An optimal control perspective on diffusion-based generative modeling

Julius Berner

Faculty of Mathematics, University of Vienna

December, 2022

Collaborators

Lorenz Richter
Zuse Institute Berlin, dida Datenschmiede GmbH

Karen Ullrich
Meta AI

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** $Y^{(i)} \sim \mathcal{D}$ (images, text, sound, ...).

Fig. 1: https://en.m.wikipedia.org/wiki/File:Cat_poster_1.jpg

Generative modeling

Task

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

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

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

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

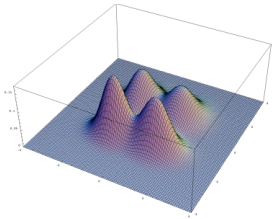


Fig. 2: https://en.wikipedia.org/wiki/File:Bimodal-bivariate-small.png

Fig. 1: https://en.m.wikipedia.org/wiki/File:Cat_poster_1.jpg

Overview of generative models

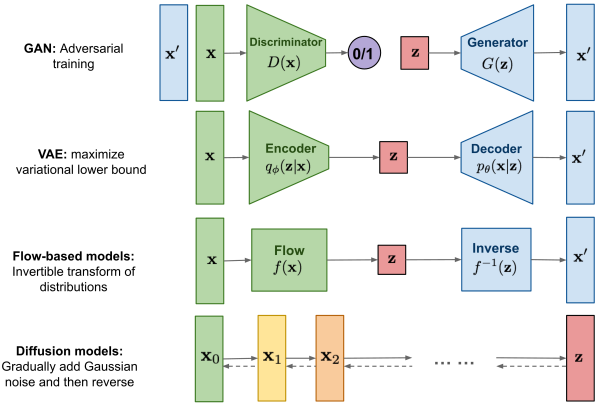
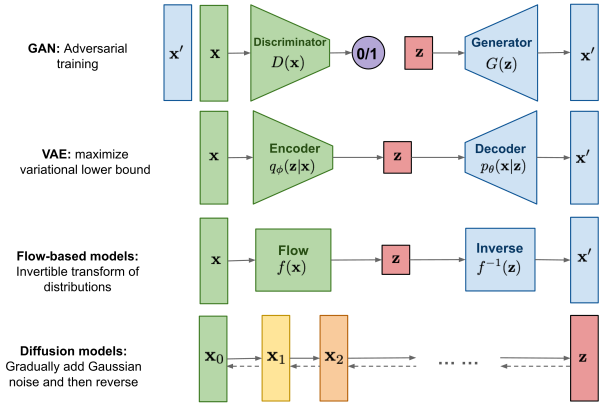


Fig.: https://lilianweng.github.io/posts/2021-07-11-diffusion-models

Overview of generative models



This talk: **Diffusion models** (originated from (denoising) diffusion probabilistic modeling and score matching with Langevin dynamics).

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015

Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020

Fig.: https://lilianweng.github.io/posts/2021-07-11-diffusion-models

Diffusion models

- **State-of-the-art** models for a variety of applications where **samples** are given.

Sampling conditioned on the text prompt "a photograph of an astronaut riding a horse" using the stable diffusion model.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2021

Diffusion models

- **State-of-the-art** models for a variety of applications where **samples** are given.
- Can be understood via **stochastic differential equations** (SDEs).

Sampling conditioned on the text prompt "a photograph of an astronaut riding a horse" using the stable diffusion model.

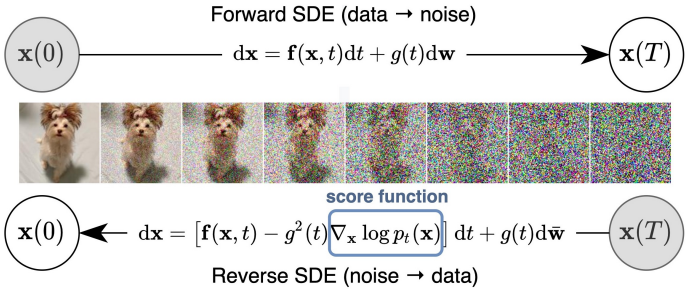
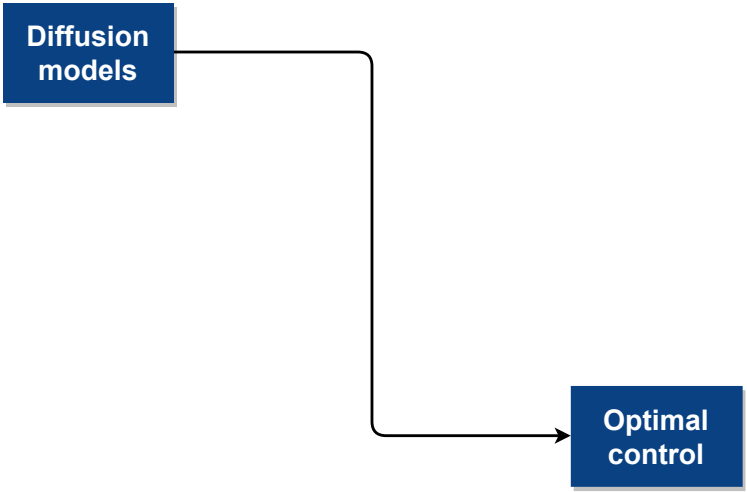


Fig.: https://yang-song.net/blog/2021/score

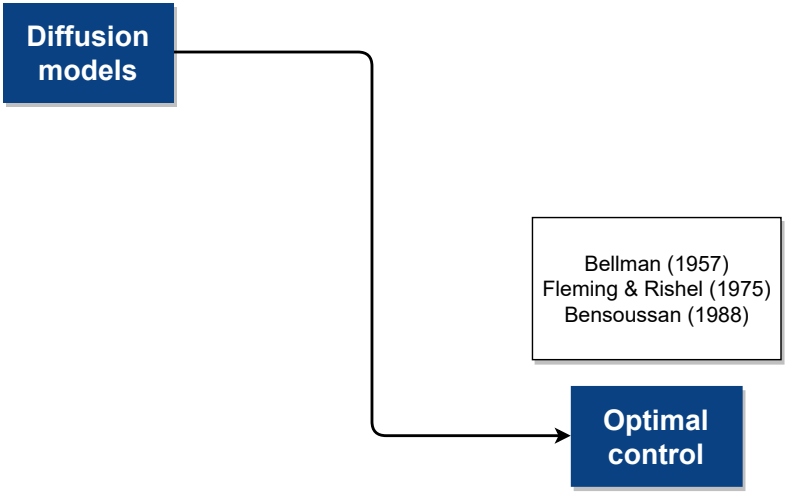
Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2021

Overview of the talk



Overview of the talk

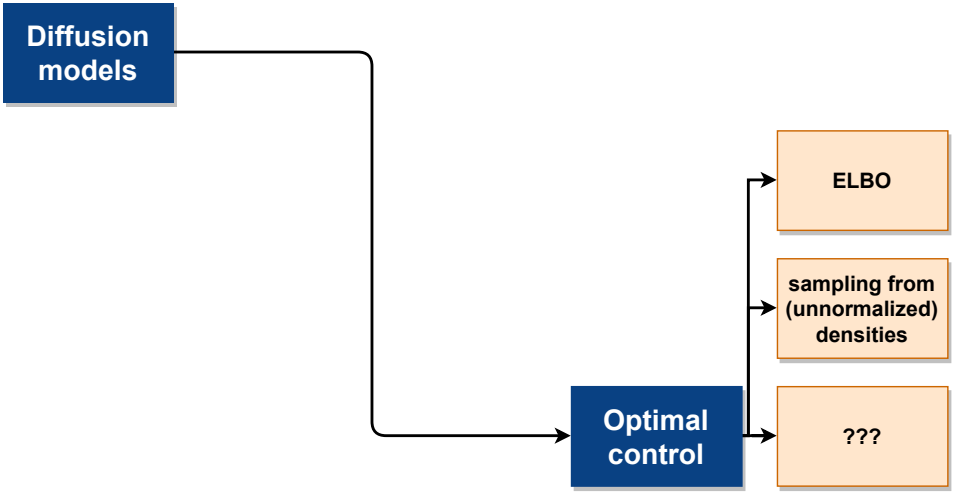


R. Bellman. *Dynamic programming*. Princeton University Press, 1957

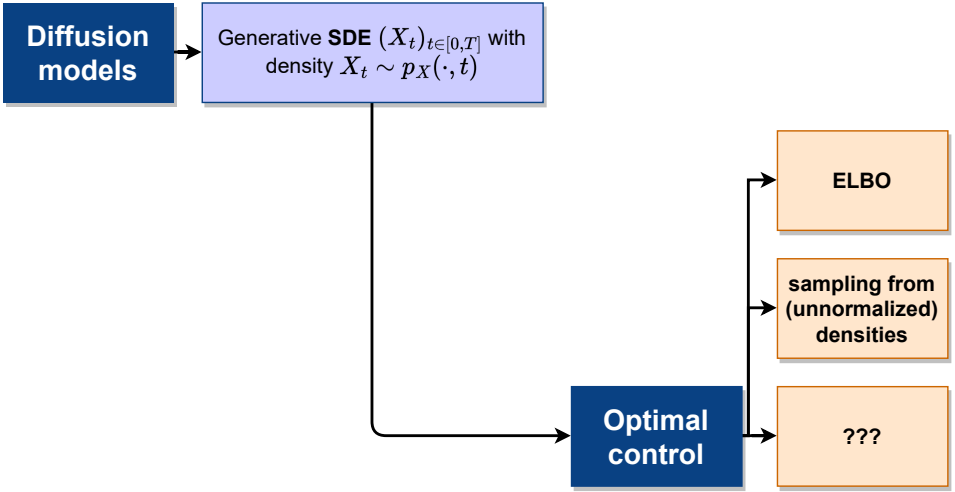
W. Fleming and R. Rishel. *Deterministic and Stochastic Optimal Control*. Applications of mathematics. Springer, 1975

A. Bensoussan. *Perturbation methods in optimal control*. Wiley, 1988

Overview of the talk



Overview of the talk



SDE-based generative modeling

- **Model:** Stochastic differential equation (SDE)

$$\mathrm{d}X_s = \underbrace{\tilde{\mu}(X_s, s)}_{\text{drift}}\,\mathrm{d}s + \underbrace{\tilde{\sigma}(s)}_{\text{diffusion}}\,\mathrm{d}\widehat{B_s}$$

$\widehat{B_s}$: d-dim. Brownian motion

SDE-based generative modeling

- **Model:** Stochastic differential equation (SDE)

$$\mathrm{d}X_s = \underbrace{\bar{\mu}(X_s, s)}_{\text{drift}}\,\mathrm{d}s + \underbrace{\bar{\sigma}(s)}_{\text{diffusion}}\,\mathrm{d}\widehat{B_s}$$

d-dim. Brownian motion

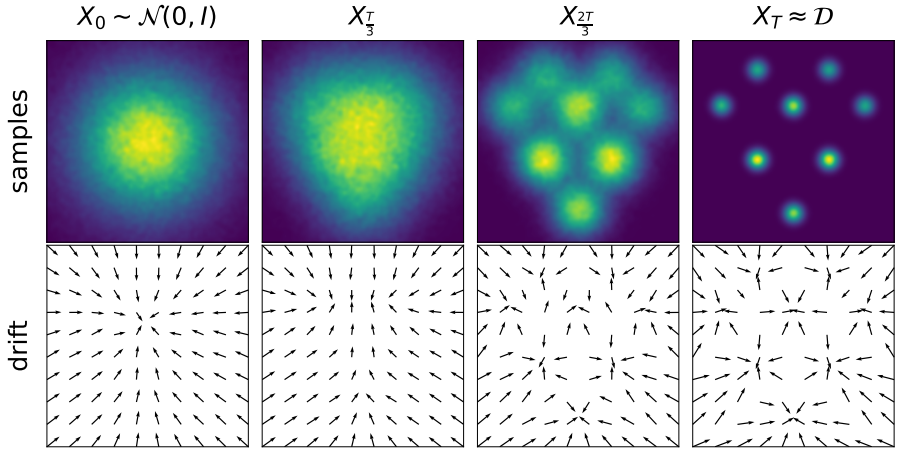
- **Simulation:** For instance, via Euler-Maruyama scheme $\hat{X}_{t_k} \approx X_{t_k}$:

$$\hat{X}_{t_{k+1}} = \hat{X}_{t_k} + \bar{\mu}(\hat{X}_{t_k}, t_k)(t_{k+1} - t_k) + \bar{\sigma}(t_k)\underbrace{(B_{t_{k+1}} - B_{t_k})}_{\sim\mathcal{N}(0, t_{k+1}-t_k)}.$$

SDE solution and Euler-Maruyama scheme with $t_k = \frac{kT}{N}$ and $N = 4, 8$.

SDE-based generative modeling

Goal: Learn the drift $\tilde{\mu}$ such that X_T approximates a given distribution $\mathcal{D}$.



Time-reversed SDE

Goal: $X_T \sim \mathcal{D}$ with **generative SDE**

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

E. Nelson. Dynamical theories of Brownian motion. *Press, Princeton, NJ*, 1967

B. D. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982

U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *The Annals of Probability*, pages 1188–1205, 1986

Time-reversed SDE

Goal: $X_T \sim \mathcal{D}$ with **generative SDE**

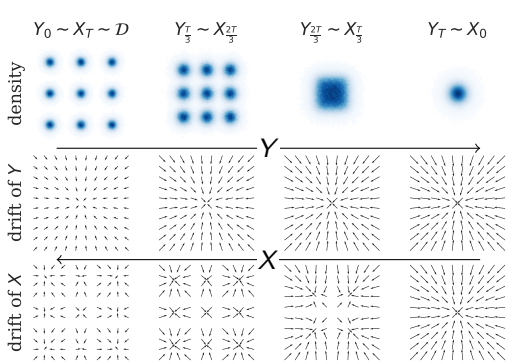
$$\mathrm{d}X_s = \bar{\mu}(X_s, s)\,\mathrm{d}s + \bar{\sigma}(s)\,\mathrm{d}B_s.$$

"Solution": Define the **inference SDE**

$$\mathrm{d}Y_s = f(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s, \;\; Y_0 \sim \mathcal{D},$$

and set $X_0 \sim Y_T$ and $\mu := \sigma\sigma^\top \nabla \log p_Y - f$.

$\Rightarrow \; p_X \overset{\text{a.e.}}{=} \overleftarrow{p}_Y$, i.e., $X_t \sim Y_{T-t}$ for $t \in [0, T]$.



Notation: $\bar{\mu}(x, t) := \mu(x, T - t)$ (time-reversal)

E. Nelson. Dynamical theories of Brownian motion. *Press, Princeton, NJ*, 1967

B. D. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982

U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *The Annals of Probability*, pages 1188–1205, 1986

Time-reversed SDE

Goal: $X_T \sim \mathcal{D}$ with **generative SDE**

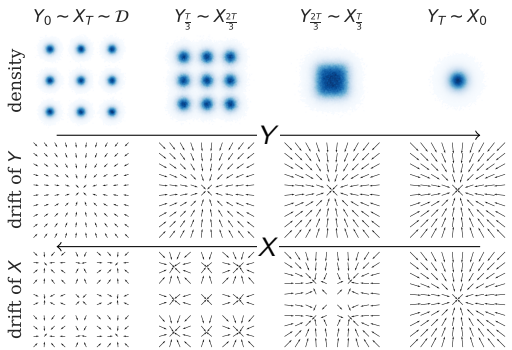
$$\mathrm{d}X_s = \overleftarrow{\mu}(X_s, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s.$$

"Solution": Define the **inference SDE**

$$\mathrm{d}Y_s = f(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s, \quad Y_0 \sim \mathcal{D},$$

and set $X_0 \sim Y_T$ and $\mu \coloneqq \sigma\sigma^\top \nabla \log p_Y - f$.

$\Rightarrow\ p_X \overset{\text{a.e.}}{=} \overleftarrow{p}_Y$, i.e., $X_t \sim Y_{T-t}$ for $t \in [0, T]$.



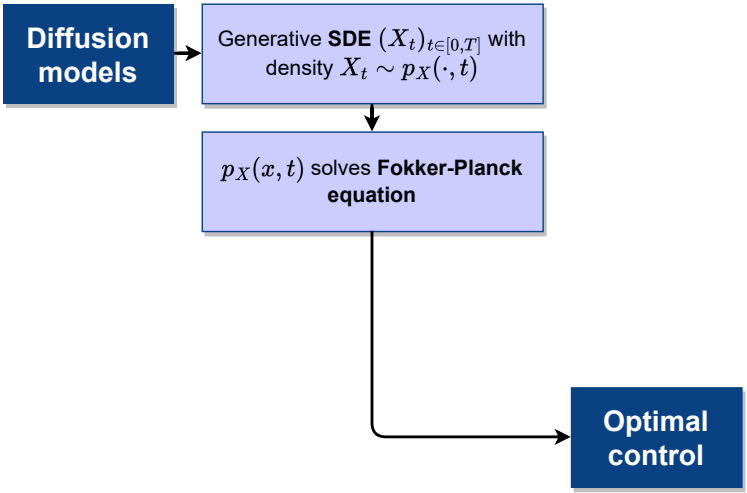
Problem: How to learn the **score** $\nabla \log p_Y$ and how to sample from $X_0 \sim Y_T$?

E. Nelson. Dynamical theories of Brownian motion. *Press, Princeton, NJ*, 1967

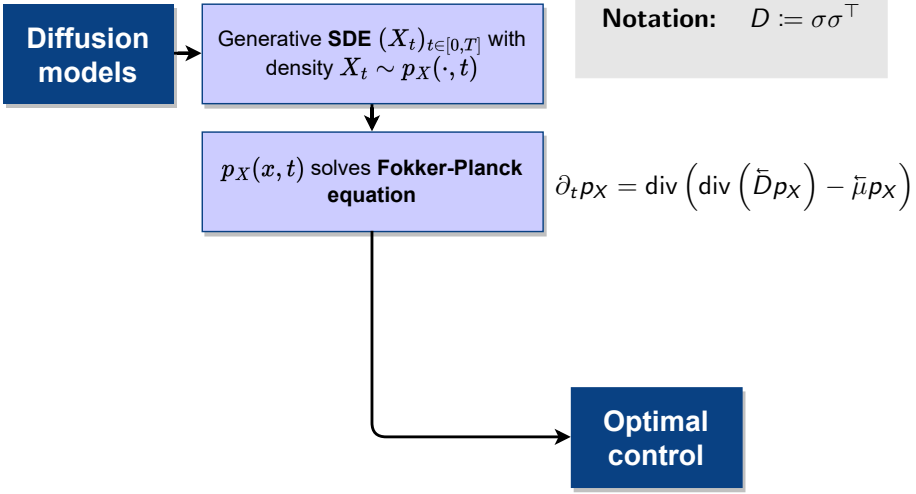
B. D. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982

U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *The Annals of Probability*, pages 1188–1205, 1986

PDE for the log-density

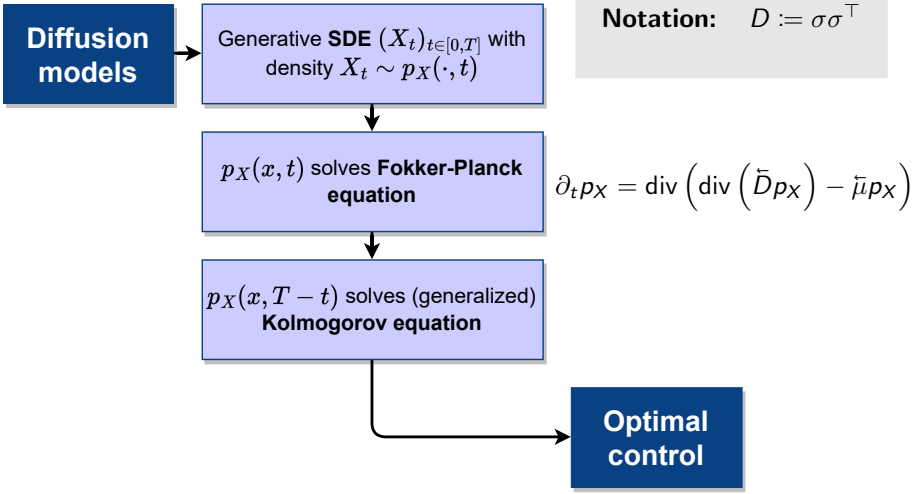


PDE for the log-density



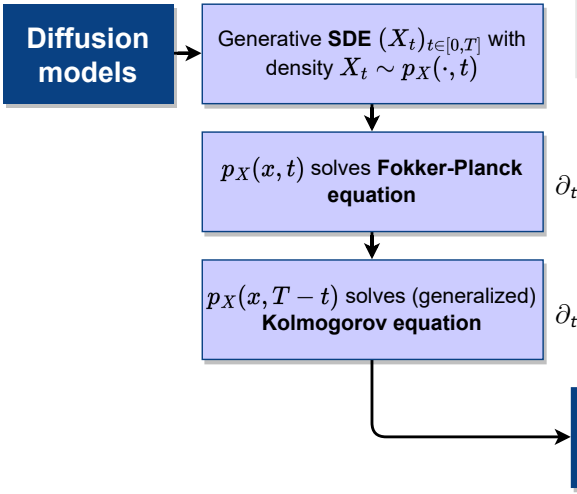
Notation: $D := \sigma\sigma^\top$

PDE for the log-density



Notation: $D \coloneqq \sigma\sigma^\top$

PDE for the log-density

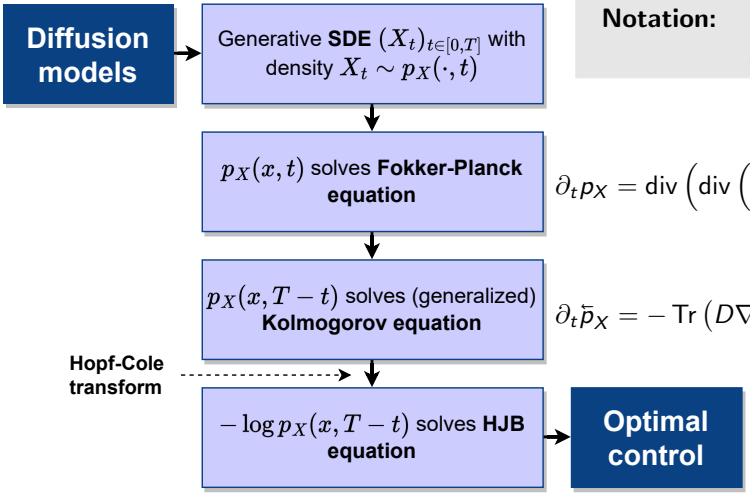


Notation: $D := \sigma\sigma^\top$
$\bar{p}_X(x,t) := p_X(x, T-t)$

$$\partial_t p_X = \operatorname{div}\left(\operatorname{div}\left(\bar{D} p_X\right) - \bar{\mu} p_X\right)$$

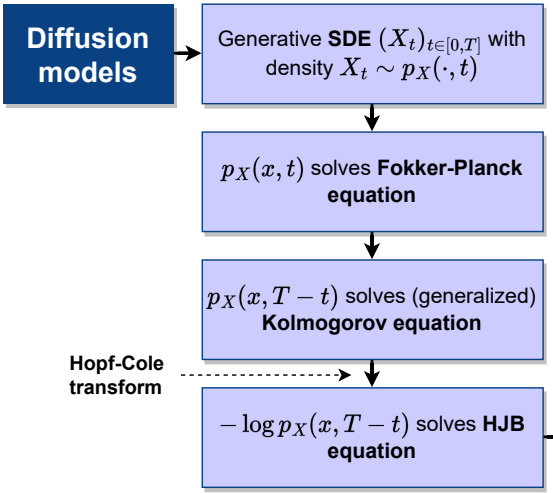
$$\partial_t \bar{p}_X = -\operatorname{Tr}\left(D\nabla^2 \bar{p}_X\right) + \mu \cdot \nabla \bar{p}_X + \operatorname{div}(\mu)\bar{p}_X$$

PDE for the log-density



Notation: $D := \sigma\sigma^\top$
$\bar{p}_X(x,t) := p_X(x, T-t)$

PDE for the log-density



Notation: $D := \sigma\sigma^\top$
$\bar{p}_X(x,t) := p_X(x, T-t)$

$$\partial_t V = -\operatorname{Tr}\left(D\nabla^2 V\right) + \mu \cdot \nabla V - \operatorname{div}(\mu) + \tfrac{1}{2}\left\|\sigma^\top \nabla V\right\|^2 \text{ where } V := -\log \bar{p}_X$$

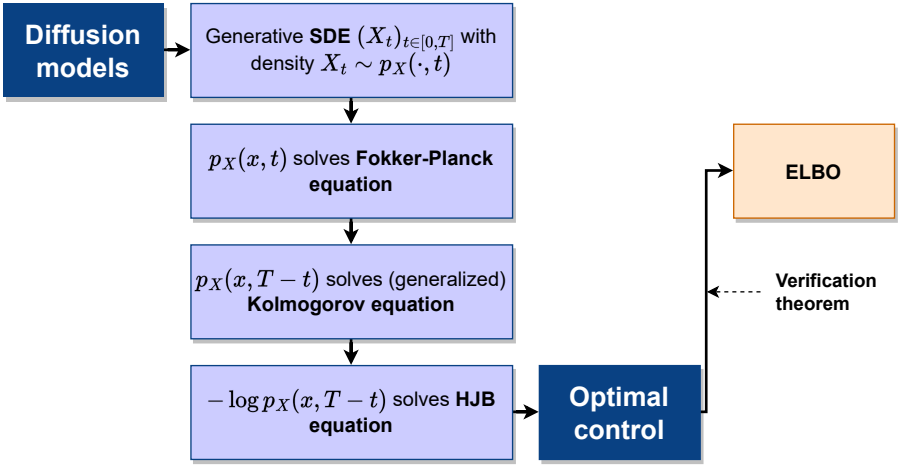
PDE for the log-density

Hamilton-Jacobi-Bellman equation for time-reversed log-density

Define $V \coloneqq -\log \overleftarrow{p}_X$ (Hopf-Cole transform), where $\overleftarrow{p}_X(x,t) \coloneqq p_X(x, T-t)$ denotes the time-reversed density of the generative SDE X. Then

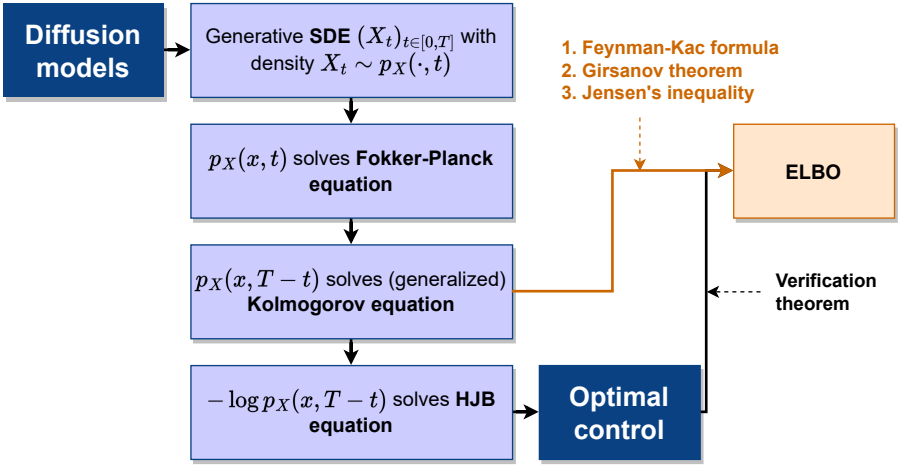
$$\partial_t V = -\operatorname{Tr}\left(D\nabla^2 V\right) + \mu \cdot \nabla V - \operatorname{div}(\mu) + \tfrac{1}{2}\left\|\sigma^\top \nabla V\right\|^2.$$

Deriving the ELBO



$$\partial_t V = -\operatorname{Tr}\left(D\nabla^2 V\right) + \mu \cdot \nabla V - \operatorname{div}(\mu) + \tfrac{1}{2}\left\|\sigma^\top \nabla V\right\|^2 \text{ where } V \coloneqq -\log \bar{p}_X$$

Deriving the ELBO



$$\partial_t V = -\operatorname{Tr}\left(D\nabla^2 V\right) + \mu \cdot \nabla V - \operatorname{div}(\mu) + \tfrac{1}{2}\left\|\sigma^\top \nabla V\right\|^2 \text{ where } V \coloneqq -\log \bar{p}_X$$

C.-W. Huang, J. H. Lim, and A. C. Courville. A variational perspective on diffusion-based generative models and score matching. *Advances in Neural Information Processing Systems*, 34, 2021

Verification theorem

Let V be the solution to the HJB equation

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

let $\mathcal{U}$ be a suitable space of controls, and let

$$\mathrm{d}Y_s^u = (\sigma u - \mu)\,(Y_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s,$$

be the controlled SDEs. Then it holds that

$$\mathbb{E}\left[V(Y_0^u, 0)\right] = \min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \left(\tfrac{1}{2}\|u\|^2 + h\right)(Y_s^u, s)\,\mathrm{d}s + V(Y_T^u, T)\right].$$

M. Pavon. Stochastic control and nonequilibrium thermodynamical systems. *Applied Mathematics and Optimization*, 19(1):187–202, 1989

P. Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991

Verification theorem

Let V be the solution to the HJB equation

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

let $\mathcal{U}$ be a suitable space of controls, and let

$$\mathrm{d}Y_s^u = \left(\sigma u - \mu\right)\left(Y_s^u, s\right)\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s,$$

be the controlled SDEs. Then it holds that

$$\mathbb{E}\left[V(Y_0^u, 0)\right] \underset{\text{Itô's lemma}}{=} \min_{u\in\mathcal{U}} \mathbb{E}\left[\underbrace{\int_0^T \left(\tfrac{1}{2}\|u\|^2 + h\right)\left(Y_s^u, s\right)\mathrm{d}s}_{\text{running costs}} + \underbrace{V(Y_T^u, T)}_{\text{terminal cost}}\right].$$

M. Pavon. Stochastic control and nonequilibrium thermodynamical systems. *Applied Mathematics and Optimization*, 19(1):187–202, 1989

P. Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991

Verification theorem

Let V be the solution to the HJB equation

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

let $\mathcal{U}$ be a suitable space of controls, and let

$$\mathrm{d}Y_s^u = (\sigma u - \mu)\,(Y_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s,$$

be the controlled SDEs. Then it holds that

$$\mathbb{E}\left[V(Y_0^u, 0)\right] = \min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \left(\tfrac{1}{2}\|u\|^2 + h\right)(Y_s^u, s)\,\mathrm{d}s + V(Y_T^u, T)\right],$$

where the unique minimum is attained by $u^* := -\sigma^\top \nabla V$.

M. Pavon. Stochastic control and nonequilibrium thermodynamical systems. *Applied Mathematics and Optimization*, 19(1):187–202, 1989

P. Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991

Verification theorem

Verification theorem

Let V be the solution to the HJB equation

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

let $\mathcal{U}$ be a suitable space of controls, and let

$$\mathrm{d}Y_s^u = \left(\sigma u - \mu\right)\left(Y_s^u, s\right)\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s,$$

be the controlled SDEs. Then it holds that

$$\mathbb{E}\left[V(Y_0^u, 0)\right] = \mathbb{E}\left[\int_0^T \left(\tfrac{1}{2}\|u\|^2 + h\right)\left(Y_s^u, s\right)\mathrm{d}s + V(Y_T^u, T)\right] - \underbrace{D_{\mathrm{KL}}(\mathbb{P}_{Y^u} | \mathbb{P}_{Y^{u^*}})}_{\text{KL-divergence on path space}},$$

where $u^* := -\sigma^\top \nabla V$.

M. Pavon. Stochastic control and nonequilibrium thermodynamical systems. *Applied Mathematics and Optimization*, 19(1):187–202, 1989

P. Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991

Verification theorem

Let V be the solution to the HJB equation

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

let $\mathcal{U}$ be a suitable space of controls, and let

$$\mathrm{d}Y_s^u = (\sigma u - \mu)(Y_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s,$$

be the controlled SDEs. Then it holds that

$$\mathbb{E}\left[V(Y_0^u, 0)\right] = \mathbb{E}\left[\int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\mu)\right)(Y_s^u, s)\,\mathrm{d}s + V(Y_T^u, T)\right] - D_{\mathrm{KL}}(\mathbb{P}_{Y^u} | \mathbb{P}_{Y^{u^*}}),$$

where $u^* := -\sigma^\top \nabla V$.

M. Pavon. Stochastic control and nonequilibrium thermodynamical systems. *Applied Mathematics and Optimization*, 19(1):187–202, 1989

P. Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991

Verification theorem

ELBO

Let $V := -\log \bar{p}_X$ (**time-reversed log-density**), i.e., the solution to the HJB equation

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

let $\mathcal{U}$ be a suitable space of controls, and let

$$\mathrm{d}Y_s^u = (\sigma u - \mu)\,(Y_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s,$$

be the controlled SDEs. Then it holds that

$$\mathbb{E}\left[\log p_{X_T}(Y_0^u)\right] = \mathbb{E}\left[-\int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\mu)\right)(Y_s^u, s)\,\mathrm{d}s + \log p_{X_0}(Y_T^u)\right] + D_{\mathrm{KL}}(\mathbb{P}_{Y^u} | \mathbb{P}_{Y^{u^*}}),$$

where $u^* := \sigma^\top \nabla \log \bar{p}_X$ (scaled **score**).

M. Pavon. Stochastic control and nonequilibrium thermodynamical systems. *Applied Mathematics and Optimization*, 19(1):187–202, 1989

P. Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991

Verification theorem

ELBO

Let $V := -\log \bar{p}_X$ (**time-reversed log-density**), i.e., the solution to the HJB equation

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

let $\mathcal{U}$ be a suitable space of controls, and let

$$\mathrm{d}Y_s^u = (\sigma u - \mu)(Y_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s,$$

be the controlled SDEs. Then it holds that

$$\underbrace{\mathbb{E}\left[\log p_{X_T}(Y_0^u)\right]}_{\text{log-likelihood}} = \mathbb{E}\left[-\int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\mu)\right)(Y_s^u, s)\,\mathrm{d}s + \log p_{X_0}(Y_T^u)\right] + \underbrace{D_{\mathrm{KL}}(\mathbb{P}_{Y^u} | \mathbb{P}_{Y^{u^*}})}_{\text{variational gap}},$$

where $u^* := \sigma^\top \nabla \log \bar{p}_X$ (scaled **score**).

M. Pavon. Stochastic control and nonequilibrium thermodynamical systems. *Applied Mathematics and Optimization*, 19(1):187–202, 1989

P. Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991

Deriving the ELBO

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

SDEs: $\mathrm{d}Y_s^u = (\sigma u - \mu)(Y_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s, \quad \mathrm{d}X_s = \overleftarrow{\mu}(X_s, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s$

Deriving the ELBO

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

SDEs: $\mathrm{d}Y_s^u = (\sigma u - \mu)(Y_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s, \quad \mathrm{d}X_s = \overleftarrow{\mu}(X_s, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s$

- **Problem:** We want to control the generative SDE X.

Deriving the ELBO

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

SDEs: $\mathrm{d}Y_s^u = (\sigma u - \mu)(Y_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s, \quad \mathrm{d}X_s = \overleftarrow{\mu}(X_s, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s$

- **Problem:** We want to control the generative SDE X.

$\Rightarrow$ **Reparametrization:** $\mu := \sigma u - f$

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

SDEs: $\mathrm{d}Y_s = f(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s, \qquad \mathrm{d}X_s^u = \left(\overleftarrow{\sigma}\overleftarrow{u} - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s$

C.-W. Huang, J. H. Lim, and A. C. Courville. A variational perspective on diffusion-based generative models and score matching. *Advances in Neural Information Processing Systems*, 34, 2021

Score-based generative modeling

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

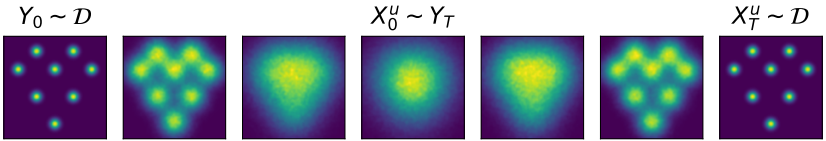
SDEs: $\underbrace{\mathrm{d}Y_s = f(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s,\ Y_0 \sim \mathcal{D}}_{\text{inference SDE}},\ \underbrace{\mathrm{d}X_s^u = \left(\overleftarrow{\sigma}\overleftarrow{u} - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s}_{\text{generative SDE}}$

Score-based generative modeling

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

SDEs: $\underbrace{\mathrm{d}Y_s = f(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s,\ Y_0 \sim \mathcal{D}}_{\text{inference SDE}},\ \underbrace{\mathrm{d}X_s^u = \left(\overleftarrow{\sigma}\overleftarrow{u} - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s}_{\text{generative SDE}}$

Time-reversal: If $X_0^u \sim Y_T$, then $u = u^* = \sigma^\top \nabla \log p_Y$ and $X_T^u \sim \mathcal{D}$.

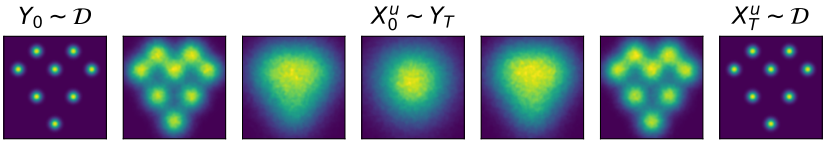


Score-based generative modeling

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

SDEs: $\underbrace{\mathrm{d}Y_s = f(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s,\ Y_0 \sim \mathcal{D}}_{\text{inference SDE}},\ \underbrace{\mathrm{d}X_s^u = \left(\overleftarrow{\sigma}\overleftarrow{u} - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s}_{\text{generative SDE}}$

Time-reversal: If $X_0^u \sim Y_T$, then $u = u^* = \sigma^\top \nabla \log p_Y$ and $X_T^u \sim \mathcal{D}$.



Questions: 1. Can the ELBO be simplified? 2. How to sample from $X_0^u \sim Y_T$?

Denoising score matching

1. Rewrite the divergence and use a MC approximation, $\tau \sim \mathcal{U}([0, T])$, of the time-integral:

Denoising score matching

1. Rewrite the divergence and use a MC approximation, $\tau \sim \mathcal{U}([0, T])$, of the time-integral:

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

P. Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011

C.-W. Huang, J. H. Lim, and A. C. Courville. A variational perspective on diffusion-based generative models and score matching. *Advances in Neural Information Processing Systems*, 34, 2021

Denoising score matching

1. Rewrite the divergence and use a MC approximation, $\tau \sim \mathcal{U}([0, T])$, of the time-integral:

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

const. ↓ ↓ const. ↓

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

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

P. Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011

C.-W. Huang, J. H. Lim, and A. C. Courville. A variational perspective on diffusion-based generative models and score matching. *Advances in Neural Information Processing Systems*, 34, 2021

Ornstein–Uhlenbeck process

2. Construct drift f and diffusion σ of Y such that $Y_T \approx \mathcal{N}(0, \mathrm{I})$:

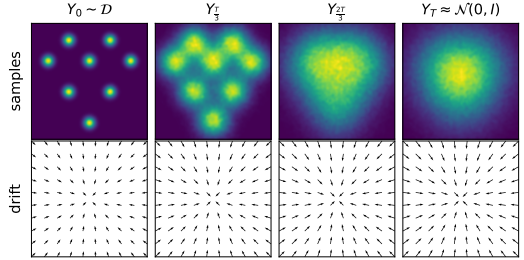
Ornstein–Uhlenbeck process

2. Construct drift f and diffusion σ of Y such that $Y_T \approx \mathcal{N}(0, \mathrm{I})$:

SDEs: $\mathrm{d}X_s^u = \left(\overleftarrow{\sigma}\overleftarrow{u} - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s, \quad X_0 \sim \mathcal{N}(0, \mathrm{I})$

$\mathrm{d}Y_s = f(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s, \quad Y_0 \sim \mathcal{D}$

f: $= b(s)Y_s$ ↑ — σ: diagonal matrix ↑



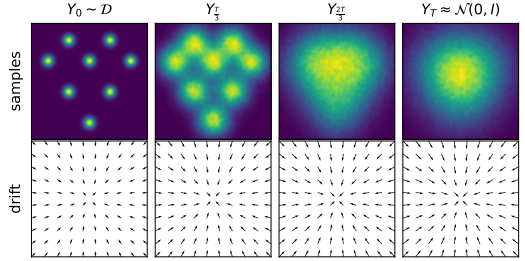
Ornstein–Uhlenbeck process

2. Construct drift f and diffusion σ of Y such that $Y_T \approx \mathcal{N}(0, \mathrm{I})$:

SDEs: $\mathrm{d}X_s^u = \left(\overleftarrow{\sigma}\overleftarrow{u} - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s, \quad X_0 \sim \mathcal{N}(0, \mathrm{I})$

$\mathrm{d}Y_s = f(Y_s, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s, \quad Y_0 \sim \mathcal{D}$

f: $\uparrow$ $= b(s)Y_s$; σ: $\uparrow$ diagonal matrix



Example: **Variance-preserving SDE** (Ornstein-Uhlenbeck process) given by $b = -\frac{1}{2}\sigma^2$ such that

$$Y_t | Y_0 \sim \mathcal{N}\left(e^{-\alpha(t)} Y_0, 1 - e^{-2\alpha(t)}\right),$$

where $\alpha(t) := \frac{1}{2}\int_0^t \sigma^2(s)\mathrm{d}s$.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020

Summary

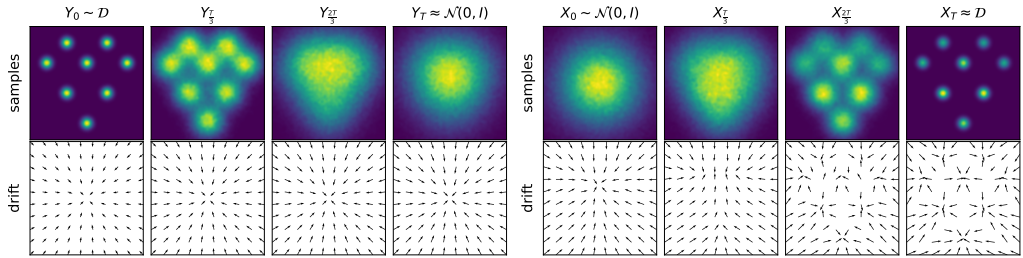
variance-preserving SDE with $Y_T \approx \mathcal{N}(0, \mathrm{I})$
for sufficiently large $\alpha(T) := \frac{1}{2}\int_0^T \sigma^2(s)\mathrm{d}s$

SDEs:

$$\mathrm{d}Y_s = -\tfrac{1}{2}\sigma^2(s) Y_s \,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s, \quad Y_0 \sim \mathcal{D}$$

$$\mathrm{d}X_s^u = \left(\overleftarrow{\sigma}\overleftarrow{u} - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s, \quad X_0 \sim \mathcal{N}(0, \mathrm{I})$$

generative SDE



Summary

variance-preserving SDE with $Y_T \approx \mathcal{N}(0, \mathrm{I})$
for sufficiently large $\alpha(T) := \frac{1}{2}\int_0^T \sigma^2(s)\mathrm{d}s$

SDEs:
$$\overbrace{\mathrm{d}Y_s = -\tfrac{1}{2}\sigma^2(s) Y_s \,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s}, \quad Y_0 \sim \mathcal{D}$$
$$\underbrace{\mathrm{d}X_s^u = \left(\overleftarrow{\sigma}\overleftarrow{u} - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s}_{\text{generative SDE}}, \quad X_0 \sim \mathcal{N}(0, \mathrm{I})$$

ELBO:
$$\overbrace{\mathbb{E}\left[\log p_{X_T^u}(Y_0)\right]}^{\text{log-likelihood}} \geq \overbrace{\mathbb{E}\left[\left\| u(Y_\tau, \tau) - \sigma(\tau)\nabla \log \underbrace{p_{Y_\tau|Y_0}(Y_\tau|Y_0)}_{= \mathcal{N}\left(e^{-\alpha(t)}Y_0,\, 1 - e^{-2\alpha(t)}\right)} \right\|^2\right]}^{\text{denoising score matching}}, \quad \tau \sim \mathcal{U}([0, T])$$

Summary

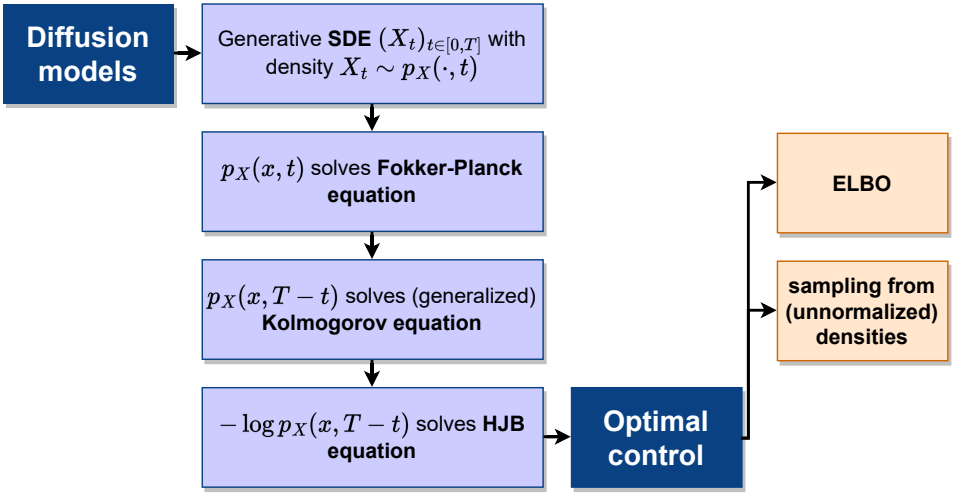
variance-preserving SDE with $Y_T \approx \mathcal{N}(0, \mathrm{I})$
for sufficiently large $\alpha(T) := \frac{1}{2}\int_0^T \sigma^2(s)\mathrm{d}s$

SDEs:
$$\overbrace{\mathrm{d}Y_s = -\tfrac{1}{2}\sigma^2(s) Y_s \,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s}, \quad Y_0 \sim \mathcal{D}$$
$$\underbrace{\mathrm{d}X_s^u = \left(\overleftarrow{\sigma}\overleftarrow{u} - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s}, \quad X_0 \sim \mathcal{N}(0, \mathrm{I})$$
generative SDE

ELBO:
$$\overbrace{\mathbb{E}\left[\log p_{X_T^u}(Y_0)\right]}^{\text{log-likelihood}} \geq \overbrace{\mathbb{E}\left[\left\| u(Y_\tau, \tau) - \sigma(\tau)\nabla \log \underbrace{p_{Y_\tau|Y_0}(Y_\tau|Y_0)}_{= \mathcal{N}\left(e^{-\alpha(t)}Y_0,\, 1 - e^{-2\alpha(t)}\right)} \right\|^2\right]}^{\text{denoising score matching}}, \quad \tau \sim \mathcal{U}([0, T])$$

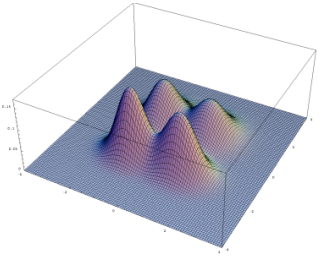
$\Rightarrow$ We arrived at (a variant of) the objective used to train diffusion models in practice!

Sampling from densities



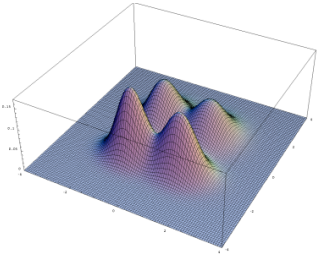
Sampling from densities

- **Setting:** Data distribution $\mathcal{D}$ admits the density $\frac{\rho}{\mathcal{Z}}$, where ρ is known but the normalizing constant $\mathcal{Z} \coloneqq \int_{\mathbb{R}^d} \rho(x)\,\mathrm{d}x$ might be intractable.



Sampling from densities

- **Setting:** Data distribution $\mathcal{D}$ admits the density $\frac{\rho}{\mathcal{Z}}$, where ρ is known but the normalizing constant $\mathcal{Z} \coloneqq \int_{\mathbb{R}^d} \rho(x)\,\mathrm{d}x$ might be intractable.



- **Problem:** Cannot use the previous ELBO as we cannot simulate Y starting at $Y_0 \sim \mathcal{D}$:

$$\mathbb{E}\left[\log p_{X_T^u}(Y_0)\right] \geq \mathbb{E}\left[\left\|u(Y_\tau, \tau) - \sigma(\tau)\nabla \log p_{Y_\tau|Y_0}(Y_\tau|Y_0)\right\|^2\right].$$

Sampling from densities

Idea: Use verification theorem on HJB equation for $V := -\log \overleftarrow{p}_Y$ (instead of $-\log \overleftarrow{p}_X$):

$$\partial_t V = -\operatorname{Tr}\left(\overleftarrow{D}\nabla^2 V\right) + \overleftarrow{f} \cdot \nabla V - \operatorname{div}(\overleftarrow{f}) + \tfrac{1}{2}\left\|\overleftarrow{\sigma}^\top \nabla V\right\|^2.$$

Sampling from densities

Idea: Use verification theorem on HJB equation for $V := -\log \overleftarrow{p}_Y$ (instead of $-\log \overleftarrow{p}_X$):

$$\partial_t V = -\operatorname{Tr}\left(\overleftarrow{D}\nabla^2 V\right) + \overleftarrow{f}\cdot\nabla V - \operatorname{div}(\overleftarrow{f}) + \tfrac{1}{2}\left\|\overleftarrow{\sigma}^\top \nabla V\right\|^2.$$

Verification theorem

Define the controlled SDEs

$$\mathrm{d}X_s^u = \left(\overleftarrow{\sigma}u - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s.$$

Then it holds that

$$\mathbb{E}\left[\log p_{Y_T}(X_0^u)\right] = \mathbb{E}\left[\log p_{Y_0}(X_T^u) - \int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\overleftarrow{f})\right)(X_s^u, s)\,\mathrm{d}s\right] + D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{X^{u^*}}).$$

where $u^* = \overleftarrow{\sigma}^\top \nabla \log \overleftarrow{p}_Y$.

Sampling from densities

- Verification theorem:

$$\mathbb{E}\left[\log p_{Y_T}(X_0^u)\right] = \mathbb{E}\left[\log p_{Y_0}(X_T^u) - \int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\overleftarrow{f})\right)(X_s^u, s)\,\mathrm{d}s\right] + D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{X^{u^*}})$$

Sampling from densities

- Verification theorem:

$$\mathbb{E}\left[\log p_{Y_T}(X_0^u)\right] = \mathbb{E}\left[\log p_{Y_0}(X_T^u) - \int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\overleftarrow{f})\right)(X_s^u, s)\,\mathrm{d}s\right] + D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{X^{u^*}})$$

- Using $p_{Y_0} = \frac{\rho}{\mathcal{Z}}$, yields the following **reverse KL objective**:

$$D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{X^{u^*}}) = \mathbb{E}\left[\int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\overleftarrow{f})\right)(X_s^u, s)\,\mathrm{d}s + \log\frac{p_{Y_T}(X_0^u)}{\rho(X_T^u)}\right] + \log\mathcal{Z}$$

Sampling from densities

- Verification theorem:

$$\mathbb{E}\left[\log p_{Y_T}(X_0^u)\right] = \mathbb{E}\left[\log p_{Y_0}(X_T^u) - \int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\overleftarrow{f})\right)(X_s^u, s)\,\mathrm{d}s\right] + D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{X^{u^*}})$$

- Using $p_{Y_0} = \frac{\rho}{\mathcal{Z}}$, yields the following **reverse KL objective**:

$$D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{X^{u^*}}) = \mathbb{E}\left[\int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\overleftarrow{f})\right)(X_s^u, s)\,\mathrm{d}s + \log \frac{\overbrace{p_{Y_T}(X_0^u)}^{\approx \mathcal{N}(X_0^u; 0, \mathrm{I})}}{\rho(X_T^u)}\right] + \underbrace{\log \mathcal{Z}}_{\text{const.}}$$

Sampling from densities

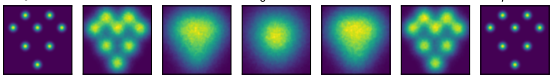
- Verification theorem:

$$\mathbb{E}\left[\log p_{Y_T}(X_0^u)\right] = \mathbb{E}\left[\log p_{Y_0}(X_T^u) - \int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\overleftarrow{f})\right)(X_s^u, s)\,\mathrm{d}s\right] + D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{X^{u^*}})$$

- Using $p_{Y_0} = \frac{\rho}{\mathcal{Z}}$, yields the following **reverse KL objective**:

$$D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{X^{u^*}}) = \mathbb{E}\left[\int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\overleftarrow{f})\right)(X_s^u, s)\,\mathrm{d}s + \log \frac{\overbrace{p_{Y_T}(X_0^u)}^{\approx \mathcal{N}(X_0^u; 0, \mathrm{I})}}{\rho(X_T^u)}\right] + \underbrace{\log \mathcal{Z}}_{\text{const.}}$$

- Remark: $D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{X^{u^*}}) = D_{\mathrm{KL}}(\mathbb{P}_{X^u}|\mathbb{P}_{\overleftarrow{Y}}) - \overbrace{D_{\mathrm{KL}}(\mathbb{P}_{X_0^u}|\mathbb{P}_{Y_T})}^{\text{prior loss}}$ with reverse time SDE $\overleftarrow{Y}$.



Time-Reversed **Di**ffusion **S**ampler

$\Rightarrow$ Diffusion-based method for (unnormalized) densities: Time-Reversed **Di**ffusion **S**ampler

DIS:
$$\min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\overleftarrow{f})\right)(X_s^u, s)\,\mathrm{d}s + \log \frac{\mathcal{N}(X_0^u;\, 0, \mathrm{I})}{\rho(X_T^u)}\right]$$
$$\mathrm{d}X_s^u = \left(\overleftarrow{\sigma} u - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s, \quad X_0^u \sim \mathcal{N}(0, \mathrm{I})$$

Time-Reversed **Di**ffusion **S**ampler

$\Rightarrow$ Diffusion-based method for (unnormalized) densities: Time-Reversed **Di**ffusion **S**ampler

DIS:
$$\min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\overleftarrow{f})\right)(X_s^u, s)\,\mathrm{d}s + \log \frac{\mathcal{N}(X_0^u;\, 0, \mathrm{I})}{\rho(X_T^u)}\right]$$
$$\mathrm{d}X_s^u = \left(\overleftarrow{\sigma} u - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s, \quad X_0^u \sim \mathcal{N}(0, \mathrm{I})$$

- Cannot simplify the objective (as in denoising score matching) since the expectation is over the controlled process X^u.

Time-Reversed **Di**ffusion **S**ampler

$\Rightarrow$ Diffusion-based method for (unnormalized) densities: Time-Reversed **Di**ffusion **S**ampler

DIS:
$$\min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\overleftarrow{f})\right)(X_s^u, s)\,\mathrm{d}s + \log \frac{\mathcal{N}(X_0^u; 0, \mathrm{I})}{\rho(X_T^u)}\right]$$
$$\mathrm{d}X_s^u = \left(\overleftarrow{\sigma} u - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s, \quad X_0^u \sim \mathcal{N}(0, \mathrm{I})$$

- Cannot simplify the objective (as in denoising score matching) since the expectation is over the controlled process X^u.

$\Rightarrow$ Discretize objective, parametrize u as neural network, simulate whole trajectory X^u, and backpropagate (or use stochastic adjoint method).

P. Kidger, J. Foster, X. C. Li, and T. Lyons. Efficient and accurate gradients for neural SDEs. *Advances in Neural Information Processing Systems*, 34:18747–18761, 2021

Time-Reversed **Di**ffusion **S**ampler

$\Rightarrow$ Diffusion-based method for (unnormalized) densities: Time-Reversed **Di**ffusion **S**ampler

DIS:
$$\min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\overleftarrow{f})\right)(X_s^u, s)\,\mathrm{d}s + \log \frac{\mathcal{N}(X_0^u; 0, \mathrm{I})}{\rho(X_T^u)}\right]$$
$$\mathrm{d}X_s^u = \left(\overleftarrow{\sigma} u - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s, \quad X_0^u \sim \mathcal{N}(0, \mathrm{I})$$

- Cannot simplify the objective (as in denoising score matching) since the expectation is over the controlled process X^u.

$\Rightarrow$ Discretize objective, parametrize u as neural network, simulate whole trajectory X^u, and backpropagate (or use stochastic adjoint method).

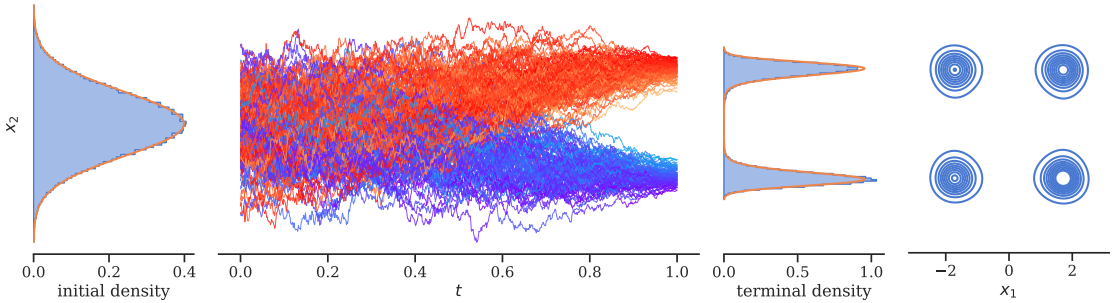
- Initialize neural network as linear interpolation between (scaled) initial and terminal scores $\overleftarrow{\sigma}^\top \nabla \log \mathcal{N}(0, \mathrm{I}) \approx \overleftarrow{\sigma}^\top \nabla \log p_{Y_T}$ and $\overleftarrow{\sigma}^\top \nabla \log \rho = \overleftarrow{\sigma}^\top \nabla \log p_{Y_0}$.

P. Kidger, J. Foster, X. C. Li, and T. Lyons. Efficient and accurate gradients for neural SDEs. *Advances in Neural Information Processing Systems*, 34:18747–18761, 2021

Time-Reversed **Di**ffusion **S**ampler

Double-Well: Sampling from a double-well (DW) with 2^5 modes in 20 dimensions:

$$\rho(x) = \exp\left(-\sum_{i=1}^{w}(x_i^2 - \delta)^2 - \frac{1}{2}\sum_{i=w+1}^{d} x_i^2\right), \quad d = 20,\ \delta = 3,\ w = 5.$$



Path Integral Sampler

- We compare against one of the standard approaches (**P**ath **I**ntegral **S**ampler) originating from the field of stochastic optimal control.

P. Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991

B. Tzen and M. Raginsky. Theoretical guarantees for sampling and inference in generative models with latent diffusions. In *Conference on Learning Theory*, pages 3084–3114. PMLR, 2019

Q. Zhang and Y. Chen. Path integral sampler: a stochastic control approach for sampling. In *International Conference on Learning Representations*, 2022

L. Richter. *Solving high-dimensional PDEs, approximation of path space measures and importance sampling of diffusions*. PhD thesis, BTU Cottbus-Senftenberg, 2021

F. Vargas, A. Ovsianas, D. Fernandes, M. Girolami, N. D. Lawrence, and N. Nüsken. Bayesian learning via neural Schrödinger–Föllmer flows. *Statistics and Computing*, 33(1):1–22, 2023

Path Integral Sampler

- We compare against one of the standard approaches (**P**ath **I**ntegral **S**ampler) originating from the field of stochastic optimal control.
- Based on Schrödinger (Half-)Bridge problem $\min_{u \in \mathcal{U}} D_{\mathrm{KL}}(\mathbb{P}_{X^u} | \mathbb{P}_{X^0})$ s.t. $X_T^u \sim \mathcal{D}$, where

$$\mathrm{d}X_s^u = \big(\sigma u + \mu\big)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s, \quad X_0^u \sim \delta_{x_0}.$$

P. Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991

B. Tzen and M. Raginsky. Theoretical guarantees for sampling and inference in generative models with latent diffusions. In *Conference on Learning Theory*, pages 3084–3114. PMLR, 2019

Q. Zhang and Y. Chen. Path integral sampler: a stochastic control approach for sampling. In *International Conference on Learning Representations*, 2022

L. Richter. *Solving high-dimensional PDEs, approximation of path space measures and importance sampling of diffusions*. PhD thesis, BTU Cottbus-Senftenberg, 2021

F. Vargas, A. Ovsianas, D. Fernandes, M. Girolami, N. D. Lawrence, and N. Nüsken. Bayesian learning via neural Schrödinger–Föllmer flows. *Statistics and Computing*, 33(1):1–22, 2023

Path Integral Sampler

- We compare against one of the standard approaches (**P**ath **I**ntegral **S**ampler) originating from the field of stochastic optimal control.
- Based on Schrödinger (Half-)Bridge problem $\min_{u \in \mathcal{U}} D_{\mathrm{KL}}(\mathbb{P}_{X^u} | \mathbb{P}_{X^0})$ s.t. $X_T^u \sim \mathcal{D}$, where

$$\mathrm{d}X_s^u = \big(\sigma u + \mu\big)(X_s^u, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}B_s, \quad X_0^u \sim \delta_{x_0}.$$

- Leads to the following objective:

$$\textbf{PIS:} \quad \min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \tfrac{1}{2}\|u(X_s^u, s)\|^2\,\mathrm{d}s + \log \frac{p_{X_T^0}(X_T^u)}{\rho(X_T^u)}\right]$$

P. Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991

B. Tzen and M. Raginsky. Theoretical guarantees for sampling and inference in generative models with latent diffusions. In *Conference on Learning Theory*, pages 3084–3114. PMLR, 2019

Q. Zhang and Y. Chen. Path integral sampler: a stochastic control approach for sampling. In *International Conference on Learning Representations*, 2022

L. Richter. *Solving high-dimensional PDEs, approximation of path space measures and importance sampling of diffusions*. PhD thesis, BTU Cottbus-Senftenberg, 2021

F. Vargas, A. Ovsianas, D. Fernandes, M. Girolami, N. D. Lawrence, and N. Nüsken. Bayesian learning via neural Schrödinger–Föllmer flows. *Statistics and Computing*, 33(1):1–22, 2023

Path Integral Sampler

- We compare against one of the standard approaches (**P**ath **I**ntegral **S**ampler) originating from the field of stochastic optimal control.
- Based on Schrödinger (Half-)Bridge problem $\min_{u \in \mathcal{U}} D_{\mathrm{KL}}(\mathbb{P}_{X^u} | \mathbb{P}_{X^0})$ s.t. $X_T^u \sim \mathcal{D}$, where

$$\mathrm{d}X_s^u = \big(\sigma u + \overbrace{\mu}^{=0}\big)(X_s^u, s)\,\mathrm{d}s + \overbrace{\sigma(s)}^{=\frac{1}{\sqrt{T}}}\,\mathrm{d}B_s, \quad X_0^u \sim \delta_{x_0}.$$

- Leads to the following objective:

$$\textbf{PIS:} \quad \min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \tfrac{1}{2}\|u(X_s^u, s)\|^2\,\mathrm{d}s + \log \frac{\overbrace{p_{X_T^0}(X_T^u)}^{\mathcal{N}(X_T^u; 0, \mathrm{I})}}{\rho(X_T^u)}\right]$$

P. Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991

B. Tzen and M. Raginsky. Theoretical guarantees for sampling and inference in generative models with latent diffusions. In *Conference on Learning Theory*, pages 3084–3114. PMLR, 2019

Q. Zhang and Y. Chen. Path integral sampler: a stochastic control approach for sampling. In *International Conference on Learning Representations*, 2022

L. Richter. *Solving high-dimensional PDEs, approximation of path space measures and importance sampling of diffusions*. PhD thesis, BTU Cottbus-Senftenberg, 2021

F. Vargas, A. Ovsianas, D. Fernandes, M. Girolami, N. D. Lawrence, and N. Nüsken. Bayesian learning via neural Schrödinger–Föllmer flows. *Statistics and Computing*, 33(1):1–22, 2023

DIS vs. PIS

DIS:
$$\min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\overleftarrow{f})\right)(X_s^u, s)\,\mathrm{d}s + \log \frac{\mathcal{N}(X_0^u; 0, \mathrm{I})}{\rho(X_T^u)}\right] \leftarrow \approx \text{density of } Y_T$$
$$\mathrm{d}X_s^u = \left(\overleftarrow{\sigma} u - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s, \quad X_0 \sim \mathcal{N}(0, \mathrm{I})$$

PIS:
$$\min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \tfrac{1}{2}\|u(X_s^u, s)\|^2\,\mathrm{d}s + \log \frac{\mathcal{N}(X_T^u; 0, \mathrm{I})}{\rho(X_T^u)}\right] \leftarrow \text{density of } X_T^0$$
$$\mathrm{d}X_s^u = \tfrac{1}{\sqrt{T}} u(X_s^u, s)\,\mathrm{d}s + \tfrac{1}{\sqrt{T}}\,\mathrm{d}B_s, \quad X_0 \sim \delta_{x_0}$$

DIS vs. PIS

DIS: $\min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \left(\frac{1}{2}\|u\|^2 + \operatorname{div}(\overleftarrow{f}) \right)(X_s^u, s)\,\mathrm{d}s + \log \frac{\mathcal{N}(X_0^u; 0, \mathrm{I})}{\rho(X_T^u)} \right]$ $\approx$ density of Y_T

$$\mathrm{d}X_s^u = \left(\overleftarrow{\sigma} u - \overleftarrow{f} \right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s, \quad X_0 \sim \mathcal{N}(0, \mathrm{I})$$

PIS: $\min_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^T \frac{1}{2}\|u(X_s^u, s)\|^2\,\mathrm{d}s + \log \frac{\mathcal{N}(X_T^u; 0, \mathrm{I})}{\rho(X_T^u)} \right]$ density of X_T^0

$$\mathrm{d}X_s^u = \frac{1}{\sqrt{T}} u(X_s^u, s)\,\mathrm{d}s + \frac{1}{\sqrt{T}}\,\mathrm{d}B_s, \quad X_0 \sim \delta_{x_0}$$

$\Rightarrow$ DIS can use different SDEs and initial distributions $X \sim \mathcal{N}(0, \nu^2 \mathrm{I})$

DIS vs. PIS

DIS:
$$\min_{u\in\mathcal{U}} \mathbb{E}\left[\int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\overleftarrow{f})\right)(X_s^u, s)\,\mathrm{d}s + \log\frac{\mathcal{N}(X_0^u; 0, \mathrm{I})}{\rho(X_T^u)}\right] \leftarrow \approx \text{density of } Y_T$$
$$\mathrm{d}X_s^u = \left(\overleftarrow{\sigma} u - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s, \quad X_0 \sim \mathcal{N}(0, \mathrm{I})$$

PIS:
$$\min_{u\in\mathcal{U}} \mathbb{E}\left[\int_0^T \tfrac{1}{2}\|u(X_s^u, s)\|^2\,\mathrm{d}s + \log\frac{\mathcal{N}(X_T^u; 0, \mathrm{I})}{\rho(X_T^u)}\right] \leftarrow \text{density of } X_T^0$$
$$\mathrm{d}X_s^u = \tfrac{1}{\sqrt{T}} u(X_s^u, s)\,\mathrm{d}s + \tfrac{1}{\sqrt{T}}\,\mathrm{d}B_s, \quad X_0 \sim \delta_{x_0}$$

$\Rightarrow$ DIS can use different SDEs and initial distributions $X \sim \mathcal{N}(0, \nu^2\mathrm{I})$

$\Rightarrow$ Allows to transfer techniques for diffusion models (noise schedules, probability flow ODE, exponential integrators, ...)

Q. Zhang and Y. Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022

DIS vs. PIS

DIS:
$$\min_{u\in\mathcal{U}} \mathbb{E}\left[\int_0^T \left(\tfrac{1}{2}\|u\|^2 + \operatorname{div}(\overleftarrow{f})\right)(X_s^u, s)\,\mathrm{d}s + \log\frac{\mathcal{N}(X_0^u; 0, \mathrm{I})}{\rho(X_T^u)}\right] \quad \leftarrow \approx \text{density of } Y_T$$
$$\mathrm{d}X_s^u = \left(\overleftarrow{\sigma} u - \overleftarrow{f}\right)(X_s^u, s)\,\mathrm{d}s + \overleftarrow{\sigma}(s)\,\mathrm{d}B_s, \quad X_0 \sim \mathcal{N}(0, \mathrm{I})$$

PIS:
$$\min_{u\in\mathcal{U}} \mathbb{E}\left[\int_0^T \tfrac{1}{2}\|u(X_s^u, s)\|^2\,\mathrm{d}s + \log\frac{\mathcal{N}(X_T^u; 0, \mathrm{I})}{\rho(X_T^u)}\right] \quad \leftarrow \text{density of } X_T^0$$
$$\mathrm{d}X_s^u = \tfrac{1}{\sqrt{T}} u(X_s^u, s)\,\mathrm{d}s + \tfrac{1}{\sqrt{T}}\,\mathrm{d}B_s, \quad X_0 \sim \delta_{x_0}$$

$\Rightarrow$ DIS can use different SDEs and initial distributions $X \sim \mathcal{N}(0, \nu^2\mathrm{I})$

$\Rightarrow$ Allows to transfer techniques for diffusion models (noise schedules, probability flow ODE, exponential integrators, ...)

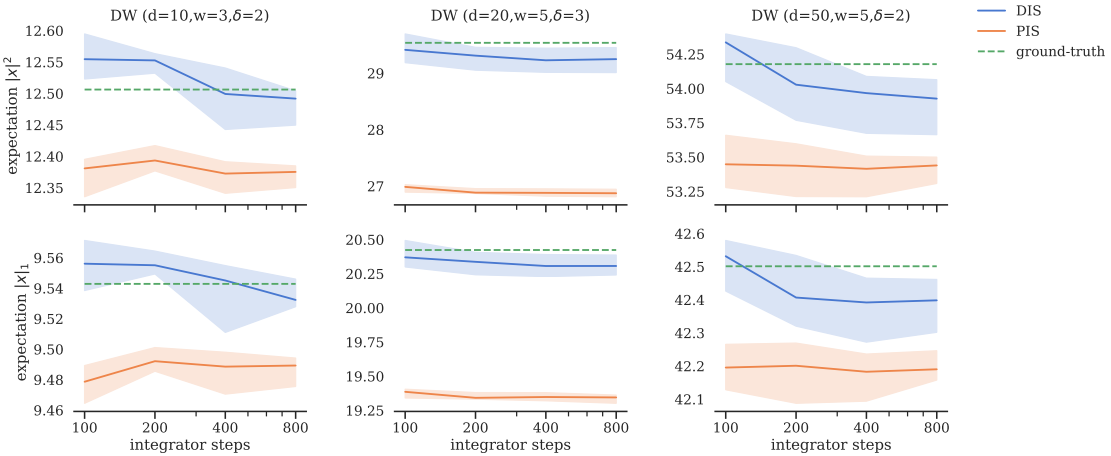
- Concurrent work: **Denoising diffusion sampler** using PIS objective with Ornstein-Uhlenbeck process X and $X_0 \sim \mathcal{N}(0, \nu^2\mathrm{I})$.

Q. Zhang and Y. Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022

F. Vargas, A. Ovsianas, D. Fernandes, M. Girolami, N. D. Lawrence, and N. Nüsken. Bayesian learning via neural Schrödinger–Föllmer flows. *Statistics and Computing*, 33(1):1–22, 2023

DIS vs. PIS

Task 1: Compute expectations $\mathbb{E}[f(X_T^u)] \approx \mathbb{E}[f(Y_0)]$ for $f(x) = \|x\|^2$ and $f(x) = \|x\|_1$ using trained DIS and PIS with varying number of steps in the Euler-Maruyama scheme.

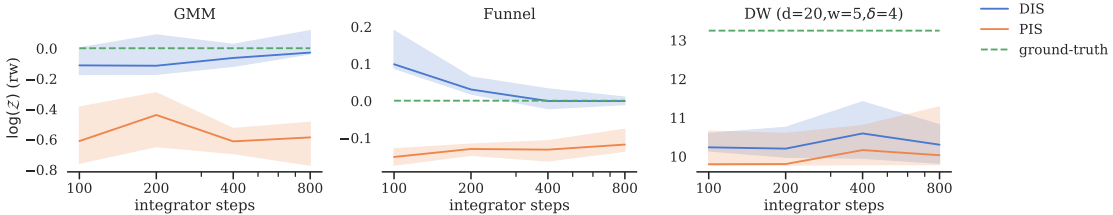


DIS vs. PIS

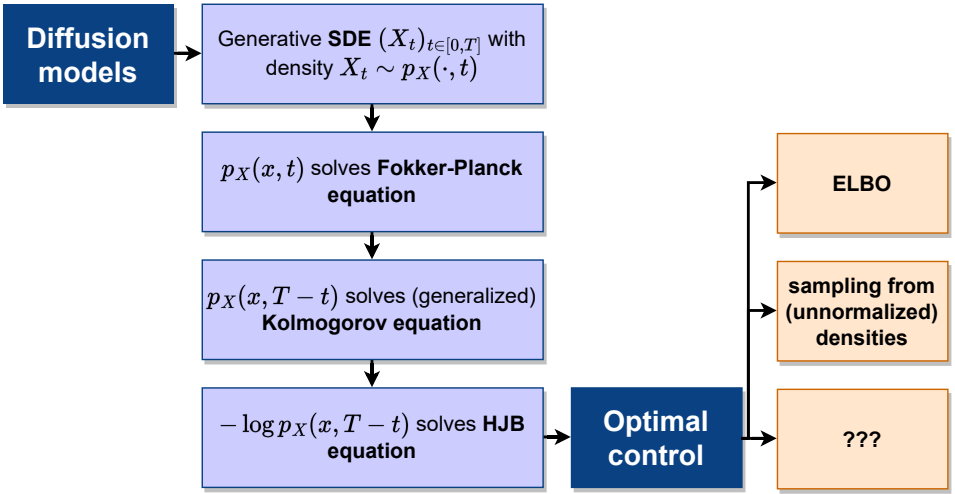
Task 2: Compute the log-normalizing constant $\log \mathcal{Z}$ for a 2-dim. Gaussian mixture model (GMM) with 9 modes on a 3×3 grid, the 10-dimensional Funnel distribution,

$$\rho(x) = \mathcal{N}(x_1; 0, \nu^2) \prod_{i=2}^{10} \mathcal{N}(x_i; 0, e^{x_1}),$$

and the double well (DW) example (using importance sampling in path space).



Summary


Diffusion models
Generative SDE $(X_t)_{t\in[0,T]}$ with density $X_t \sim p_X(\cdot, t)$
$p_X(x,t)$ solves Fokker-Planck equation
$p_X(x, T-t)$ solves (generalized) Kolmogorov equation
$-\log p_X(x, T-t)$ solves HJB equation
Optimal control
ELBO
sampling from (unnormalized) densities
???

Thank you for your attention!

julius.berner@univie.ac.at, lorenz.richter@dida.do, karenu@meta.com