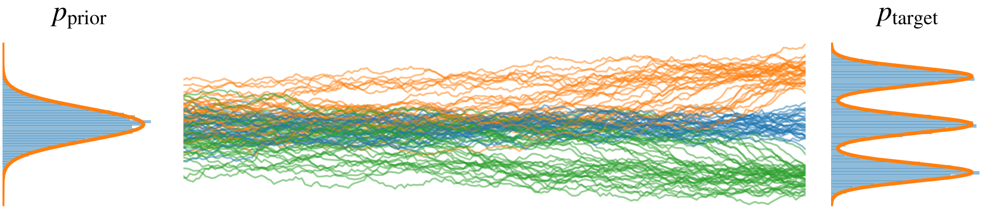
Goal: We want to sample from distribution $p_{\mathrm{target}} = \rho / \mathcal{Z}$.



Goal: We want to sample from distribution $p_{\text{target}} = \rho/\mathcal{Z}$.



Goal: We want to sample from distribution $p_{\mathrm{target}} = \rho/\mathcal{Z}$.

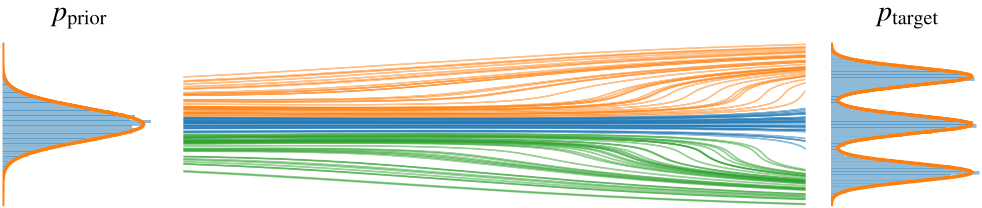


SDE

Setting:

$$X_0 \sim p_{\mathrm{prior}} \qquad \mathrm{d}X_s = \mu(X_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s$$

Goal: We want to sample from distribution $p_{\text{target}} = \rho/\mathcal{Z}$.

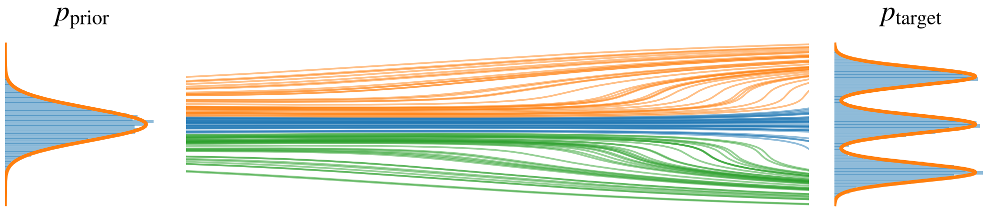


SDE **ODE**

Setting:

$X_0 \sim p_{\text{prior}}$ $\mathrm{d}X_s = \mu(X_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s$ $\mathrm{d}X_s = \mu(X_s, s)\,\mathrm{d}s$

Goal: We want to sample from distribution $p_{\text{target}} = \rho/\mathcal{Z}$.



SDE **ODE**

Setting:

$X_0 \sim p_{\text{prior}}$ $\quad \mathrm{d}X_s = \mu(X_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s \quad \mathrm{d}X_s = \mu(X_s, s)\,\mathrm{d}s$

Idea: Learn μ s.t. $X_T \sim p_{\text{target}}$.

Attempt I: PDE perspective

Attempt I: PDE perspective

- Considering the density of X_t, denoted by $p_X(\cdot, t)$, leads to the following PDEs:
  - **SDE:** Fokker-Planck equation

$$\partial_t p_X + \operatorname{div}(p_X \mu) - \frac{1}{2} \operatorname{Tr}(\sigma \sigma^\top \nabla^2 p_X) = 0,$$

Attempt I: PDE perspective

- Considering the density of X_t, denoted by $p_X(\cdot, t)$, leads to the following PDEs:
  - **SDE:** Fokker-Planck equation

$$\partial_t p_X + \operatorname{div}(p_X \mu) - \frac{1}{2} \operatorname{Tr}(\sigma\sigma^\top \nabla^2 p_X) = 0,$$

  - **ODE:** Continuity equation

$$\partial_t p_X + \operatorname{div}(p_X \mu) = 0,$$

Attempt I: PDE perspective

- Considering the density of X_t, denoted by $p_X(\cdot, t)$, leads to the following PDEs:
  - **SDE:** Fokker-Planck equation
  $$\partial_t p_X + \operatorname{div}(p_X \mu) - \frac{1}{2} \operatorname{Tr}(\sigma\sigma^\top \nabla^2 p_X) = 0,$$
  - **ODE:** Continuity equation
  $$\partial_t p_X + \operatorname{div}(p_X \mu) = 0,$$

  with boundary conditions $p_X(\cdot, 0) = p_{\mathrm{prior}}$ and $p_X(\cdot, T) = p_{\mathrm{target}}$.

Attempt I: PDE perspective

- Considering the density of X_t, denoted by $p_X(\cdot, t)$, leads to the following PDEs:
  - **SDE:** Fokker-Planck equation
$$\partial_t p_X + \operatorname{div}(p_X \mu) - \frac{1}{2} \operatorname{Tr}(\sigma\sigma^\top \nabla^2 p_X) = 0,$$
  - **ODE:** Continuity equation
$$\partial_t p_X + \operatorname{div}(p_X \mu) = 0,$$

  with boundary conditions $p_X(\cdot, 0) = p_{\mathrm{prior}}$ and $p_X(\cdot, T) = p_{\mathrm{target}}$.

- **Idea**: Identify pairs (μ, p_X) that fulfill the above PDEs.

- **Variational formulation of PDEs:** Consider loss functionals

$$\mathcal{L} : C(\mathbb{R}^d \times [0, T], \mathbb{R}^d) \times C(\mathbb{R}^d \times [0, T], \mathbb{R}) \to \mathbb{R}_{\geq 0}$$

that are zero if and only if a pair (μ, p_X) fulfills the corresponding PDE.

Attempt I: PDE perspective – learning the evolutions

- **Variational formulation of PDEs:** Consider loss functionals

$$\mathcal{L} : C(\mathbb{R}^d \times [0, T], \mathbb{R}^d) \times C(\mathbb{R}^d \times [0, T], \mathbb{R}) \to \mathbb{R}_{\geq 0}$$

  that are zero if and only if a pair (μ, p_X) fulfills the corresponding PDE.

- For numerical stability, we consider the PDEs in log-space, $V := \log p_X$, yielding
  - **SDE:**

$$\mathcal{R}_{\mathrm{logFP}}(\mu, V) := \partial_t V + \operatorname{div}(\mu) + \nabla V \cdot \mu - \tfrac{1}{2}\|\sigma^\top \nabla V\|^2 - \tfrac{1}{2}\operatorname{Tr}(\sigma\sigma^\top \nabla^2 V) = 0,$$

  - **ODE:**

$$\mathcal{R}_{\mathrm{logCE}}(\mu, V) := \partial_t V + \operatorname{div}(\mu) + \nabla V \cdot \mu = 0.$$

Attempt I: PDE perspective – general vs. constrained dynamics

- Learning μ and p_X simultaneously typically leads to non-unique solutions.

Attempt I: PDE perspective – general vs. constrained dynamics

- Learning μ and p_X simultaneously typically leads to non-unique solutions.
- Adding constraints allows for unique solutions:
  - **Annealing:** Fix p_X (i.e. a path of densities) and consider

$$\mathcal{R}^{\mathrm{anneal}}_{\mathrm{logFP}}(\widetilde{\mu}) := \mathcal{R}_{\mathrm{logFP}}(\widetilde{\mu}, V), \qquad \mathcal{R}^{\mathrm{anneal}}_{\mathrm{logCE}}(\widetilde{\mu}) := \mathcal{R}_{\mathrm{logCE}}(\widetilde{\mu}, V).$$

Attempt I: PDE perspective – general vs. constrained dynamics

- Learning μ and p_X simultaneously typically leads to non-unique solutions.
- Adding constraints allows for unique solutions:
  - **Annealing:** Fix p_X (i.e. a path of densities) and consider

$$\mathcal{R}_{\mathrm{logFP}}^{\mathrm{anneal}}(\widetilde{\mu}) := \mathcal{R}_{\mathrm{logFP}}(\widetilde{\mu}, V), \qquad \mathcal{R}_{\mathrm{logCE}}^{\mathrm{anneal}}(\widetilde{\mu}) := \mathcal{R}_{\mathrm{logCE}}(\widetilde{\mu}, V).$$

  - **Score-based generative modeling:** Fix $\mu = \sigma\sigma^\top \nabla V - f$ and consider

$$\mathcal{R}_{\mathrm{score}}(\widetilde{V}) := \mathcal{R}_{\mathrm{logFP}}(\sigma\sigma^\top \nabla \widetilde{V} - f, \widetilde{V}).$$

Attempt I: PDE perspective – general vs. constrained dynamics

- Learning μ and p_X simultaneously typically leads to non-unique solutions.
- Adding constraints allows for unique solutions:
  - **Annealing:** Fix p_X (i.e. a path of densities) and consider

$$\mathcal{R}^{\mathrm{anneal}}_{\mathrm{logFP}}(\widetilde{\mu}) := \mathcal{R}_{\mathrm{logFP}}(\widetilde{\mu}, V), \qquad \mathcal{R}^{\mathrm{anneal}}_{\mathrm{logCE}}(\widetilde{\mu}) := \mathcal{R}_{\mathrm{logCE}}(\widetilde{\mu}, V).$$

  - **Score-based generative modeling:** Fix $\mu = \sigma\sigma^\top \nabla V - f$ and consider

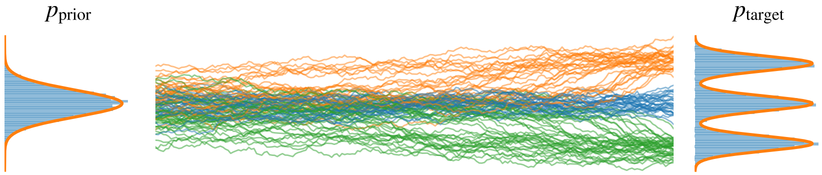
$$\mathcal{R}_{\mathrm{score}}(\widetilde{V}) := \mathcal{R}_{\mathrm{logFP}}(\sigma\sigma^\top \nabla \widetilde{V} - f, \widetilde{V}).$$

  - **Optimal transport & Schrödinger bridges:** Additionally minimize $\mathbb{E}\left[\frac{1}{2}\int_0^T \|\mu(X_s, s)\|^2 \mathrm{d}s\right]$. Find $\mu = \nabla \Phi$, where Φ solves

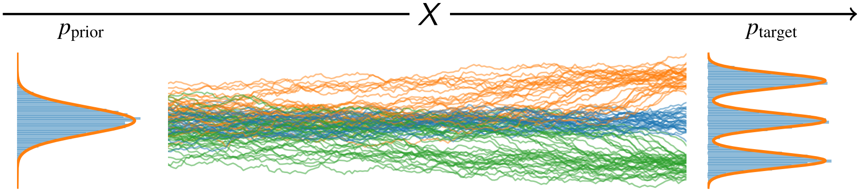
$$\mathcal{R}^{\mathrm{SB}}_{\mathrm{HJB}}(\Phi) := \partial_t \Phi + \tfrac{1}{2}\|\nabla \Phi\|^2 + \tfrac{1}{2}\operatorname{Tr}(\sigma\sigma^\top \nabla^2 \Phi) = 0, \qquad \mathcal{R}^{\mathrm{OT}}_{\mathrm{HJB}}(\Phi) := \partial_t \Phi + \tfrac{1}{2}\|\nabla \Phi\|^2 = 0.$$

Attempt II: Time-reversals and path space measures

Attempt II: Time-reversals and path space measures



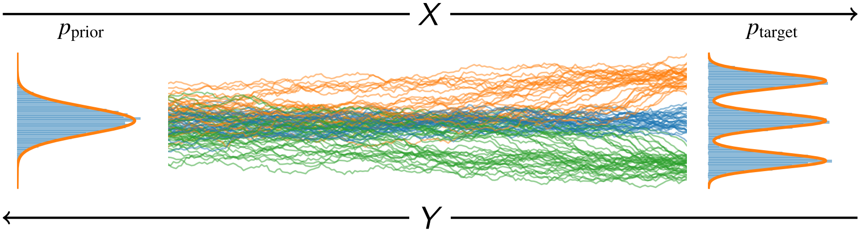
Attempt II: Time-reversals and path space measures



- **Setting:** Consider forward and reverse SDE:

$$\mathrm{d}X_s = \widetilde{\mu}_F(X_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \quad X_0 \sim p_{\text{prior}},$$

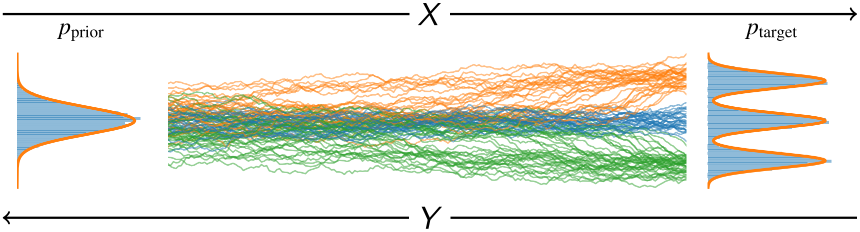
Attempt II: Time-reversals and path space measures



- **Setting:** Consider forward and reverse SDE:

$$\mathrm{d}X_s = \widetilde{\mu}_F(X_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \quad X_0 \sim p_{\text{prior}},$$
$$\mathrm{d}Y_s = \widetilde{\mu}_B(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\overleftarrow{\mathrm{d}}W_s, \quad Y_T \sim p_{\text{target}}.$$

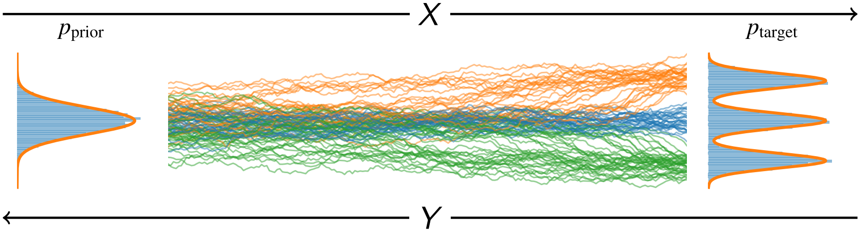
Attempt II: Time-reversals and path space measures



- **Setting:** Consider forward and reverse SDE:

$$\mathrm{d}X_s = \widetilde{\mu}_F(X_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \quad X_0 \sim p_{\text{prior}},$$
$$\mathrm{d}Y_s = \widetilde{\mu}_B(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\overleftarrow{\mathrm{d}}W_s, \quad Y_T \sim p_{\text{target}}.$$

- **Idea:** Learn $\widetilde{\mu}_F, \widetilde{\mu}_B$ s.t. X is time-reversal of Y, implying $X_T \sim p_{\text{target}}$, $Y_0 \sim p_{\text{prior}}$.

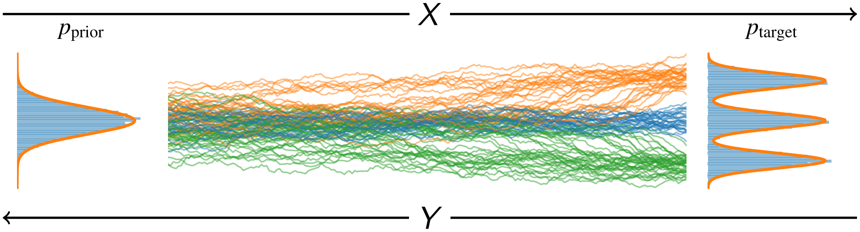


- **Setting:** Consider forward and reverse SDE:

$$\mathrm{d}X_s = \widetilde{\mu}_F(X_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \quad X_0 \sim p_{\text{prior}},$$
$$\mathrm{d}Y_s = \widetilde{\mu}_B(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\overleftarrow{\mathrm{d}}W_s, \quad Y_T \sim p_{\text{target}}.$$

- **Idea:** Learn $\widetilde{\mu}_F, \widetilde{\mu}_B$ s.t. X is time-reversal of Y, implying $X_T \sim p_{\text{target}}$, $Y_0 \sim p_{\text{prior}}$.
  - Learn $\widetilde{\mu}_F$ and $\widetilde{\mu}_B$ simultaneously: $\mathcal{L}^{\text{BSDE}}_{\text{Bridge}}(\widetilde{\mu}_F, \widetilde{\mu}_B)$

Attempt II: Time-reversals and path space measures

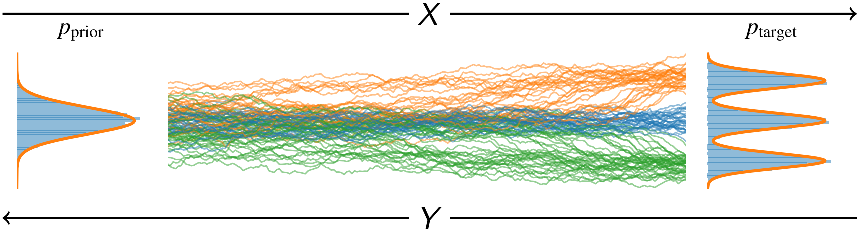


- **Setting:** Consider forward and reverse SDE:

$$\mathrm{d}X_s = \widetilde{\mu}_F(X_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \quad X_0 \sim p_{\mathrm{prior}},$$
$$\mathrm{d}Y_s = \widetilde{\mu}_B(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\overleftarrow{\mathrm{d}}W_s, \quad Y_T \sim p_{\mathrm{target}}.$$

- **Idea:** Learn $\widetilde{\mu}_F, \widetilde{\mu}_B$ s.t. X is time-reversal of Y, implying $X_T \sim p_{\mathrm{target}}$, $Y_0 \sim p_{\mathrm{prior}}$.
  - Learn $\widetilde{\mu}_F$ and $\widetilde{\mu}_B$ simultaneously: $\mathcal{L}^{\mathrm{BSDE}}_{\mathrm{Bridge}}(\widetilde{\mu}_F, \widetilde{\mu}_B)$
  - Fixe annealing p_X: $\mathcal{L}^{\mathrm{BSDE}}_{\mathrm{CMCD}}(\widetilde{\mu}_F)$

Attempt II: Time-reversals and path space measures



- **Setting:** Consider forward and reverse SDE:

$$\mathrm{d}X_s = \widetilde{\mu}_F(X_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \quad X_0 \sim p_{\mathrm{prior}},$$

$$\mathrm{d}Y_s = \widetilde{\mu}_B(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\overleftarrow{\mathrm{d}}W_s, \quad Y_T \sim p_{\mathrm{target}}.$$

- **Idea:** Learn $\widetilde{\mu}_F, \widetilde{\mu}_B$ s.t. X is time-reversal of Y, implying $X_T \sim p_{\mathrm{target}}$, $Y_0 \sim p_{\mathrm{prior}}$.
  - Learn $\widetilde{\mu}_F$ and $\widetilde{\mu}_B$ simultaneously: $\mathcal{L}^{\mathrm{BSDE}}_{\mathrm{Bridge}}(\widetilde{\mu}_F, \widetilde{\mu}_B)$
  - Fixe annealing p_X: $\mathcal{L}^{\mathrm{BSDE}}_{\mathrm{CMCD}}(\widetilde{\mu}_F)$
  - Fix $\widetilde{\mu}_B$ suitably: $\mathcal{L}^{\mathrm{BSDE}}_{\mathrm{DIS}}(\widetilde{\mu}_F)$

Attempt II: Time-reversals and path space measures

- We consider the SDEs

$$\mathrm{d}X_s = \widetilde{\mu}_F(X_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \quad X_0 \sim p_{\mathrm{prior}},$$
$$\mathrm{d}Y_s = \widetilde{\mu}_B(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\overleftarrow{\mathrm{d}}W_s, \quad Y_T \sim p_{\mathrm{target}}.$$

Attempt II: Time-reversals and path space measures

- We consider the SDEs

$$\mathrm{d}X_s = \widetilde{\mu}_F(X_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \quad X_0 \sim p_{\mathrm{prior}},$$
$$\mathrm{d}Y_s = \widetilde{\mu}_B(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\overleftarrow{\mathrm{d}}W_s, \quad Y_T \sim p_{\mathrm{target}}.$$

- **Path space perspective:** Consider path measures $\mathbb{P}_{X^{\mu_F}}$ and $\mathbb{P}_{\overleftarrow{Y}^{\mu_B}}$.

Attempt II: Time-reversals and path space measures

- We consider the SDEs

$$\mathrm{d}X_s = \widetilde{\mu}_F(X_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \quad X_0 \sim p_{\mathrm{prior}},$$
$$\mathrm{d}Y_s = \widetilde{\mu}_B(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\overleftarrow{\mathrm{d}}W_s, \quad Y_T \sim p_{\mathrm{target}}.$$

- **Path space perspective:** Consider path measures $\mathbb{P}_{X^{\mu_F}}$ and $\mathbb{P}_{\overleftarrow{Y}^{\mu_B}}$.
- Identify drifts μ_F, μ_B via divergence of those measures

$$\mu_F, \mu_B \in \operatorname*{arg\,min}_{\widetilde{\mu}_F, \widetilde{\mu}_B} D\big(\mathbb{P}_{X^{\widetilde{\mu}_F}} \big| \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}\big).$$

Attempt II: Time-reversals and path space measures

- We consider the SDEs

$$\mathrm{d}X_s = \widetilde{\mu}_F(X_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s, \quad X_0 \sim p_{\mathrm{prior}},$$
$$\mathrm{d}Y_s = \widetilde{\mu}_B(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\overleftarrow{\mathrm{d}}W_s, \quad Y_T \sim p_{\mathrm{target}}.$$

- **Path space perspective:** Consider path measures $\mathbb{P}_{X^{\mu_F}}$ and $\mathbb{P}_{\overleftarrow{Y}^{\mu_B}}$.
- Identify drifts μ_F, μ_B via divergence of those measures

$$\mu_F, \mu_B \in \operatorname*{arg\,min}_{\widetilde{\mu}_F, \widetilde{\mu}_B} D\left(\mathbb{P}_{X^{\widetilde{\mu}_F}} \middle| \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}\right).$$

Proposition (Log-likelihood for path measures)

$$\log \frac{\mathrm{d}\mathbb{P}_{X^{\widetilde{\mu}_F}}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}}(X^{\widetilde{\mu}_R}) = \int_0^T \left(\sigma^{-2}(\widetilde{\mu}_F + \widetilde{\mu}_B) \cdot \left(\widetilde{\mu}_R + \frac{\widetilde{\mu}_B - \widetilde{\mu}_F}{2}\right) + \nabla \cdot \widetilde{\mu}_B\right)(X_s^{\widetilde{\mu}_R}, s)\,\mathrm{d}s$$
$$+ \int_0^T \sigma^{-1}(\widetilde{\mu}_F + \widetilde{\mu}_B)(X_s^{\widetilde{\mu}_R}, s) \cdot \mathrm{d}W_s + \log \frac{p_{\mathrm{prior}}(X_0^{\widetilde{\mu}_R})}{p_{\mathrm{target}}(X_T^{\widetilde{\mu}_R})}$$

Connections and equivalences: divergences and loss functions

- We need to choose a loss $\mathcal{L}$ or a divergence D.

- We need to choose a loss $\mathcal{L}$ or a divergence D.
- First choice: BSDE-based losses for stochastic evolutions.

- We need to choose a loss $\mathcal{L}$ or a divergence D.
- First choice: BSDE-based losses for stochastic evolutions.
- **BSDE loss:** stochastic representation of PDE via Itô's formula. For the process

$$\mathrm{d}X_s = \mu(X_s, s)\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s.$$

and a PDE

$$\partial_t V + \tfrac{1}{2}\operatorname{Tr}\left(\sigma\sigma^\top \nabla^2 V\right) + \mu \cdot \nabla V + h(\cdot, \cdot, V, \nabla V) = 0$$

it holds

$$\mathcal{R}_{\mathrm{BSDE}}(V) = V(X_0, 0) - V(X_T, T) + \int_0^T \Big(\partial_s V + \tfrac{1}{2}\operatorname{Tr}(\sigma\sigma^\top \nabla^2 V) + \mu \cdot \nabla V\Big)(X_s, s)\,\mathrm{d}s + \int_0^T \sigma^\top \nabla V(X_s, s) \cdot \mathrm{d}W_s = 0.$$

Connections and equivalences: Time-reversals and BSDEs

- We need to choose a loss $\mathcal{L}$ or a divergence D.
- First choice: BSDE-based losses for stochastic evolutions.
- **BSDE loss:** stochastic representation of PDE via Itô's formula. For the process

$$\mathrm{d}X_s = \mu(X_s, s)\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s.$$

and a PDE

$$\partial_t V + \frac{1}{2}\operatorname{Tr}\left(\sigma\sigma^\top \nabla^2 V\right) + \mu \cdot \nabla V + h(\cdot, \cdot, V, \nabla V) = 0$$

it holds

$$\mathcal{R}_{\mathrm{BSDE}}(V) = V(X_0, 0) - V(X_T, T) - \int_0^T h(\cdot, \cdot, V, \nabla V)(X_s, s)\,\mathrm{d}s + \int_0^T \sigma^\top \nabla V(X_s, s) \cdot \mathrm{d}W_s = 0.$$

- We need to choose a loss $\mathcal{L}$ or a divergence D.
- First choice: BSDE-based losses for stochastic evolutions.
- **BSDE loss:** stochastic representation of PDE via Itô's formula. For the process

$$\mathrm{d}X_s = \mu(X_s, s)\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s.$$

and a PDE

$$\partial_t V + \tfrac{1}{2}\operatorname{Tr}\left(\sigma\sigma^\top \nabla^2 V\right) + \mu \cdot \nabla V + h(\cdot, \cdot, V, \nabla V) = 0$$

it holds

$$\mathcal{R}_{\mathrm{BSDE}}(V) = V(X_0, 0) - V(X_T, T) - \int_0^T h(\cdot, \cdot, V, \nabla V)(X_s, s)\,\mathrm{d}s + \int_0^T \sigma^\top \nabla V(X_s, s) \cdot \mathrm{d}W_s = 0.$$

- We can now consider the loss

$$\mathcal{L}_{\mathrm{BSDE}}(\widetilde{V}) = \mathbb{E}\left[\left(\mathcal{R}_{\mathrm{BSDE}}(\widetilde{V})(X)\right)^2\right],$$

where the expectation is over different realizations of the process X.

- BSDE-based losses are equivalent to a particular divergence between path space measures:

$$D^{\widetilde{\mu}_R}_{\mathrm{BSDE}}(\mathbb{P}_{X^{\widetilde{\mu}_F}} | \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}) = \mathbb{E}\left[\left(\log \frac{\mathrm{d}\mathbb{P}_{X^{\widetilde{\mu}_F}}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}}(X^{\widetilde{\mu}_R})\right)^2\right] = \mathcal{L}^{\mathrm{BSDE}}_{\mathrm{logFP}}(\widetilde{\mu}, \widetilde{V}) = \mathcal{L}^{\mathrm{BSDE}}_{\mathrm{Bridge}}(\widetilde{\mu}_F, \widetilde{\mu}_B)$$

- BSDE-based losses are equivalent to a particular divergence between path space measures:

$$D^{\widetilde{\mu}_R}_{\mathrm{BSDE}}(\mathbb{P}_{X^{\widetilde{\mu}_F}}|\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}) = \mathbb{E}\left[\left(\log \frac{\mathrm{d}\mathbb{P}_{X^{\widetilde{\mu}_F}}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}}(X^{\widetilde{\mu}_R})\right)^2\right] = \mathcal{L}^{\mathrm{BSDE}}_{\mathrm{logFP}}(\widetilde{\mu}, \widetilde{V}) = \mathcal{L}^{\mathrm{BSDE}}_{\mathrm{Bridge}}(\widetilde{\mu}_F, \widetilde{\mu}_B)$$

Proposition (Equivalence to trajectory-based methods)

The BSDE versions of our losses are equivalent to previously existing losses:

- BSDE-based losses are equivalent to a particular divergence between path space measures:

$$D^{\widetilde{\mu}_R}_{\mathrm{BSDE}}\left(\mathbb{P}_{X^{\widetilde{\mu}_F}} \middle| \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}\right) = \mathbb{E}\left[\left(\log \frac{\mathrm{d}\mathbb{P}_{X^{\widetilde{\mu}_F}}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}}(X^{\widetilde{\mu}_R})\right)^2\right] = \mathcal{L}^{\mathrm{BSDE}}_{\mathrm{logFP}}(\widetilde{\mu}, \widetilde{V}) = \mathcal{L}^{\mathrm{BSDE}}_{\mathrm{Bridge}}(\widetilde{\mu}_F, \widetilde{\mu}_B)$$

Proposition (Equivalence to trajectory-based methods)

The BSDE versions of our losses are equivalent to previously existing losses:

1. *Assuming the reparametrization $\sigma\sigma^\top \nabla \widetilde{V} = \widetilde{\mu}_F - \widetilde{\mu}_B$, it holds*

$$\mathcal{L}^{\mathrm{BSDE}}_{\mathrm{logFP}}(\widetilde{\mu}_F, \widetilde{V}) = \mathcal{L}^{\mathrm{BSDE}}_{\mathrm{Bridge}}(\widetilde{\mu}_F, \widetilde{\mu}_B).$$

Connections and equivalences: Time-reversals and BSDEs

- BSDE-based losses are equivalent to a particular divergence between path space measures:

$$D^{\widetilde{\mu}_R}_{\mathrm{BSDE}}\left(\mathbb{P}_{X^{\widetilde{\mu}_F}} \middle| \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}\right) = \mathbb{E}\left[\left(\log \frac{\mathrm{d}\mathbb{P}_{X^{\widetilde{\mu}_F}}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}}(X^{\widetilde{\mu}_R})\right)^2\right] = \mathcal{L}^{\mathrm{BSDE}}_{\mathrm{logFP}}(\widetilde{\mu}, \widetilde{V}) = \mathcal{L}^{\mathrm{BSDE}}_{\mathrm{Bridge}}(\widetilde{\mu}_F, \widetilde{\mu}_B)$$

Proposition (Equivalence to trajectory-based methods)

The BSDE versions of our losses are equivalent to previously existing losses:

1. *Assuming the reparametrization* $\sigma\sigma^\top \nabla \widetilde{V} = \widetilde{\mu}_F - \widetilde{\mu}_B$, *it holds*

$$\mathcal{L}^{\mathrm{BSDE}}_{\mathrm{logFP}}(\widetilde{\mu}_F, \widetilde{V}) = \mathcal{L}^{\mathrm{BSDE}}_{\mathrm{Bridge}}(\widetilde{\mu}_F, \widetilde{\mu}_B).$$

2. *It holds*

$$\mathcal{L}^{\mathrm{anneal,BSDE}}_{\mathrm{logFP}}(\widetilde{\mu}_F) = \mathcal{L}^{\mathrm{BSDE}}_{\mathrm{CMCD}}(\widetilde{\mu}_F).$$

Connections and equivalences: Time-reversals and BSDEs

- BSDE-based losses are equivalent to a particular divergence between path space measures:

$$D_{\mathrm{BSDE}}^{\widetilde{\mu}_R}(\mathbb{P}_{X^{\widetilde{\mu}_F}}|\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}) = \mathbb{E}\left[\left(\log \frac{\mathrm{d}\mathbb{P}_{X^{\widetilde{\mu}_F}}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}}(X^{\widetilde{\mu}_R})\right)^2\right] = \mathcal{L}_{\mathrm{logFP}}^{\mathrm{BSDE}}(\widetilde{\mu}, \widetilde{V}) = \mathcal{L}_{\mathrm{Bridge}}^{\mathrm{BSDE}}(\widetilde{\mu}_F, \widetilde{\mu}_B)$$

Proposition (Equivalence to trajectory-based methods)

The BSDE versions of our losses are equivalent to previously existing losses:

1. *Assuming the reparametrization* $\sigma\sigma^\top \nabla \widetilde{V} = \widetilde{\mu}_F - \widetilde{\mu}_B$, *it holds*

$$\mathcal{L}_{\mathrm{logFP}}^{\mathrm{BSDE}}(\widetilde{\mu}_F, \widetilde{V}) = \mathcal{L}_{\mathrm{Bridge}}^{\mathrm{BSDE}}(\widetilde{\mu}_F, \widetilde{\mu}_B).$$

2. *It holds*

$$\mathcal{L}_{\mathrm{logFP}}^{\mathrm{anneal,BSDE}}(\widetilde{\mu}_F) = \mathcal{L}_{\mathrm{CMCD}}^{\mathrm{BSDE}}(\widetilde{\mu}_F).$$

3. *Assuming the reparametrization* $\sigma\sigma^\top \nabla \widetilde{V} = \widetilde{\mu}_F - f$, *it holds*

$$\mathcal{L}_{\mathrm{score}}^{\mathrm{BSDE}}(\widetilde{V}) = \mathcal{L}_{\mathrm{DIS}}^{\mathrm{BSDE}}(\widetilde{\mu}_F).$$

Connections and equivalences: Path measures and optimal control

- $D = D_{\mathrm{KL}}(\mathbb{P}_{X^{\widetilde{\mu}_F}} | \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}) = \mathbb{E}\left[\log \frac{\mathrm{d}\mathbb{P}_{X^{\widetilde{\mu}_F}}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}}(X^{\widetilde{\mu}_F})\right]$ leads to stochastic optimal control:

Connections and equivalences: Path measures and optimal control

- $D = D_{\mathrm{KL}}(\mathbb{P}_{X^{\widetilde{\mu}_F}} | \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}) = \mathbb{E}\left[\log \frac{\mathrm{d}\mathbb{P}_{X^{\widetilde{\mu}_F}}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}}(X^{\widetilde{\mu}_F})\right]$ leads to stochastic optimal control:

Proposition (Verification theorem, time-reversed diffusion sampling (DIS))

Set $\widetilde{\mu}_F := f + \sigma u$, i.e. let X^u be defined by

$$\mathrm{d}X_s^u = (f + \sigma u)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s,$$

and fix $\widetilde{\mu}_B = f$. Consider the loss

$$\mathcal{L}(u) = D_{\mathrm{KL}}(\mathbb{P}_{X^u} | \mathbb{P}_{X^{u^*}}) = D_{\mathrm{KL}}(\mathbb{P}_{X^u} | \mathbb{P}_{\overleftarrow{Y}}) - D_{\mathrm{KL}}(\mathbb{P}_{X_0^u} | \mathbb{P}_{Y_T}),$$

where $\mathbb{P}_{X^u}$ denotes the path space measure of X^u etc. Then it holds that

$$-\log \mathcal{Z} = \min_{u \in \mathcal{U}} \mathcal{L}(u) := \min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \left(\frac{1}{2}\|u\|^2 - \mathrm{div}(f)\right)(X_s^u, s)\,\mathrm{d}s + \log \frac{p_{Y_T}(X_0^u)}{\rho(X_T^u)}\right],$$

where the unique minimum is attained by $u^* := \sigma^\top \nabla \log \overleftarrow{p}_Y$.

Path space perspective: Recovering score-based generative modeling

- **Detour:** let us consider the case with available data samples, but no density ρ

Path space perspective: Recovering score-based generative modeling

- **Detour:** let us consider the case with available data samples, but no density ρ
- The previous loss is not feasible $\rightarrow$ we cannot minimize a divergence directly

Path space perspective: Recovering score-based generative modeling

- **Detour:** let us consider the case with available data samples, but no density ρ
- The previous loss is not feasible $\to$ we cannot minimize a divergence directly
- **Trick:** instead of $D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{\overleftarrow{Y}})$ let us consider $D_{\mathrm{KL}}(\mathbb{P}_Y|\mathbb{P}_{\overleftarrow{X}^u}) = \mathbb{E}\left[\log \frac{\mathrm{d}\mathbb{P}_Y}{\mathrm{d}\mathbb{P}_{\overleftarrow{X}^u}}(Y)\right]$ (which is possible since we have data samples)

Path space perspective: Recovering score-based generative modeling

- **Detour:** let us consider the case with available data samples, but no density ρ
- The previous loss is not feasible $\rightarrow$ we cannot minimize a divergence directly
- **Trick:** instead of $D_{\mathrm{KL}}(\mathbb{P}_{X^u} | \mathbb{P}_{\overleftarrow{Y}})$ let us consider $D_{\mathrm{KL}}(\mathbb{P}_{Y} | \mathbb{P}_{\overleftarrow{X}^u}) = \mathbb{E}\left[\log \frac{\mathrm{d}\mathbb{P}_Y}{\mathrm{d}\mathbb{P}_{\overleftarrow{X}^u}}(Y)\right]$ (which is possible since we have data samples), yielding

$$\mathcal{L}(u) = \min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \left(\frac{1}{2}\|u\|^2 + \mathrm{div}(\sigma u - f)\right)(Y_s, s)\,\mathrm{d}s + \log \frac{p_{X_T^u}(Y_0)}{p_{X_0^u}(Y_T)}\right].$$

Path space perspective: Recovering score-based generative modeling

- **Detour:** let us consider the case with available data samples, but no density ρ
- The previous loss is not feasible $\rightarrow$ we cannot minimize a divergence directly
- **Trick:** instead of $D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{\overleftarrow{Y}})$ let us consider $D_{\mathrm{KL}}(\mathbb{P}_{Y}|\mathbb{P}_{\overleftarrow{X}^u}) = \mathbb{E}\left[\log \frac{\mathrm{d}\mathbb{P}_Y}{\mathrm{d}\mathbb{P}_{\overleftarrow{X}^u}}(Y)\right]$ (which is possible since we have data samples), yielding

$$\mathcal{L}(u) = \min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \left(\frac{1}{2}\|u\|^2 + \operatorname{div}(\sigma u - f)\right)(Y_s, s)\,\mathrm{d}s + \log \frac{p_{X_T^u}(Y_0)}{p_{X_0^u}(Y_T)}\right].$$

- Not tractable since $p_{X_T^u}$ is not known

Path space perspective: Recovering score-based generative modeling

- **Detour:** let us consider the case with available data samples, but no density ρ
- The previous loss is not feasible $\rightarrow$ we cannot minimize a divergence directly
- **Trick:** instead of $D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{\overleftarrow{Y}})$ let us consider $D_{\mathrm{KL}}(\mathbb{P}_Y|\mathbb{P}_{\overleftarrow{X}^u}) = \mathbb{E}\left[\log \frac{\mathrm{d}\mathbb{P}_Y}{\mathrm{d}\mathbb{P}_{\overleftarrow{X}^u}}(Y)\right]$ (which is possible since we have data samples), yielding

$$\mathcal{L}(u) = \min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \left(\frac{1}{2}\|u\|^2 + \mathrm{div}(\sigma u - f)\right)(Y_s, s)\,\mathrm{d}s + \log \frac{p_{X_T^u}(Y_0)}{p_{X_0^u}(Y_T)}\right].$$

- Not tractable since $p_{X_T^u}$ is not known$\rightarrow$ as a remedy, maximize ELBO

$$\mathbb{E}\left[\log p_{X_T^u}(Y_0)\right] \geq \mathbb{E}\left[\log p_{X_0^u}(Y_T) - \int_0^T \left(\frac{1}{2}\|u\|^2 + \mathrm{div}(\sigma u - f)\right)(Y_s, s)\,\mathrm{d}s\right]$$

Path space perspective: Recovering score-based generative modeling

- **Denoising score matching:** Rewrite the divergence and apply Monte Carlo approximation of the time-integral, using $\tau \sim \mathcal{U}([0, T])$, so that no trajectories are needed anymore:

Path space perspective: Recovering score-based generative modeling

- **Denoising score matching:** Rewrite the divergence and apply Monte Carlo approximation of the time-integral, using $\tau \sim \mathcal{U}([0, T])$, so that no trajectories are needed anymore:

$$\mathbb{E}\left[\log p_{X_T^u}(Y_0)\right] \geq \mathbb{E}\left[\log p_{X_0^u}(Y_T) - \int_0^T \left(\frac{1}{2}\|u\|^2 + \operatorname{div}(\sigma u - f)\right)(Y_s, s)\,\mathrm{d}s\right]$$

Path space perspective: Recovering score-based generative modeling

- **Denoising score matching:** Rewrite the divergence and apply Monte Carlo approximation of the time-integral, using $\tau \sim \mathcal{U}([0, T])$, so that no trajectories are needed anymore:

Divergence theorem: $u \cdot \sigma^\top \nabla \log p_{Y_s|Y_0}$ (in expectation)

$$\mathbb{E}\left[\log p_{X_T^u}(Y_0)\right] \geq \mathbb{E}\left[\underbrace{\log p_{X_0^u}(Y_T)}_{\text{const.}} - \int_0^T \left(\frac{1}{2}\|u\|^2 + \operatorname{div}(\sigma u) - \underbrace{\operatorname{div}(f)}_{\text{const.}}\right)(Y_s, s)\,\mathrm{d}s\right]$$

- **Denoising score matching:** Rewrite the divergence and apply Monte Carlo approximation of the time-integral, using $\tau \sim \mathcal{U}([0, T])$, so that no trajectories are needed anymore:

Divergence theorem: $u \cdot \sigma^\top \nabla \log p_{Y_s|Y_0}$ (in expectation)

$$\mathbb{E}\left[\log p_{X_T^u}(Y_0)\right] \geq \mathbb{E}\left[\underbrace{\log p_{X_0^u}(Y_T)}_{\text{const.}} - \int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\sigma u) - \underbrace{\operatorname{div}(f)}_{\text{const.}}\right)(Y_s, s)\,\mathrm{d}s\right]$$

$$= \frac{T}{2}\underbrace{\mathbb{E}\left[\left\|u(Y_\tau, \tau) - \sigma^\top(\tau)\nabla \log p_{Y_\tau|Y_0}(Y_\tau|Y_0)\right\|^2\right]}_{\text{denoising score matching}} + \text{const.}$$

- We propose a novel divergence:

Definition (Log-variance divergence)

$$D_{\mathrm{LV}}^{\widetilde{\mu}_R}\left(\mathbb{P}_{X^{\widetilde{\mu}_F}}, \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}\right) := \operatorname{Var}\left(\log \frac{\mathrm{d}\mathbb{P}_{X^{\widetilde{\mu}_F}}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}}(X^{\widetilde{\mu}_R})\right)$$

Path space perspective: Novel divergences

- We propose a novel divergence:

Definition (Log-variance divergence)

$$D_{\mathrm{LV}}^{\widetilde{\mu}_R}\left(\mathbb{P}_{X^{\widetilde{\mu}_F}}, \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}\right) := \mathrm{Var}\left(\log \frac{\mathrm{d}\mathbb{P}_{X^{\widetilde{\mu}_F}}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}}(X^{\widetilde{\mu}_R})\right)$$

- In principle arbitrary choice for $\widetilde{\mu}_R$ allows to balance exploration and exploitation.

Path space perspective: Novel divergences

- We propose a novel divergence:

Definition (Log-variance divergence)

$$D_{\mathrm{LV}}^{\widetilde{\mu}_R}(\mathbb{P}_{X^{\widetilde{\mu}_F}}, \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}) := \mathrm{Var}\left(\log \frac{\mathrm{d}\mathbb{P}_{X^{\widetilde{\mu}_F}}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}}(X^{\widetilde{\mu}_R})\right)$$

- In principle arbitrary choice for $\widetilde{\mu}_R$ allows to balance exploration and exploitation.
- No differentiation through the SDE solver.

Path space perspective: Novel divergences

- We propose a novel divergence:

Definition (Log-variance divergence)

$$D_{\mathrm{LV}}^{\widetilde{\mu}_R}(\mathbb{P}_{X^{\widetilde{\mu}_F}}, \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}) := \mathrm{Var}\left(\log \frac{\mathrm{d}\mathbb{P}_{X^{\widetilde{\mu}_F}}}{\mathrm{d}\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}}(X^{\widetilde{\mu}_R})\right)$$

- In principle arbitrary choice for $\widetilde{\mu}_R$ allows to balance exploration and exploitation.
- No differentiation through the SDE solver.

Proposition (Equivalence with KL divergence)

$$\frac{1}{2}\left(\frac{\delta}{\delta\widetilde{\mu}_F} D_{\mathrm{LV}}^{\widetilde{\mu}_R}(\mathbb{P}_{X^{\widetilde{\mu}_F}}, \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}})\Big|_{\widetilde{\mu}_R=\widetilde{\mu}_F}\right) = \frac{\delta}{\delta\widetilde{\mu}_F} D_{\mathrm{KL}}(\mathbb{P}_{X^{\widetilde{\mu}_F}} | \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}})$$

The log-variance divergence: Variance reduction

Proposition (Control variate)

$$\frac{1}{2}\left(\frac{\delta}{\delta\widetilde{\mu}_F}\widehat{D}_{\mathrm{LV}}^{\widetilde{\mu}_R}(\mathbb{P}_{X^{\widetilde{\mu}_F}}, \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}})\Big|_{\widetilde{\mu}_R=\widetilde{\mu}_F}\right) \quad \textit{is a control variate version of} \quad \frac{\delta}{\delta\widetilde{\mu}_F}\widehat{D}_{\mathrm{KL}}(\mathbb{P}_{X^{\widetilde{\mu}_F}}|\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}).$$

The log-variance divergence: Variance reduction

Proposition (Control variate)

$$\frac{1}{2}\left(\frac{\delta}{\delta\widetilde{\mu}_F}\widehat{D}^{\widetilde{\mu}_R}_{\mathrm{LV}}(\mathbb{P}_{X^{\widetilde{\mu}_F}}, \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}})\Big|_{\widetilde{\mu}_R=\widetilde{\mu}_F}\right) \quad \textit{is a control variate version of} \quad \frac{\delta}{\delta\widetilde{\mu}_F}\widehat{D}_{\mathrm{KL}}(\mathbb{P}_{X^{\widetilde{\mu}_F}}|\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}).$$

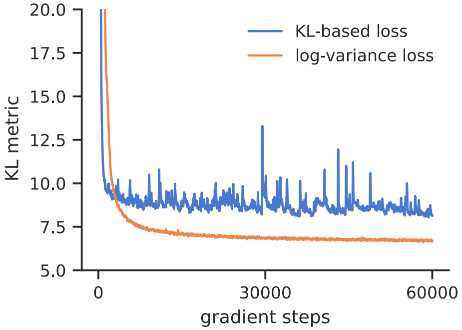
- $X^{\mathrm{CV}} = X + C$, where $\mathbb{E}[C] = 0$

The log-variance divergence: Variance reduction

Proposition (Control variate)

$$\frac{1}{2}\left(\frac{\delta}{\delta\widetilde{\mu}_F}\widehat{D}^{\widetilde{\mu}_R}_{\mathrm{LV}}(\mathbb{P}_{X^{\widetilde{\mu}_F}},\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}})\Big|_{\widetilde{\mu}_R=\widetilde{\mu}_F}\right) \quad \textit{is a control variate version of} \quad \frac{\delta}{\delta\widetilde{\mu}_F}\widehat{D}_{\mathrm{KL}}(\mathbb{P}_{X^{\widetilde{\mu}_F}}|\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}).$$

- $X^{\mathrm{CV}} = X + C$, where $\mathbb{E}[C] = 0$

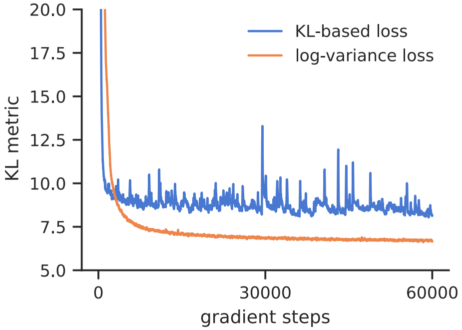


The log-variance divergence: Variance reduction

Proposition (Control variate)

$$\frac{1}{2}\left(\frac{\delta}{\delta \widetilde{\mu}_F}\widehat{D}_{\mathrm{LV}}^{\widetilde{\mu}_R}(\mathbb{P}_{X^{\widetilde{\mu}_F}}, \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}})\Big|_{\widetilde{\mu}_R=\widetilde{\mu}_F}\right) \quad \textit{is a control variate version of} \quad \frac{\delta}{\delta \widetilde{\mu}_F}\widehat{D}_{\mathrm{KL}}(\mathbb{P}_{X^{\widetilde{\mu}_F}}|\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}).$$

- $X^{\mathrm{CV}} = X + C$, where $\mathbb{E}[C] = 0$
- This leads to variance reduction in the estimated gradient.

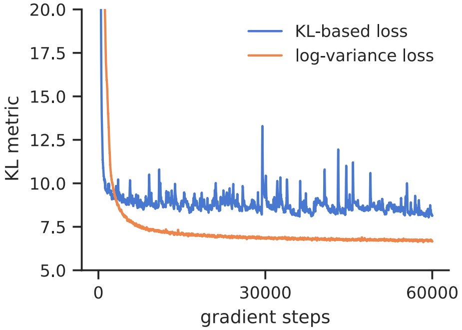


The log-variance divergence: Variance reduction

Proposition (Control variate)

$$\frac{1}{2}\left(\frac{\delta}{\delta\widetilde{\mu}_F}\widehat{D}^{\widetilde{\mu}_R}_{\mathrm{LV}}(\mathbb{P}_{X^{\widetilde{\mu}_F}},\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}})\Big|_{\widetilde{\mu}_R=\widetilde{\mu}_F}\right) \quad \textit{is a control variate version of} \quad \frac{\delta}{\delta\widetilde{\mu}_F}\widehat{D}_{\mathrm{KL}}(\mathbb{P}_{X^{\widetilde{\mu}_F}}|\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}).$$

- $X^{\mathrm{CV}} = X + C$, where $\mathbb{E}[C] = 0$
- This leads to variance reduction in the estimated gradient.
- Usually implying faster and better convergence of gradient based optimization.



The log-variance divergence: Robustness properties

Proposition (Robustness at solution)

$$\mathrm{Var}\left(\frac{\delta}{\delta\widetilde{\mu}_F}\Big|_{\widetilde{\mu}_F=\mu_F}\widehat{D}_{\mathrm{LV}}^{\widetilde{\mu}_R}(\mathbb{P}_{X^{\widetilde{\mu}_F}},\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}})\right)=0, \qquad \mathrm{Var}\left(\frac{\delta}{\delta\widetilde{\mu}_B}\Big|_{\widetilde{\mu}_B=\mu_B}\widehat{D}_{\mathrm{LV}}^{\widetilde{\mu}_R}(\mathbb{P}_{X^{\widetilde{\mu}_F}},\mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_F}})\right)=0.$$

The log-variance divergence: Robustness properties

Proposition (Robustness at solution)

$$\mathrm{Var}\left(\frac{\delta}{\delta \widetilde{\mu}_F}\Big|_{\widetilde{\mu}_F = \mu_F} \widehat{D}^{\widetilde{\mu}_R}_{\mathrm{LV}}(\mathbb{P}_{X^{\widetilde{\mu}_F}}, \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_B}}) \right) = 0, \qquad \mathrm{Var}\left(\frac{\delta}{\delta \widetilde{\mu}_B}\Big|_{\widetilde{\mu}_B = \mu_B} \widehat{D}^{\widetilde{\mu}_R}_{\mathrm{LV}}(\mathbb{P}_{X^{\widetilde{\mu}_F}}, \mathbb{P}_{\overleftarrow{Y}^{\widetilde{\mu}_F}}) \right) = 0.$$

Proposition (Robustness in high dimensions)

$$\frac{\sqrt{\mathrm{Var}\left(\widehat{D}^{w}_{\mathrm{LV}}\left(\bigotimes_{i=1}^{d} \mathbb{P}_i, \bigotimes_{i=1}^{d} \mathbb{Q}_i \right)\right)}}{D^{w}_{\mathrm{LV}}\left(\bigotimes_{i=1}^{d} \mathbb{P}_i, \bigotimes_{i=1}^{d} \mathbb{Q}_i \right)}$$

can be bounded uniformly in d.

- **Alternative:** PINN-based losses for stochastic and deterministic evolutions.

A simulation-free attempt based on PDEs: PINN-based losses

- **Alternative:** PINN-based losses for stochastic and deterministic evolutions.

- Consider parametrization (encoding the boundary conditions)

$$\widetilde{V}_{\varphi,z}(\cdot, t) = \frac{t}{T} \log \frac{\rho_{\mathrm{target}}}{z(t)} + \left(1 - \frac{t}{T}\right) \log p_{\mathrm{prior}} + \frac{t}{T}\left(1 - \frac{t}{T}\right) \varphi(\cdot, t),$$

where we learn z and φ, cf. Máté & Fleuret, 2023.

A simulation-free attempt based on PDEs: PINN-based losses

- **Alternative:** PINN-based losses for stochastic and deterministic evolutions.

- Consider parametrization (encoding the boundary conditions)

$$\widetilde{V}_{\varphi,z}(\cdot, t) = \tfrac{t}{T} \log \tfrac{\rho_{\mathrm{target}}}{z(t)} + \left(1 - \tfrac{t}{T}\right) \log p_{\mathrm{prior}} + \tfrac{t}{T}\left(1 - \tfrac{t}{T}\right) \varphi(\cdot, t),$$

  where we learn z and φ, cf. Máté & Fleuret, 2023.

- We can then minimize

$$\mathcal{L}(\widetilde{\mu}, \widetilde{V}) = \mathbb{E}\left[\left(\mathcal{R}(\widetilde{\mu}, \widetilde{V})(\xi, \tau)\right)^2\right],$$

  where $(\xi, \tau) \sim \nu$ are sampled from a measure ν, e.g. $\nu = \mathrm{Unif}(\Omega \times [0, T]), \Omega \subset \mathbb{R}^d$.

Numerical implications of the losses

- BSDE-bassed losses:

Numerical implications of the losses

- BSDE-bassed losses:
  - Neither second-order nor time derivatives have to be computed.

Numerical implications of the losses

- BSDE-bassed losses:
  - Neither second-order nor time derivatives have to be computed.
  - Gradients of the solutions (usually corresponding to the learned drift) can be learned directly.

Numerical implications of the losses

- BSDE-bassed losses:
  - Neither second-order nor time derivatives have to be computed.
  - Gradients of the solutions (usually corresponding to the learned drift) can be learned directly.
  - Only stochastic dynamics can be approached.

Numerical implications of the losses

- BSDE-bassed losses:
  - Neither second-order nor time derivatives have to be computed.
  - Gradients of the solutions (usually corresponding to the learned drift) can be learned directly.
  - Only stochastic dynamics can be approached.
- PINN-based losses:

Numerical implications of the losses

- BSDE-bassed losses:
  - Neither second-order nor time derivatives have to be computed.
  - Gradients of the solutions (usually corresponding to the learned drift) can be learned directly.
  - Only stochastic dynamics can be approached.

- PINN-based losses:
  - Can be readily applied to deterministic evolutions.

Numerical implications of the losses

- BSDE-bassed losses:
  - Neither second-order nor time derivatives have to be computed.
  - Gradients of the solutions (usually corresponding to the learned drift) can be learned directly.
  - Only stochastic dynamics can be approached.

- PINN-based losses:
  - Can be readily applied to deterministic evolutions.
  - Simulation-free, no time-discretization.

Numerical implications of the losses

- BSDE-bassed losses:
  - Neither second-order nor time derivatives have to be computed.
  - Gradients of the solutions (usually corresponding to the learned drift) can be learned directly.
  - Only stochastic dynamics can be approached.

- PINN-based losses:
  - Can be readily applied to deterministic evolutions.
  - Simulation-free, no time-discretization.
  - Off-policy training comes by design.

Numerical implications of the losses

- BSDE-bassed losses:
  - Neither second-order nor time derivatives have to be computed.
  - Gradients of the solutions (usually corresponding to the learned drift) can be learned directly.
  - Only stochastic dynamics can be approached.

- PINN-based losses:
  - Can be readily applied to deterministic evolutions.
  - Simulation-free, no time-discretization.
  - Off-policy training comes by design.
  - Need to know "essential support" of target density.

Numerical implications of the losses

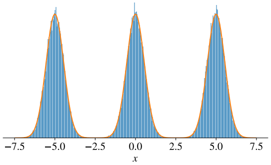
- BSDE-bassed losses:
  - Neither second-order nor time derivatives have to be computed.
  - Gradients of the solutions (usually corresponding to the learned drift) can be learned directly.
  - Only stochastic dynamics can be approached.
- PINN-based losses:
  - Can be readily applied to deterministic evolutions.
  - Simulation-free, no time-discretization.
  - Off-policy training comes by design.
  - Need to know "essential support" of target density.
  - Training is sensitive to hyperparameter tuning.

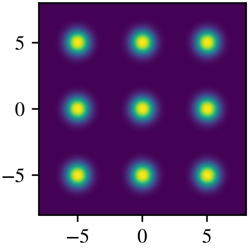
Numerical examples

- We consider the following losses:

Method	Stochastic	Deterministic	BSDE version	Unique
General bridge	$\mathcal{L}_{\mathrm{logFP}}(\widetilde{\mu}, \widetilde{V})$	$\mathcal{L}_{\mathrm{logCE}}(\widetilde{\mu}, \widetilde{V})$	Bridge	✗
Prescribed bridge	$\mathcal{L}^{\mathrm{anneal}}_{\mathrm{logFP}}(\widetilde{\mu})$	$\mathcal{L}^{\mathrm{anneal}}_{\mathrm{logCE}}(\widetilde{\mu})$	CMCD	✓
Score-based	$\mathcal{L}_{\mathrm{score}}(\widetilde{V})$		DIS	✓
SB & OT	$\mathcal{L}_{\mathrm{SB}}(\widetilde{\mu}, \widetilde{V})$	$\mathcal{L}_{\mathrm{OT}}(\widetilde{\mu}, \widetilde{V})$		✓

Numerical examples: Gaussian mixture ($d = 2$, 9 modes)

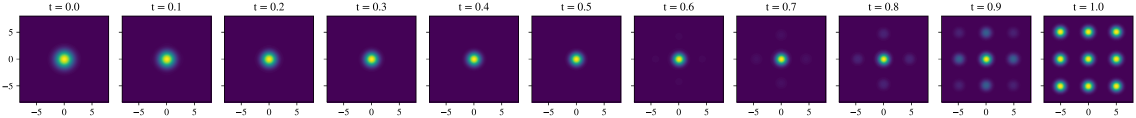




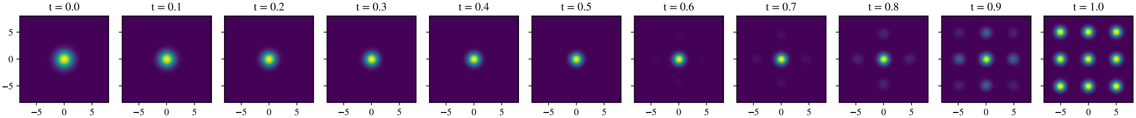
Problem	Method	Loss	$\Delta \log Z \downarrow$	$\mathcal{W}_\gamma^2 \downarrow$	ESS $\uparrow$	Δ std $\downarrow$	sec./it. $\downarrow$
GMM	PIS-KL		1.094	0.467	0.0051	1.937	0.503
($d = 2$)	PIS-LV		0.046	**0.020**	0.9093	0.023	0.500
	DIS-KL		1.551	0.064	0.0226	2.522	0.565
	DIS-LV		0.056	**0.020**	0.8660	0.004	0.536
	SDE	$\mathcal{L}_{\mathrm{logFP}}$	**0.000**	**0.020**	**1.0000**	0.004	0.011
	SDE-anneal	$\mathcal{L}_{\mathrm{logFP}}^{\mathrm{anneal}}$	5.364	0.172	0.1031	0.209	0.062
	SDE-score	$\mathcal{L}_{\mathrm{score}}$	0.009	**0.020**	0.9818	0.096	0.013
	SB	$\mathcal{L}_{\mathrm{SB}}$	0.002	**0.020**	0.9959	0.050	0.017
	ODE	$\mathcal{L}_{\mathrm{logCE}}$	**0.000**	**0.020**	**1.0000**	**0.003**	**0.008**
	ODE-anneal	$\mathcal{L}_{\mathrm{logCE}}^{\mathrm{anneal}}$	4.227	0.044	0.1427	0.753	0.020
	OT	$\mathcal{L}_{\mathrm{OT}}$	0.005	0.057	0.9932	0.065	0.080

Numerical examples: Gaussian mixture ($d = 2$, 9 modes)

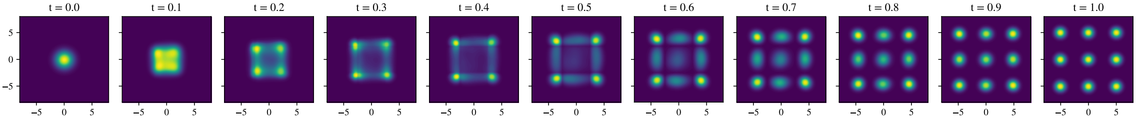
- Geometric annealing path can be suboptimal ($\mathcal{L}_{\mathrm{logCE}}^{\mathrm{anneal}}$):



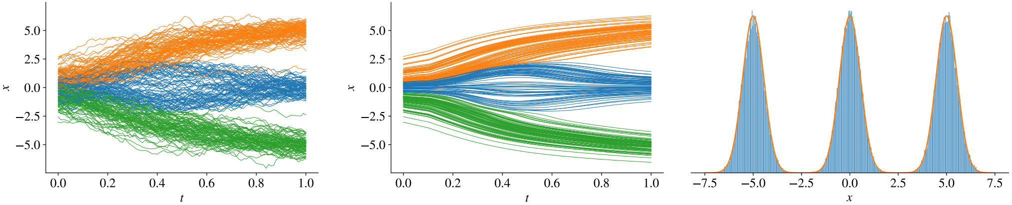
- Geometric annealing path can be suboptimal ($\mathcal{L}_{\mathrm{logCE}}^{\mathrm{anneal}}$):



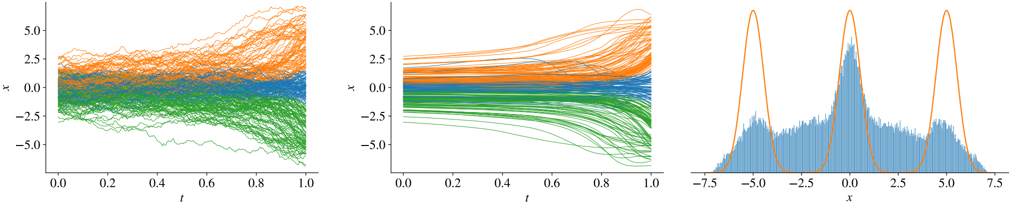
- The learned path seems to be more appropriate ($\mathcal{L}_{\mathrm{logCE}}$):



Numerical examples: Gaussian mixture ($d = 2$, 9 modes)

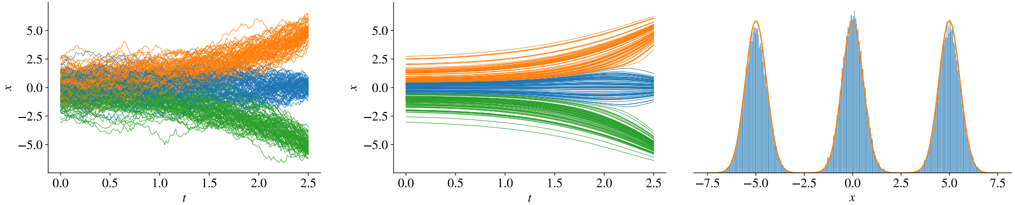


$\mathcal{L}_{\mathrm{logFP}}$

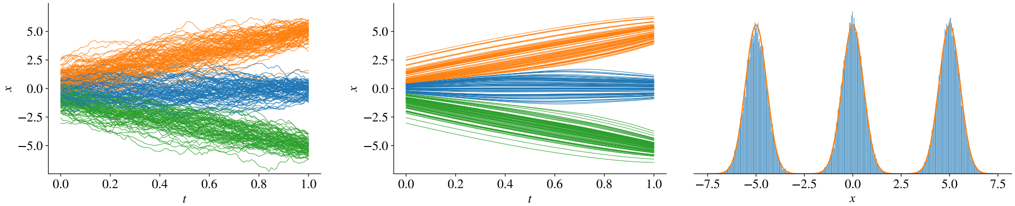


$\mathcal{L}_{\mathrm{logFP}}^{\mathrm{anneal}}$

Numerical examples: Gaussian mixture ($d = 2$, 9 modes)



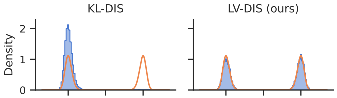
$\mathcal{L}_{\mathrm{score}}$

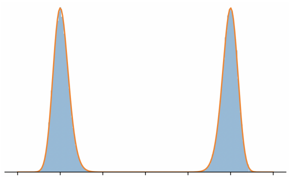


$\mathcal{L}_{\mathrm{SB}}$

Numerical examples: Double well ($d = 5$, 32 modes)

$$\rho(x) := \exp\left(-\sum_{i=1}^{5}(x_i^2 - 4)^2\right)$$

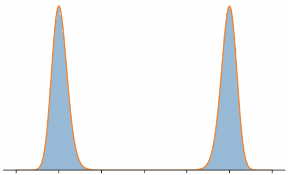




Problem	Method	Loss	$\Delta \log Z \downarrow$	$\mathcal{W}_\gamma^2 \downarrow$	ESS $\uparrow$	Δ std $\downarrow$	sec./it. $\downarrow$
MW	PIS-KL		3.567	1.699	0.0004	1.409	0.441
($d = 5, m = 5, \delta = 4$)	PIS-LV		0.214	0.121	0.6744	0.001	0.402
	DIS-KL		1.462	1.175	0.0012	0.431	0.490
	DIS-LV		0.375	0.120	0.4519	0.001	0.437
	SDE	$\mathcal{L}_{\mathrm{logFP}}$	0.161	0.123	0.8167	0.016	0.017
	SDE-anneal	$\mathcal{L}_{\mathrm{logFP}}^{\mathrm{anneal}}$	0.842	0.257	0.3464	0.004	0.014
	SDE-score	$\mathcal{L}_{\mathrm{score}}$	3.969	0.427	0.0124	0.004	0.026
	SB	$\mathcal{L}_{\mathrm{SB}}$	7.855	0.328	0.0314	0.045	0.029
	ODE	$\mathcal{L}_{\mathrm{logCE}}$	**0.000**	**0.118**	**0.9993**	**0.000**	**0.008**
	ODE-anneal	$\mathcal{L}_{\mathrm{logCE}}^{\mathrm{anneal}}$	0.025	0.121	0.9506	0.005	0.010
	OT	$\mathcal{L}_{\mathrm{OT}}$	0.010	0.120	0.9862	0.002	0.020

Numerical examples: Double well ($d = 50$, 32 modes)

$$\rho(x) := \exp\left(-\sum_{i=1}^{5}(x_i^2 - 2)^2 - \frac{1}{2}\sum_{i=6}^{50} x_i^2\right)$$



Problem	Method	Loss	$\Delta \log Z \downarrow$	$\mathcal{W}_\gamma^2 \downarrow$	ESS $\uparrow$	Δ std $\downarrow$	sec./it. $\downarrow$
MW	PIS-KL		0.101	6.821	0.8172	0.001	0.479
($d = 50, m = 5, \delta = 2$)	PIS-LV		0.087	6.823	0.8453	**0.000**	0.416
	DIS-KL		1.785	6.854	0.0225	0.009	0.522
	DIS-LV		1.783	6.855	0.0227	0.009	0.450
	SDE	$\mathcal{L}_{\mathrm{logFP}}$	0.038	6.820	0.9511	0.001	0.050
	SDE-anneal	$\mathcal{L}_{\mathrm{logFP}}^{\mathrm{anneal}}$	0.270	6.899	0.9171	0.021	0.067
	SDE-score	$\mathcal{L}_{\mathrm{score}}$	1.989	**6.803**	0.1065	0.016	0.053
	SB	$\mathcal{L}_{\mathrm{SB}}$	189.71	7.552	0.0106	0.051	0.053
	ODE	$\mathcal{L}_{\mathrm{logCE}}$	**0.003**	6.815	**0.9937**	0.002	**0.023**
	ODE-anneal	$\mathcal{L}_{\mathrm{logCE}}^{\mathrm{anneal}}$	1.759	6.821	0.2100	0.017	0.043
	OT	$\mathcal{L}_{\mathrm{OT}}$	0.104	6.824	0.9027	0.001	0.043

Numerical examples: ϕ^4 field theory

$$\rho(\phi) = \exp\left(-\sum_{x\in\Lambda}\left(-2\kappa\sum_{\widehat{\mu}=1}^{2}\phi(x)\phi(x+\widehat{\mu}) + (1-2\lambda)\phi(x)^2 + \lambda\phi(x)^4\right)\right)$$

Numerical examples: ϕ^4 field theory

$$\rho(x) = \exp\left(-\sum_{i=1}^{d}\left(-2\kappa(x_i x_{i-L} + x_i x_{i+1}) + (1-2\lambda)x_i^2 + \lambda x_i^4\right)\right)$$

Numerical examples: ϕ^4 field theory

$$\rho(x) = \exp\left(- \sum_{i=1}^{d} \left(-2\kappa(x_i x_{i-L} + x_i x_{i+1}) + (1 - 2\lambda) x_i^2 + \lambda x_i^4 \right) \right)$$

- Difficulty depends non-trivially on the choices of κ and λ.

Numerical examples: ϕ^4 field theory

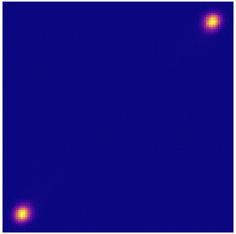
$$\rho(x) = \exp\left(-\sum_{i=1}^{d}\left(-2\kappa(x_i x_{i-L} + x_i x_{i+1}) + (1-2\lambda)x_i^2 + \lambda x_i^4\right)\right)$$

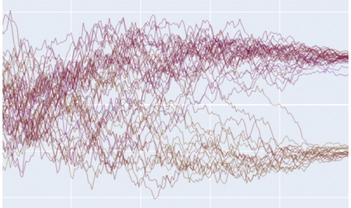
- Difficulty depends non-trivially on the choices of κ and λ.
- Preliminary results, not incorporating any symmetries (with $\lambda = 0.022$), using Hamiltonian dynamics and learned priors.

Numerical examples: ϕ^4 field theory

$$\rho(x) = \exp\left(-\sum_{i=1}^{d}\left(-2\kappa(x_i x_{i-L} + x_i x_{i+1}) + (1-2\lambda)x_i^2 + \lambda x_i^4\right)\right)$$

- Difficulty depends non-trivially on the choices of κ and λ.
- Preliminary results, not incorporating any symmetries (with $\lambda = 0.022$), using Hamiltonian dynamics and learned priors.





κ	dimension	$\Delta \log Z \downarrow$	ESS $\uparrow$
0.2	16×8	0.0002	0.91
	64×8	0.0057	0.08
0.5	16×8	0.0176	0.79
	64×8	0.0169	0.05

Conclusion

- We have established optimal control, PDE and path space perspectives on generative modeling.

Conclusion

- We have established optimal control, PDE and path space perspectives on generative modeling.
- This allows to carry over respective methods and theory to generative modeling.

Conclusion

- We have established optimal control, PDE and path space perspectives on generative modeling.
- This allows to carry over respective methods and theory to generative modeling.
- We introduced algorithms to sample from an (unnormalized) density, which are already competitive to MCMC/SMC.

Conclusion

- We have established optimal control, PDE and path space perspectives on generative modeling.
- This allows to carry over respective methods and theory to generative modeling.
- We introduced algorithms to sample from an (unnormalized) density, which are already competitive to MCMC/SMC.
- The log-variance divergence outperforms the KL divergence.

Conclusion

- We have established optimal control, PDE and path space perspectives on generative modeling.
- This allows to carry over respective methods and theory to generative modeling.
- We introduced algorithms to sample from an (unnormalized) density, which are already competitive to MCMC/SMC.
- The log-variance divergence outperforms the KL divergence.
- PINNs seem to be suitable for learning dynamical systems for sampling.

Conclusion

- We have established optimal control, PDE and path space perspectives on generative modeling.
- This allows to carry over respective methods and theory to generative modeling.
- We introduced algorithms to sample from an (unnormalized) density, which are already competitive to MCMC/SMC.
- The log-variance divergence outperforms the KL divergence.
- PINNs seem to be suitable for learning dynamical systems for sampling.
- Often, non-uniqueness helps to find a “better” solution.

Outlook

- **General framework:** (stochastic) normalizing flows and GFlowNets can be incorporated, however, continuous-time perspective allows for more flexibility.

Outlook

- **General framework:** (stochastic) normalizing flows and GFlowNets can be incorporated, however, continuous-time perspective allows for more flexibility.
- **SMC:** Annealed importance sampling and resampling can be naturally integrated. (Diffusion model version of CRAFT.)

Outlook

- **General framework:** (stochastic) normalizing flows and GFlowNets can be incorporated, however, continuous-time perspective allows for more flexibility.
- **SMC:** Annealed importance sampling and resampling can be naturally integrated. (Diffusion model version of CRAFT.)
- **Hamiltonian dynamics:** underdamped versions can be considered and lead to improved performance.

Thank you for your attention!

richter@zib.de

References:

- N. Nüsken, L.R. *Solving high-dimensional Hamilton–Jacobi–Bellman PDEs using neural networks: perspectives from the theory of controlled diffusions and measures on path space*. Partial Differential Equations and Applications, 2021.
- J. Berner, L.R., K. Ullrich. *An optimal control perspective on diffusion-based generative modeling*. TMLR, 2024.
- L.R., J. Berner. *Improved sampling via learned diffusions*. ICLR, 2024.
- J. Sun, J. Berner, L.R. et al. *Dynamical measure transport and neural PDE solvers for sampling*. arXiv:2407.07873, 2024.