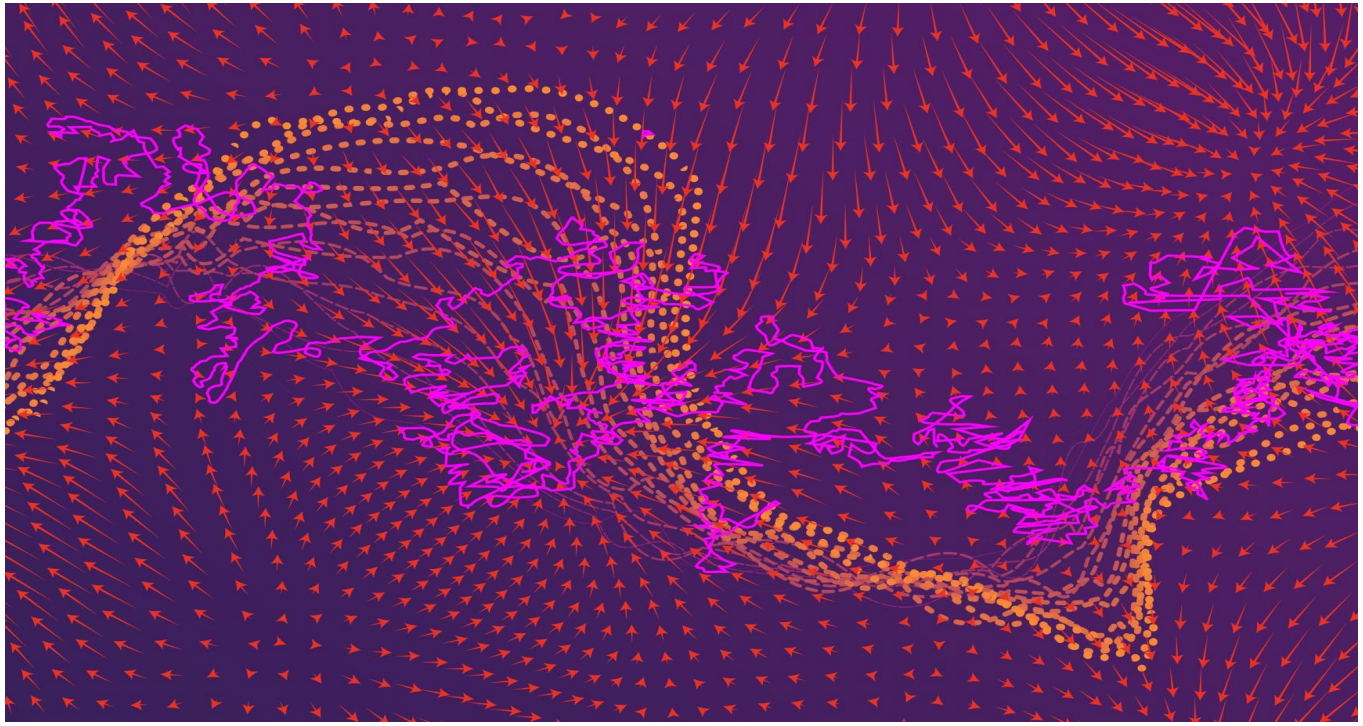
SDE and Wasserstein Gradient Flow

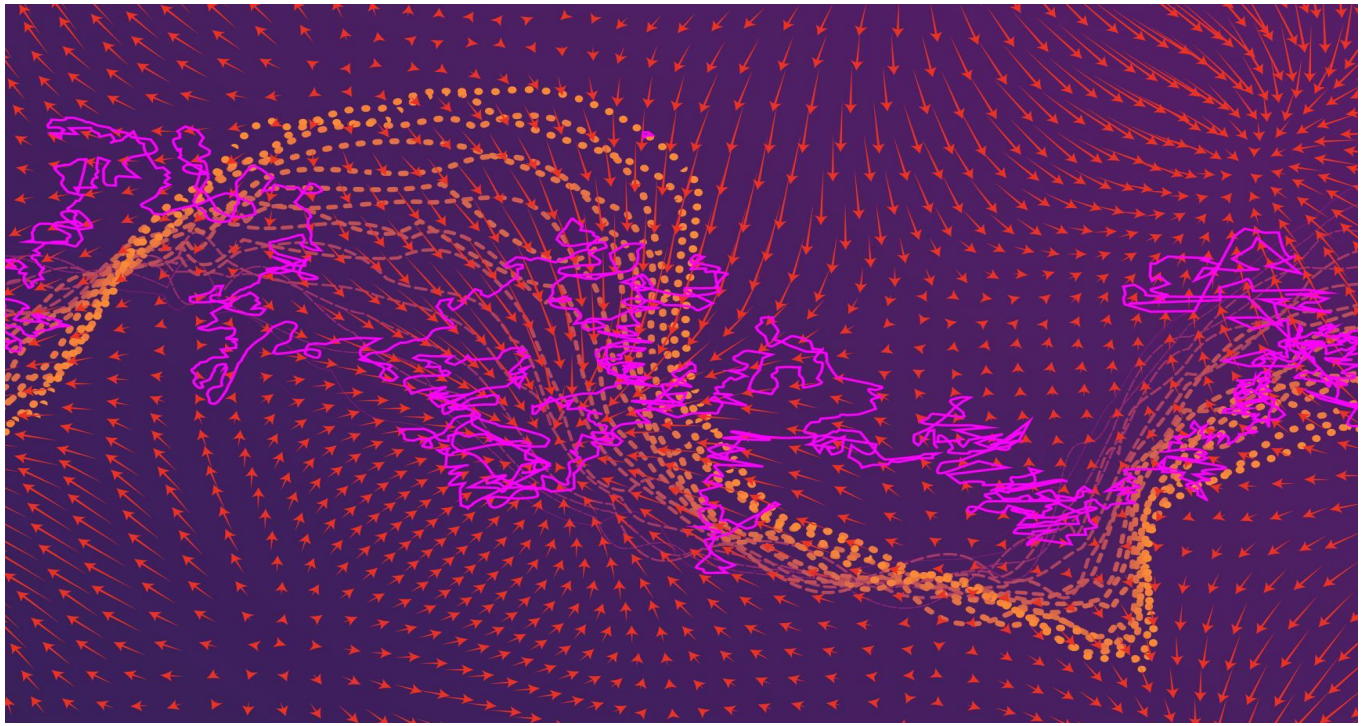
- A unified framework

Bangyan Liao

liaobangyan@westlake.edu.cn



WESTLAKE UNIVERSITY



Introduction

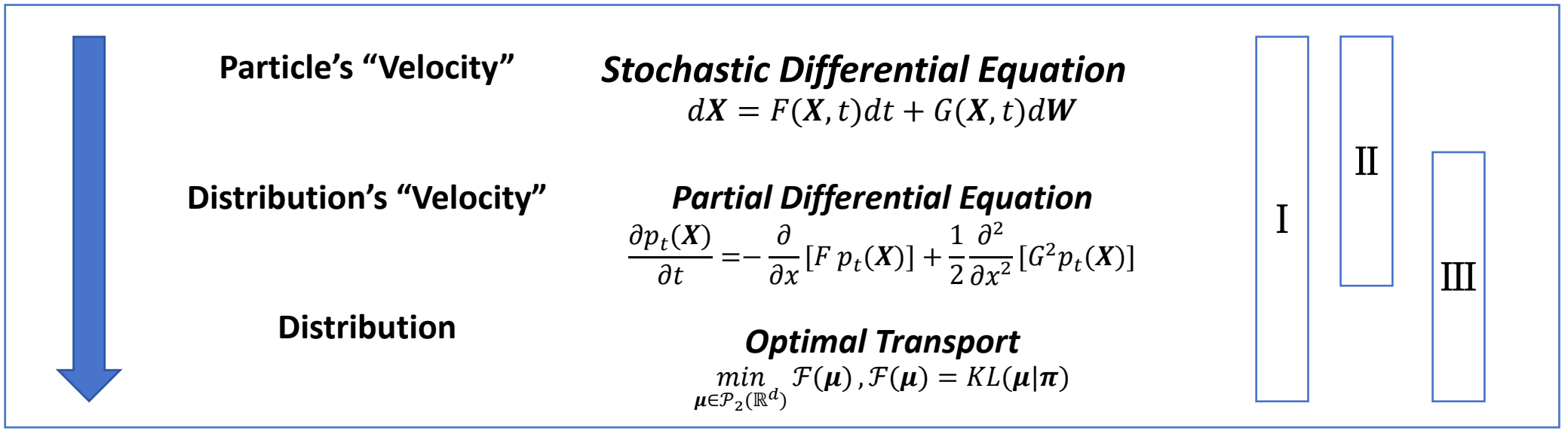
About this tutorial

Introduction

Objective

- Familiarize you with these concepts and build an intuitive understanding
- Provide self-consistent definitions, theories and proofs
- Offer a unified theoretical framework for different topics

We view *Stochastic Differential Equation* (SDE), *Partial Differential Equation* (PDE), and *Optimal Transport* as a unified framework.



Brownian Motion - A first glance

- Brownian Motion and Ito Calculas
- Brownian Motion and Heat Equation
- Heat Equation and Entropy

How to reverse a diffusion process - A second glance

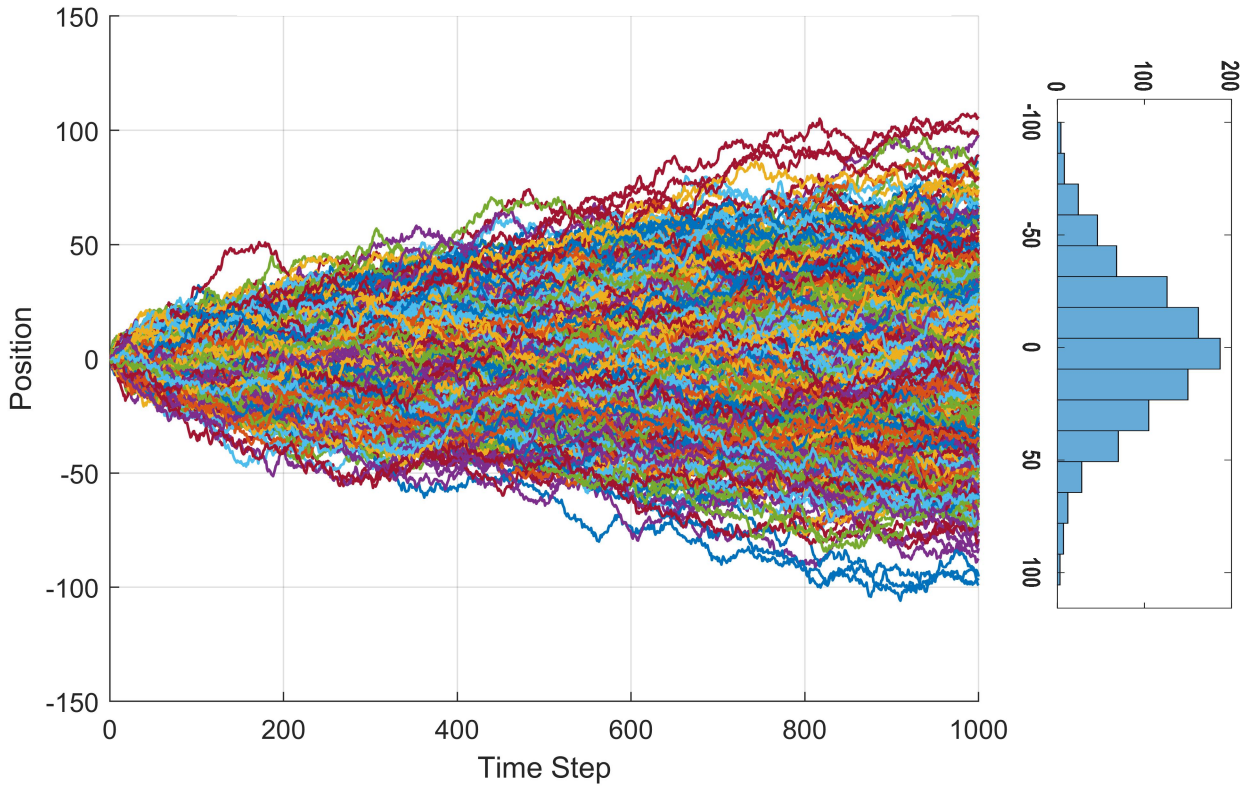
- Stochastic Differential Equation (SDE)
- Fokker-Planck Equation
- Reverse-time SDE and Probabilistic Flow ODE

Optimization over $\mathcal{P}_2(\mathbb{R}^d)$ - A third glance

- Optimization over $\mathbb{R}^d$
- Geometry of $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$
- Wasserstein Gradient Flow

A unified framework - Conclusion

Outline



Brownian Motion - A first glance

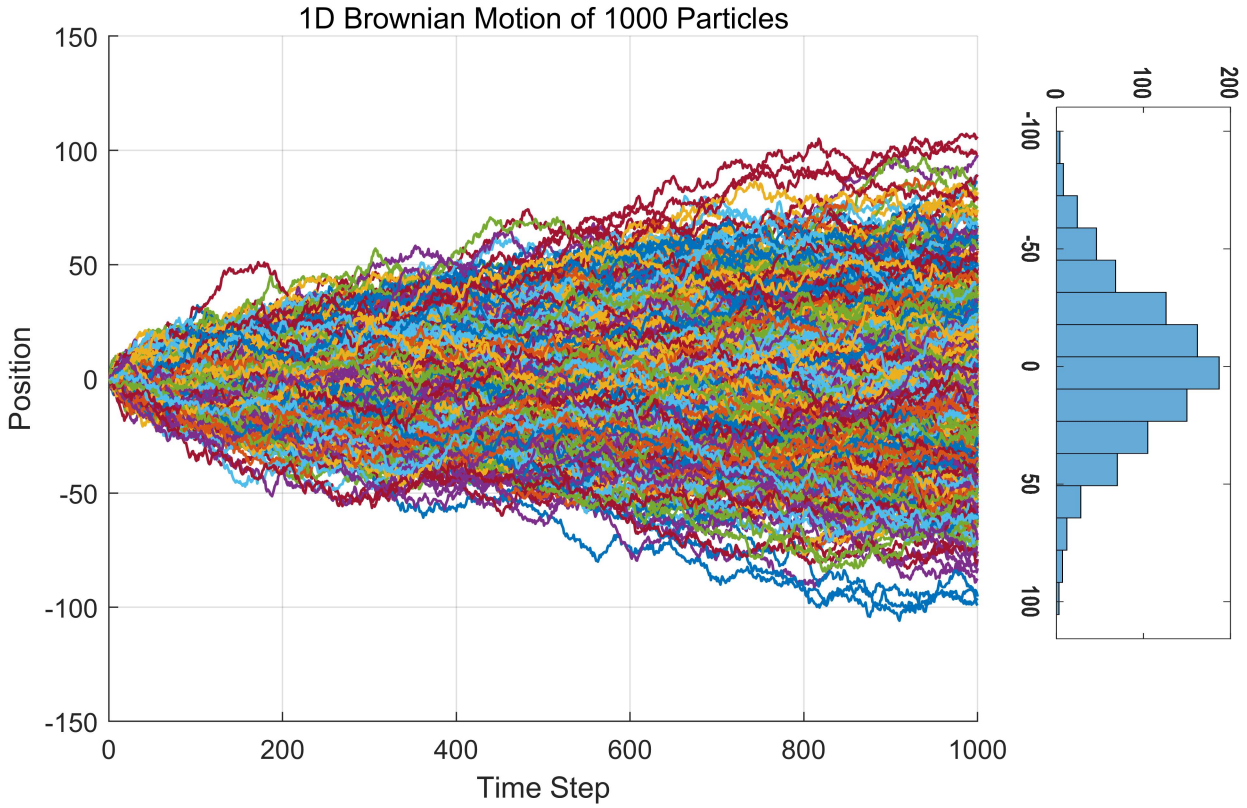
- Brownian Motion and Ito Calculas
- Brownian Motion and Heat Equation
- Heat Equation and Entropy

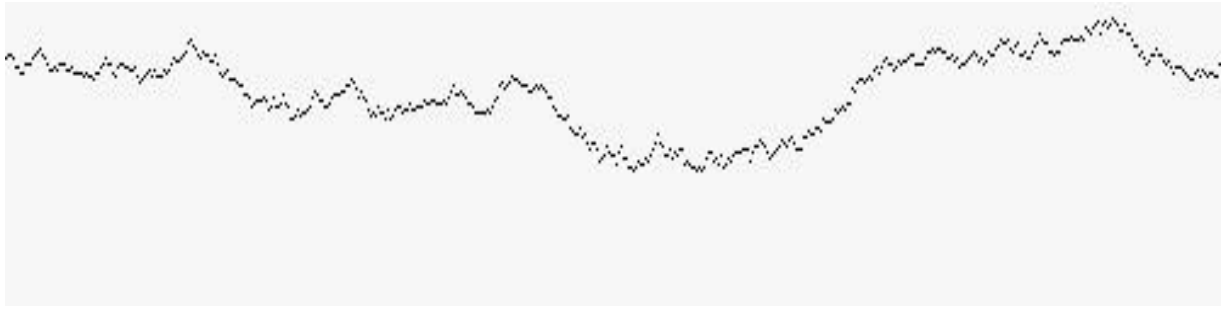
Brownian Motion

Definition. A real-valued stochastic process $W(\cdot)$ is called a Brownian motion (or Wiener process) if
(1) $W(0) = 0$
(2) $W(t) - W(s)$ is $\mathcal{N}(0, \mathrm{t} - \mathrm{s})$ for all $\mathrm{t} \geq \mathrm{s} \geq 0$
(3) for all times $0 \leq t_1 \leq t_2 \leq \cdots \leq t_n$, the random variables $W(t_1), W(t_2) - W(t_1), \cdots, W(t_n) - W(t_{n-1})$ are independent.

Property.

$$\mathbb{E}(W(t)) = 0, \mathbb{E}(W^2(t)) = t$$
$$dWdW \approx dt \text{ (quadratic variation)}$$





Brownian Motion and Ito Calculas

Brownian Motion - A first glance

Ito Integral

Theorem. For all constants $a, b \in \mathbb{R}$ and for all $G, H \in \mathbb{L}^2(0,T)$, we have

$$(1)\ \int_0^T aG + bH\, dW = a\int_0^T G\, dW + b\int_0^T H\, dW$$

$$(2)\ \mathbb{E}\left[\int_0^T G\, dW\right] = 0$$

$$(3)\ \mathbb{E}\left[\left(\int_0^T G\, dW\right)^2\right] = \mathbb{E}\left[\int_0^T G^2\, dt\right]$$

$$(4)\ \mathbb{E}\left[\int_0^T G\, dW \int_0^T H\, dW\right] = \mathbb{E}\left[\int_0^T GH\, dt\right]$$

$\mathbb{L}^2(0,T)$ denotes the space of all real-valued, progressively measurable stochastic processes $G(\cdot)$ such that $\mathbb{E}\left[\int_0^T G^2\, dt\right] < \infty$

$\mathbb{L}^1(0,T)$ denotes the space of all real-valued, progressively measurable stochastic processes $F(\cdot)$ such that $\mathbb{E}\left[\int_0^T |F|\, dt\right] < \infty$

Ito Lemma (Ito's chain rule)[1]

Theorem. Suppose that $X(\cdot)$ has a stochastic differential

$$dX = Fdt + GdW$$

for $F \in \mathbb{L}^1(0,T), G \in \mathbb{L}^2(0,T)$. Assume u: $\mathbb{R} \times [0,T] \to \mathbb{R}$, $u = (x,t)$, is continuous and that its partical derivatives exist and are continuous. Then $Y(t) = u(X(t),t)$ has a stochastic function.

$$du(X,t) = \frac{\partial u}{\partial t}dt + \frac{\partial u}{\partial x}dX + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}G^2 dt$$
$$= \left(\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}F + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}G^2\right)dt + \frac{\partial u}{\partial x}GdW$$

[1] https://en.wikipedia.org/wiki/It%C3%B4%27s_lemma

Heat Equation

Theorem. Assume that the $\mathbb{R}^m$-valued stochastic process $\boldsymbol{X}(\cdot)$ follows the d-dimensional Brownian motion. Then the corresponding $\mathbb{R}^d$ probabilistic distribution $p(\boldsymbol{X})$ evolves following the heat equation:

$$\frac{\partial p}{\partial t} = \frac{1}{2}\Delta p$$

Δu is the laplace operator which is the trace of hessian matrix, e.g.

$$\Delta p(x_1, x_2, x_3) = \frac{\partial^2 p}{\partial x_1^2} + \frac{\partial^2 p}{\partial x_2^2} + \frac{\partial^2 p}{\partial x_3^2}$$

Proof. We denote $p(\boldsymbol{X}, t+dt)$ as the probabilistic distribution $p(\boldsymbol{X}, t)$ evolves from time t to $t+dt$ and $k(\cdot) \sim \mathcal{N}(0, dt)$ as the guassian kernel.

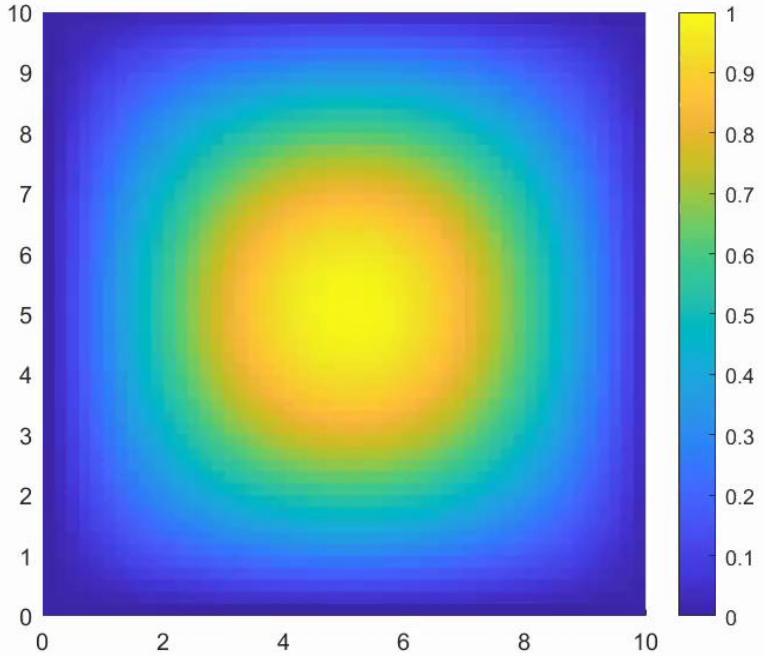
$$p(\boldsymbol{X}, t+dt) = \int_{-\infty}^{+\infty} k(\bar{x}) \int_{-\infty}^{+\infty} k(\bar{y})\ p(x-\bar{x}, y-\bar{y}, t)\, d\bar{x}d\bar{y}$$

$$= \int_{-\infty}^{+\infty} k(\bar{x}) \int_{-\infty}^{+\infty} k(\bar{y}) \left(p - \frac{\partial p}{\partial x}\bar{x} + \frac{1}{2}\frac{\partial^2 p}{\partial x^2}\bar{x}^2 - \frac{\partial p}{\partial y}\bar{y} + \frac{1}{2}\frac{\partial^2 p}{\partial y^2}\bar{y}^2 + \frac{1}{2}\frac{\partial^2 p}{\partial x \partial y}\bar{x}\bar{y} + \cdots \right) d\bar{x}d\bar{y}$$

$$= p \iint k(\bar{x})k(\bar{y}) + \int_{-\infty}^{+\infty} k(\bar{x}) \left(-\frac{\partial p}{\partial x}\bar{x} + \frac{1}{2}\frac{\partial^2 p}{\partial x^2}\bar{x}^2 \right) d\bar{x}$$

$$+ \int_{-\infty}^{+\infty} k(\bar{y}) \left(-\frac{\partial p}{\partial y}\bar{y} + \frac{1}{2}\frac{\partial^2 p}{\partial y^2}\bar{y}^2 \right) d\bar{y} + \int_{-\infty}^{+\infty} k(\bar{x}) \int_{-\infty}^{+\infty} k(\bar{y}) \left(\frac{1}{2}\frac{\partial^2 p}{\partial x \partial y}\bar{x}\bar{y} \right) d\bar{x}d\bar{y}$$

■



Proof. We denote $p(\boldsymbol{X}, t+dt)$ as the probabilistic distribution $p(\boldsymbol{X}, t)$ evolves from time t to $t+dt$ and $k(\cdot) \sim \mathcal{N}(0, dt)$ as the guassian kernel.

$$p(\boldsymbol{X}, t+dt) = \int_{-\infty}^{+\infty} k(\bar{x}) \int_{-\infty}^{+\infty} k(\bar{y})\ p(x-\bar{x}, y-\bar{y}, t)\, d\bar{x}d\bar{y}$$

$$= \int_{-\infty}^{+\infty} k(\bar{x}) \int_{-\infty}^{+\infty} k(\bar{y}) \left(p - \frac{\partial p}{\partial x}\bar{x} + \frac{1}{2}\frac{\partial^2 p}{\partial x^2}\bar{x}^2 - \frac{\partial p}{\partial y}\bar{y} + \frac{1}{2}\frac{\partial^2 p}{\partial y^2}\bar{y}^2 + \frac{1}{2}\frac{\partial^2 p}{\partial x \partial y}\bar{x}\bar{y} + \cdots \right) d\bar{x}d\bar{y}$$

$$= p \iint k(\bar{x})k(\bar{y})d\bar{x}d\bar{y} + \int_{-\infty}^{+\infty} k(\bar{x}) \left(-\frac{\partial p}{\partial x}\bar{x} + \frac{1}{2}\frac{\partial^2 p}{\partial x^2}\bar{x}^2 \right) d\bar{x}$$

$$+ \int_{-\infty}^{+\infty} k(\bar{y}) \left(-\frac{\partial p}{\partial y}\bar{y} + \frac{1}{2}\frac{\partial^2 p}{\partial y^2}\bar{y}^2 \right) d\bar{y} + \frac{1}{2}\frac{\partial^2 p}{\partial x \partial y} \int_{-\infty}^{+\infty} \bar{x}k(\bar{x})d\bar{x} \int_{-\infty}^{+\infty} \bar{y}k(\bar{y})d\bar{y}$$

$$= p + \frac{1}{2}\frac{\partial^2 p}{\partial x^2}dt + \frac{1}{2}\frac{\partial^2 p}{\partial y^2}dt$$

$$\int_{-\infty}^{+\infty} k(\bar{x})d\bar{x} = 1$$

$$\int_{-\infty}^{+\infty} \bar{x}k(\bar{x})d\bar{x} = 0$$

$$\int_{-\infty}^{+\infty} \bar{x}^2 k(\bar{x})d\bar{x} = dt$$

Then

$$\frac{\partial p}{\partial t} = \frac{p(\boldsymbol{X}, t+dt) - p(\boldsymbol{X}, t)}{dt} = \frac{1}{2}\frac{\partial^2 p}{\partial x^2} + \frac{1}{2}\frac{\partial^2 p}{\partial y^2}$$

■

Heat Equation and Entropy

Brownian Motion - A first glance

Entropy

Definition. The entropy of marginal probabilistic distribution $p(x,t)$ at time t is defined as

$$H(t) = -\int p(x,t)\, log p(x,t) dx$$

Property. The probability distribution that conforms to the diffusion described by the heat equation is necessarily a process of entropy increase namely

$$\frac{dH(t)}{dt} \geq 0, \quad if \ \frac{\partial p}{\partial t} = \frac{1}{2}\Delta p$$

Proof (dim = 1). Given that

$$\frac{dH(t)}{dt} = -\int (\, log p(x,t) \, + \, 1) \, \frac{\partial p}{\partial t} dx$$

Substitute in the heat equation

$$\frac{dH(t)}{dt} = -\frac{1}{2}\int (\, log p(x,t) \, + \, 1) \, \frac{\partial^2 p}{\partial x^2} \, dx$$

with integration-by-parts

$$\frac{dH(t)}{dt} = \frac{1}{2}\int \frac{1}{p}\left(\frac{\partial p}{\partial x}\right)^2 \geq 0$$

■

Formally, the heat equation can be derived from the continuity equation and the wasserstion gradient flow of the entropy.

A first glance

Brownian Motion - A first glance

Brownian Motion

$$d\boldsymbol{X} = d\boldsymbol{W}$$

Brownian motion describes the particle velocity (particle velocity field).

Brownian motion realizes the Heat Equation

Heat Equation

$$\frac{\partial p}{\partial t} = \frac{1}{2}\Delta p$$

Heat equation describes the probabilisty density changes in a short time interval on a specific localization. (probabilistic distribution evolution).

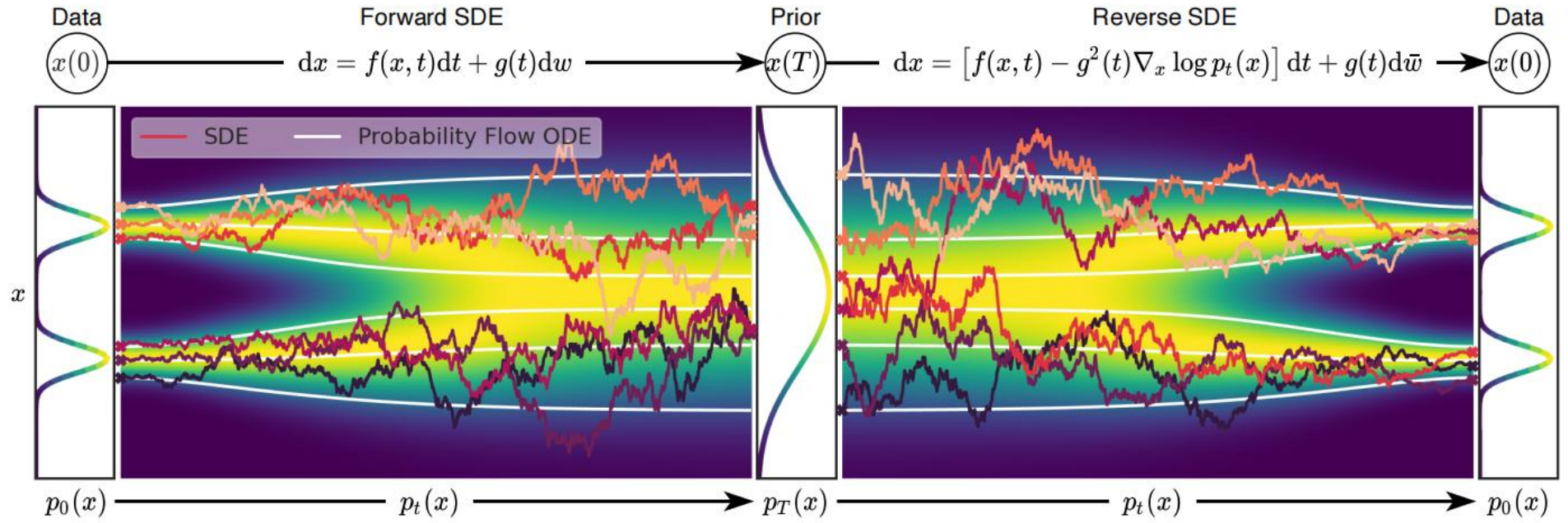
Heat Equation is the continuity equation of Wasserstein Gradient Flow

Entropy

$$H(p) = -\int p(x) log p(x) dx$$

Entropy describes the disorder degree or the information.

Brownian motion realizes the Wasserstein Gradient Flow of the Entropy



How to reverse a diffusion process - A second glance

- Stochastic Diffusion Equation (SDE)
- Fokker-Planck Equation
- Reverse-time SDE and Probabilistic Flow ODE

Stochastic Differential Equation

How to reverse a diffusion process - A second glance

Stochastic Differential Equation

Definition. We say that $\mathbb{R}^m$-valued stochastic process $\boldsymbol{X}(\cdot)$ is a solution of the Ito stochastic differential equation

$$\begin{cases} d\boldsymbol{X} = F(\boldsymbol{X}, t)dt + G(\boldsymbol{X}, t)d\boldsymbol{W} \\ \boldsymbol{X}(0) = \boldsymbol{X_0} \end{cases}$$

where $\boldsymbol{W}(\cdot)$ be an n-dimensional Brownian motion and $\mathrm{F} \in \mathbb{L}^1(0,T)\colon \mathbb{R}^m \to \mathbb{R}^m$, $\mathrm{G} \in \mathbb{L}^2(0,T)\colon \mathbb{R}^m \to \mathbb{R}^{m\times n}$

Ito Lemma (Ito's chain rule) in high dimension

Theorem. For a $\mathbb{R}^m$-valued stochastic process $\boldsymbol{X}(\cdot)$: $d\boldsymbol{X} = F(\boldsymbol{X}, t)dt + G(\boldsymbol{X}, t)d\boldsymbol{W}$, the chain rule of $u(\boldsymbol{X}(t), t)$ can be defined as

$$du(\boldsymbol{X}, t) = \frac{\partial u}{\partial t}dt + \sum_{i=1}^{m}\frac{\partial u}{\partial x_i}d\boldsymbol{X}^i + \frac{1}{2}\sum_{i,j=1}^{m}\frac{\partial^2 u}{\partial x_i \partial x_j}\sum_{l=1}^{n}G^{il}G^{jl}\,dt$$

$$= \left(\frac{\partial u}{\partial t} + \sum_{i=1}^{m}\frac{\partial u}{\partial x_i}F^i + \frac{1}{2}\sum_{i,j=1}^{m}\frac{\partial^2 u}{\partial x_i \partial x_j}\sum_{l=1}^{n}G^{il}G^{jl}\right)dt + \sum_{i=1}^{m}\frac{\partial u}{\partial x_i}\sum_{j=1}^{n}G^{ij}\,d\boldsymbol{W}^j$$

Fokker-Planck Equation

How to reverse a diffusion process - A second glance

Fokker-Planck Equation (Kolmogorov Forward Equation)

Definition. Given the stochastic process $\boldsymbol{X}(\cdot)$: $d\boldsymbol{X} = F(\boldsymbol{X}, t)dt + G(\boldsymbol{X}, t)d\boldsymbol{W}$. The marginal probability density $p_t(\boldsymbol{X}(t))$ at time t evolves according to Fokker-Planck equation

$$\frac{\partial p_t(\boldsymbol{X})}{\partial t} = -\sum_{i=1}^{m} \frac{\partial}{\partial x_i}[F_i(\boldsymbol{X}, t)\, p_t(\boldsymbol{X})] + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \frac{\partial^2}{\partial x_i \partial x_j}\left[\sum_{k=1}^{n} G_{ik}(\boldsymbol{X}, t) G_{jk}(\boldsymbol{X}, t)\, p_t(\boldsymbol{X})\right]$$

when $m = 1$ and $n = 1$

$$\frac{\partial p_t(\boldsymbol{X})}{\partial t} = -\frac{\partial}{\partial x}[F(\boldsymbol{X}, t)\, p_t(\boldsymbol{X})] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[G^2(\boldsymbol{X}, t)\, p_t(\boldsymbol{X})]$$

Remark. 1. The evolution of the marginal probability density is **deterministic**.

2. The heat function $\frac{\partial p}{\partial t} = \frac{1}{2}\Delta p$ is a special case of Fokker-Planck Equation.

Fokker-Planck Equation

How to reverse a diffusion process - A second glance

Fokker-Planck Equation Proof when $m = 1$ and $n = 1$ [1]

Proof. For some stochastic process $\boldsymbol{X}(\cdot)$: $d\boldsymbol{X} = F(\boldsymbol{X}, t)dt + G(\boldsymbol{X}, t)d\boldsymbol{W}$, the $du(\boldsymbol{X})$ can be defined through Ito's lemma for some function u: $\mathbb{R} \times [0, T] \to \mathbb{R}$

$$du(\boldsymbol{X}) = \left(\frac{\partial u}{\partial x}F + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}G^2\right)dt + \frac{\partial u}{\partial x}Gd\boldsymbol{W}$$

for some probabilistic distributions $p(\boldsymbol{X}, t)$

$$\frac{d}{dt}\int u(x)\, p_t(x)\, dx = \frac{d}{dt}\int u(\boldsymbol{X})\, p_0(x)\, dx = \int \frac{\partial u}{\partial x}F\, p_t(x)\, dx + \int \frac{1}{2}\frac{\partial^2 u}{\partial x^2}G^2 p_t(x)\, dx \quad (1)$$

besides,

$$\frac{d}{dt}\int u(x)\, p_t(x)\, dx = \int \frac{\partial p_t(x)}{\partial t}\, u(x)\, dx \quad (2)$$

setting (2) equals to (1) yields

$$\int \frac{\partial p_t(x)}{\partial t}\, u(x)\, dx = \int \frac{\partial u}{\partial x}F\, p_t(x)\, dx + \int \frac{1}{2}\frac{\partial^2 u}{\partial x^2}G^2 p_t(x)\, dx \quad (3)$$

perform integration-by-parts to the right hand side of (3)

$$\int \frac{\partial p_t(x)}{\partial t}\, u(x)\, dx = -\int u(x)\frac{\partial}{\partial x}[F\, p(x,t)]\, dx + \int \frac{1}{2}u(x)\frac{\partial^2}{\partial x^2}[G^2 p(x,t)]\, dx \quad (4)$$

from (4) we can get the final Fokker-Planck Equation

$$\frac{\partial p_t(\boldsymbol{X})}{\partial t} = -\frac{\partial}{\partial x}[F(\boldsymbol{X}, t)\, p_t(\boldsymbol{X})] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[G^2(\boldsymbol{X}, t)\, p_t(\boldsymbol{X})]$$

https://www.thphys.uni-heidelberg.de/~wolschin/statsem21_6s.pdf

■

Reverse-time SDE and Probabilistic Flow ODE

How to reverse a diffusion process - A second glance

Reverse-time SDE

Definition. Given the stochastic process $\boldsymbol{X}(\cdot)$: $d\boldsymbol{X} = F(\boldsymbol{X}, t)dt + G(\boldsymbol{X}, t)d\boldsymbol{W}$ and the marginal probability density $p_t(\boldsymbol{X}(t))$ at time t, the reverse-time stochastic process is defined as

$$d\boldsymbol{X} = \{F(\boldsymbol{X}, \tilde{t}) - \nabla \cdot [G(\boldsymbol{X}, \tilde{t})G(\boldsymbol{X}, \tilde{t})^T] - G(\boldsymbol{X}, \tilde{t})G(\boldsymbol{X}, \tilde{t})^T \nabla_x \, log \, p_{\tilde{t}}(x)\}d\tilde{t} + G(\boldsymbol{X}, \tilde{t})d\widetilde{\boldsymbol{W}}$$

when $n = 1$ and $G(\boldsymbol{X}, t) = G(t)$

$$d\boldsymbol{X} = [F(\boldsymbol{X}, \tilde{t}) - G^2(\tilde{t}) \, \nabla_x \, log \, p_{\tilde{t}}(x)]d\tilde{t} + G(\tilde{t})d\widetilde{\boldsymbol{W}}$$

where $\widetilde{\boldsymbol{W}}(\cdot)$ represents the standard Wiener process when time flows backwards, and $d\tilde{t}$ is an infinitesimal negative timestep from T to 0.

Probabilistic Flow ODE

Definition. For each reverse-time stochastic process, the probabilistic flow ODE can be defined as followed whose trajectories share the marginal probability densities $p_t(\boldsymbol{X}(t))$.

$$d\boldsymbol{X} = \left\{F(\boldsymbol{X}, \tilde{t}) - \frac{1}{2}\nabla \cdot [G(\boldsymbol{X}, \tilde{t})G(\boldsymbol{X}, \tilde{t})^T] - \frac{1}{2}G(\boldsymbol{X}, \tilde{t})G(\boldsymbol{X}, \tilde{t})^T \nabla_x \, log \, p_{\tilde{t}}(x)\right\} d\tilde{t}$$

when $n = 1$ and $G(\boldsymbol{X}, t) = G(t)$

$$d\boldsymbol{X} = \left[F(\boldsymbol{X}, \tilde{t}) - \frac{1}{2}G^2(\tilde{t}) \, \nabla_x \, log \, p_{\tilde{t}}(x)\right] d\tilde{t}$$

where $d\tilde{t}$ is an infinitesimal negative timestep from T to 0.

Reverse-time SDE and Probabilistic Flow ODE

How to reverse a diffusion process - A second glance

Reverse-time SDE when $n = 1$ and $G(\boldsymbol{X}, t) = G(t)$

Proof. For some stochastic process $\boldsymbol{X}(\cdot)$: $d\boldsymbol{X} = F(\boldsymbol{X}, t)dt + G(t)d\boldsymbol{W}$, the corresponding Fokker-Planck equation is defined as

$$\frac{\partial p_t(\boldsymbol{X})}{\partial t} = -\frac{\partial}{\partial x}[F(\boldsymbol{X}, t)\, p_t(\boldsymbol{X})] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[G^2(t)\, p_t(\boldsymbol{X})]$$

We also define the reverse-time stochastic process $\boldsymbol{Y}(\cdot)$: $d\boldsymbol{Y} = F(\boldsymbol{Y}, \tilde{t})d\tilde{t} + G(\tilde{t})d\widetilde{\boldsymbol{W}}$, and the corresponding $q_t(\boldsymbol{Y})$ is defined as

$$\frac{\partial q_t(\boldsymbol{Y})}{\partial t} = -\frac{\partial p_{T-t}(\boldsymbol{X})}{\partial t} = \frac{\partial}{\partial x}[F(\boldsymbol{X}, T-t)\, p_{T-t}(\boldsymbol{X})] - \frac{1}{2}\frac{\partial^2}{\partial x^2}[G^2(T-t)\, p_{T-t}(\boldsymbol{X})]$$

$$= \frac{\partial}{\partial x}\left[F(\boldsymbol{X}, T-t)\, p_{T-t}(\boldsymbol{X}) - G^2(T-t)\, \frac{\partial}{\partial x} p_{T-t}(\boldsymbol{X})\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[G^2(T-t)\, p_{T-t}(\boldsymbol{X})]$$

$$= \frac{\partial}{\partial x}[\,[F(\boldsymbol{X}, T-t) - G^2(T-t)\, \nabla_x \log p_{T-t}(x)\,]\, p_{T-t}(\boldsymbol{X})] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[G^2(T-t)\, p_{T-t}(\boldsymbol{X})]$$

$$= \frac{\partial}{\partial y}[\,[F(\boldsymbol{X}, \tilde{t}) - G^2(\tilde{t})\, \nabla_x \log p_{\tilde{t}}(x)\,]\, q_t(\boldsymbol{Y})] + \frac{1}{2}\frac{\partial^2}{\partial y^2}[G^2(\tilde{t})\, q_t(\boldsymbol{Y})]$$

$$F(\boldsymbol{Y}, \tilde{t}) = F(\boldsymbol{X}, \tilde{t}) - G^2(\tilde{t})\, \nabla_x \log p_{\tilde{t}}(x)\,, \quad G(\tilde{t}) = G(\tilde{t})$$

■

Reverse-time SDE and Probabilistic Flow ODE

How to reverse a diffusion process - A second glance

Probabilistic Flow ODE when $n = 1$ and $G(\boldsymbol{X}, t) = G(t)$

Proof. For some stochastic process $\boldsymbol{X}(\cdot)$: $d\boldsymbol{X} = F(\boldsymbol{X}, t)dt + G(t)d\boldsymbol{W}$, the corresponding Fokker-Planck equation is defined as

$$\frac{\partial p_t(\boldsymbol{X})}{\partial t} = -\frac{\partial}{\partial x}[F(\boldsymbol{X}, t)\, p_t(\boldsymbol{X})] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[G^2(t)\, p_t(\boldsymbol{X})]$$

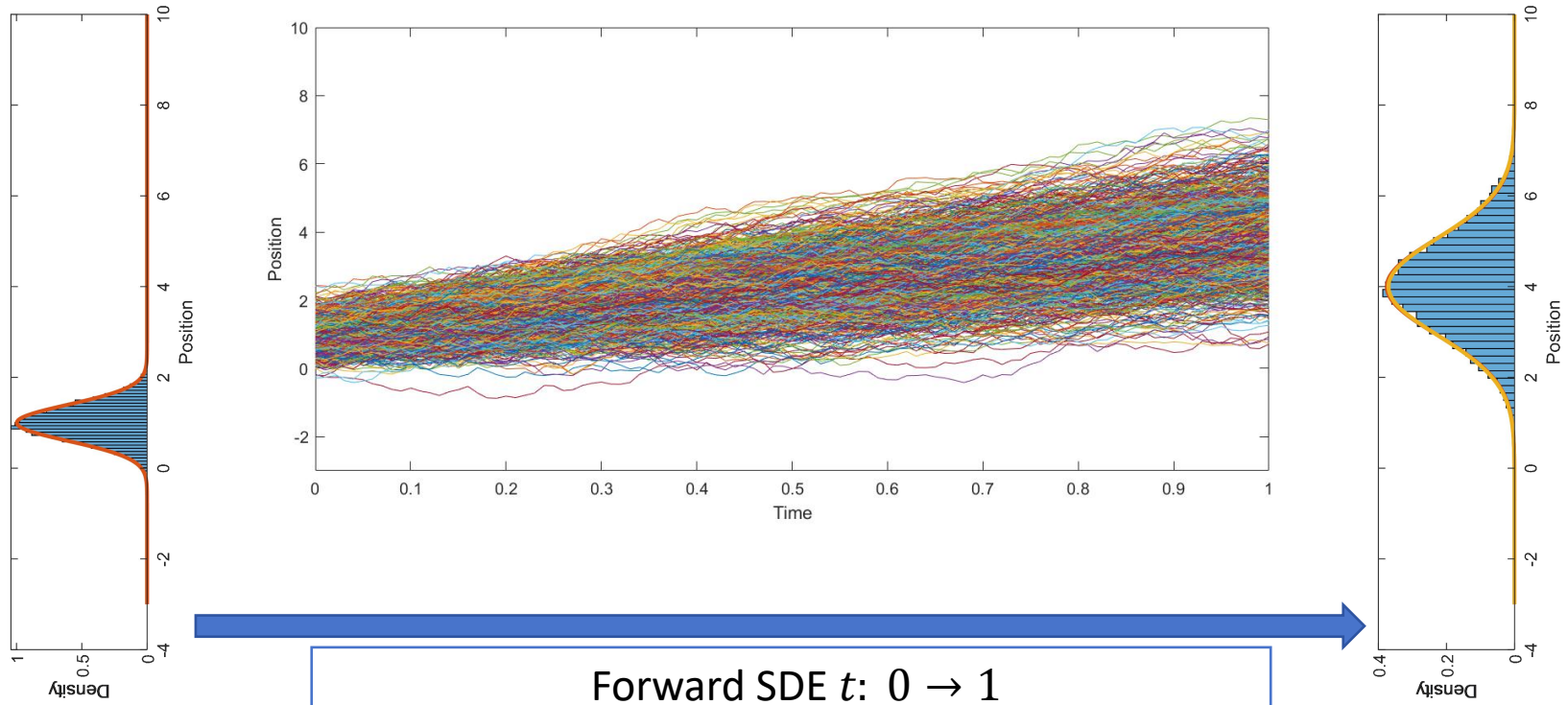
We also define the reverse-time ode process $\boldsymbol{Y}(\cdot)$: $d\boldsymbol{Y} = F(\boldsymbol{Y}, \tilde{t})d\tilde{t}$, and the corresponding $q_{\tilde{t}}(\boldsymbol{Y})$ is defined as

$$\begin{aligned}
\frac{\partial q_t(\boldsymbol{Y})}{\partial t} &= -\frac{\partial p_{T-t}(\boldsymbol{X})}{\partial t} = \frac{\partial}{\partial x}[F(\boldsymbol{X}, T-t)\, p_{T-t}(\boldsymbol{X})] - \frac{1}{2}\frac{\partial^2}{\partial x^2}[G^2(T-t)\, p_{T-t}(\boldsymbol{X})] \\
&= \frac{\partial}{\partial x}\left[F(\boldsymbol{X}, T-t)\, p_{T-t}(\boldsymbol{X}) - \frac{1}{2}G^2(T-t)\, \frac{\partial}{\partial x} p_{T-t}(\boldsymbol{X})\right] \\
&= \frac{\partial}{\partial x}\left[\left[F(\boldsymbol{X}, T-t) - \frac{1}{2}G^2(T-t)\, \nabla_x \log p_{T-t}(x)\right] p_{T-t}(\boldsymbol{X})\right] \\
&= \frac{\partial}{\partial y}\left[\left[F(\boldsymbol{X}, \tilde{t}) - \frac{1}{2}G^2(\tilde{t})\, \nabla_x \log p_{\tilde{t}}(x)\right] q_t(\boldsymbol{Y})\right] \\
F(\boldsymbol{Y}, \tilde{t}) &= F(\boldsymbol{X}, \tilde{t}) - \frac{1}{2}G^2(\tilde{t})\, \nabla_x \log p_{\tilde{t}}(x)
\end{aligned}$$

■

A second glance

How to reverse a diffusion process - A second glance



Forward SDE t: $0 \rightarrow 1$

$X(\cdot)$: $dX = 3dt + dW$

Initial Distribution

$x_0 \sim \mathcal{N}(1, 0.4)$

Target Distribution

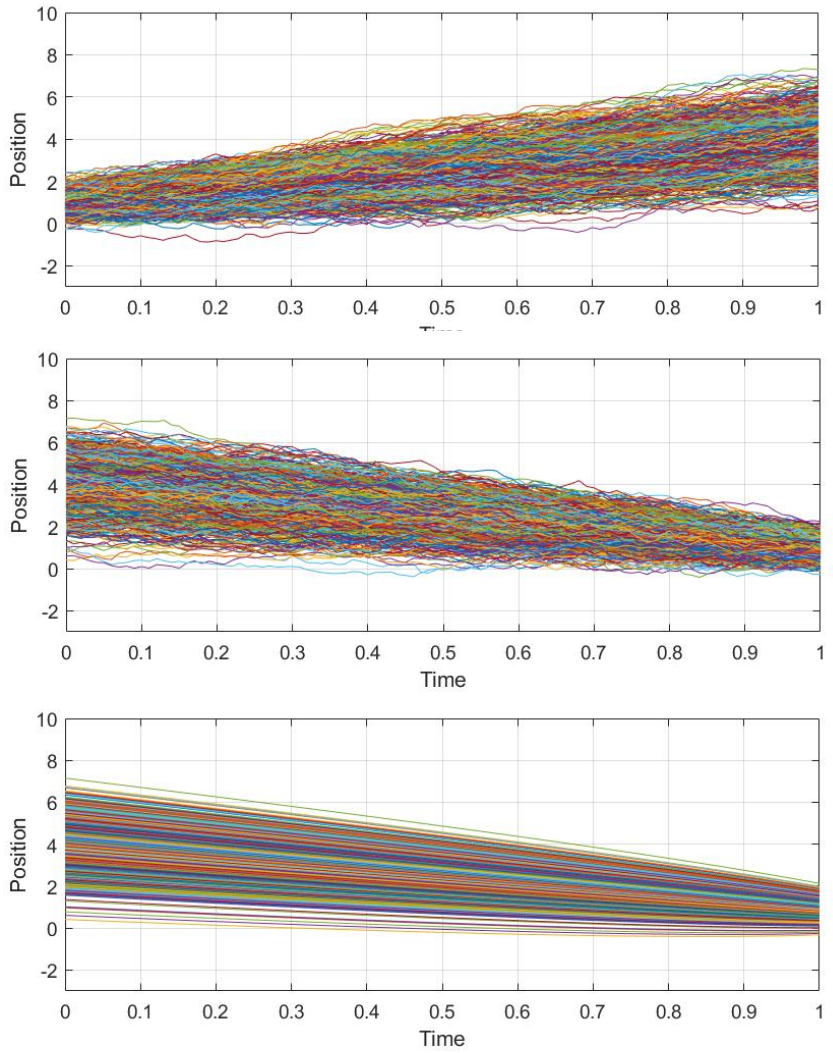
$x_T \sim \mathcal{N}(4, 1.077)$

Fokker-Planck Equation

$$\frac{\partial p_t(\boldsymbol{X})}{\partial t} = -3\frac{\partial}{\partial x}[p_t(\boldsymbol{X})] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[p_t(\boldsymbol{X})]$$

A second glance

How to reverse a diffusion process - A second glance



Forward SDE t: $0 \rightarrow 1$

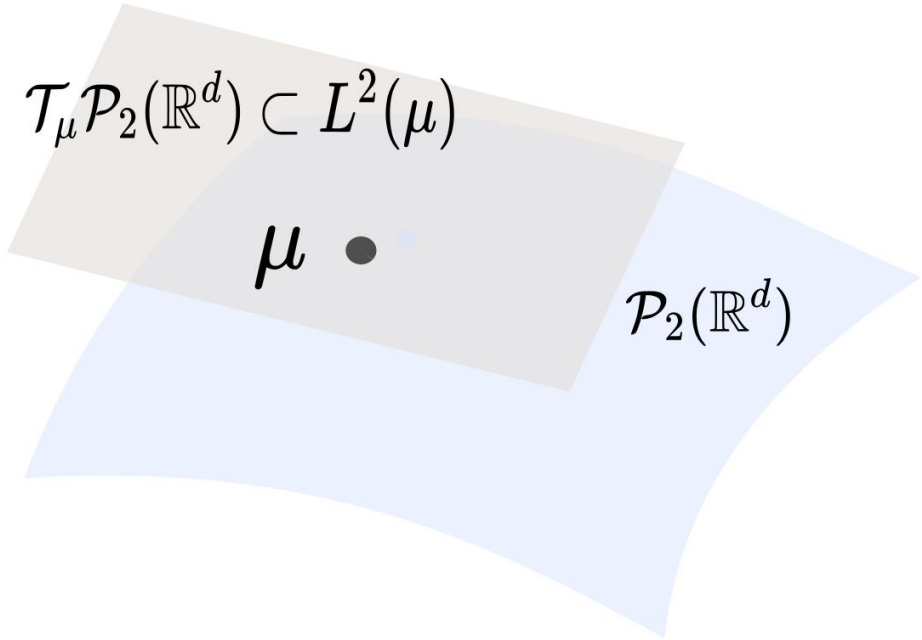
$$X(\cdot):\ dX = 3dt + dW$$

Reverse SDE t: $1 \rightarrow 0$

$$Y(\cdot): dY = [3 - \nabla_x \log p_{\tilde{t}}(x)]d\tilde{t} + d\widetilde{W}$$

Probabilistic Flow ODE t: $1 \rightarrow 0$

$$Y(\cdot): dY = \left[3 - \frac{1}{2}\nabla_x \log p_{\tilde{t}}(x)\right] d\tilde{t}$$



Optimization over $\mathcal{P}_2(\mathbb{R}^d)$ - A third glance

- Optimization over $\mathbb{R}^d$
- Geometry of $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$
- Wasserstein Gradient Flow

Optimization over $\mathbb{R}^d$

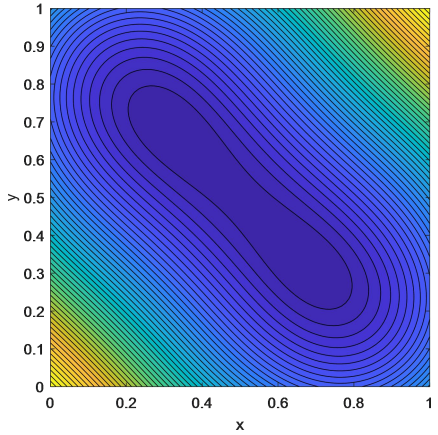
Optimization over $\mathcal{P}_2(\mathbb{R}^d)$ - A third glance

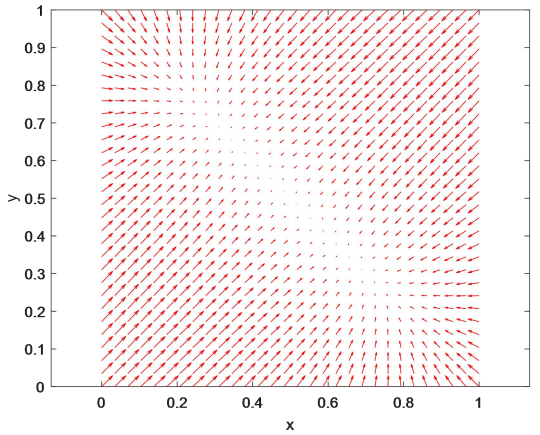
Optimization Problem

$$\min_{x \in \mathbb{R}^d} V(x)$$

where V: $\mathbb{R}^d \to \mathbb{R}$ describes the loss function

Example.





$$V(x,y) = cos(\pi x) * cos(\pi y) + 4(x-0.5)^2 + 4(y-0.5)^2$$

$$\nabla V(x,y) = \begin{bmatrix} -\pi * sin(\pi x) * cos(\pi y) + 8(x-0.5) \\ -\pi * cos(\pi x) * sin(\pi y) + 8(y-0.5) \end{bmatrix}$$

Euclidean Gradient ∇V

Definition.

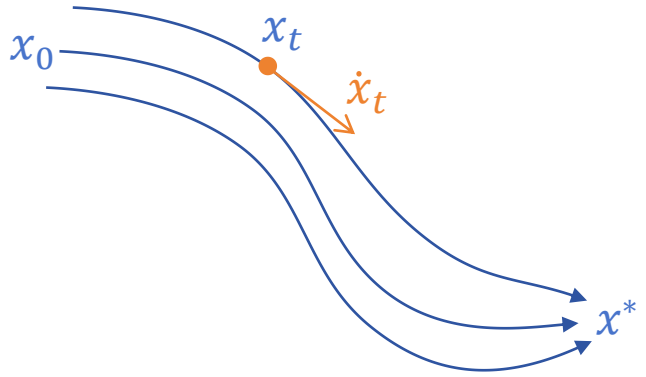
$$d\,V(x)[v] = \lim_{\varepsilon \to 0} \frac{V(x+\varepsilon v) - V(x)}{\varepsilon} = \langle \nabla V, v \rangle$$

Euclidean Gradient Flow

The gradient flow describes the evolution of a point in the gradient vector field. We can represent it by a curve $x_t = x(t)$ that changes over time, and this curve satisfies the following ordinary differential equation (ODE)

Definition.

$$\dot{x}_t = \frac{\partial x_t}{\partial t} = -\nabla V(x_t)$$



Geometry of $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$

Optimization over $\mathcal{P}_2(\mathbb{R}^d)$ - A third glance

Pushforward Measure

Definition. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $T\colon\ \mathbb{R}^d \to \mathbb{R}^d$ a measurable map. The pushforward measure $T_{\#}\mu$ is characterized by:

$$X \sim \mu \implies T(X) \sim T_{\#}\mu$$

and if $T_{\#}\mu = \nu$, and $T: X \to Y$

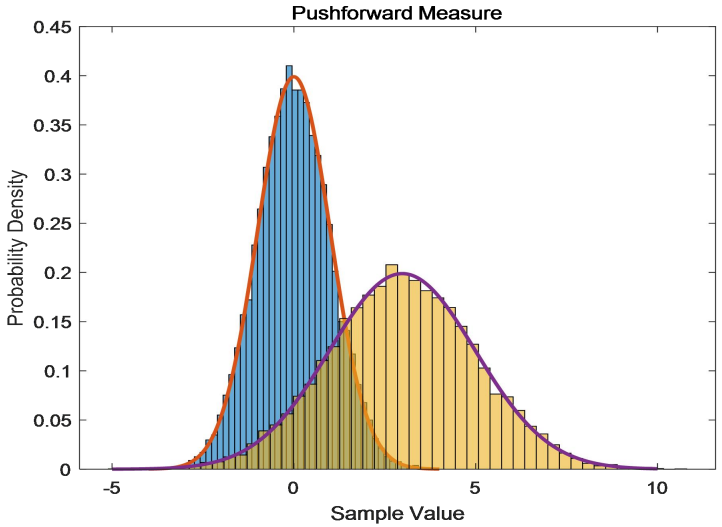
$$\nu(A) = \mu(T^{-1}(A)), \forall A \subset Y \; Borel$$

Remarks. 1) $Id_{\#}\mu = \mu$ where Id denotes the identity map

2) $T \in L^2(\mu)$

3) $T_{\#}\mu = \nu \iff \int_Y \psi(y)\, d\nu(y) = \int_X \psi(T(x))\, d\mu(x) \quad \forall \psi\colon Y \to \mathbb{R}\; Borel \; and \; bounded$

Example.



$\mu = \mathcal{N}(0,1)$

$T(X) = 2X + 3$

$T_{\#}\mu = \mathcal{N}(3,2)$

Geometry of $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$

Optimization over $\mathcal{P}_2(\mathbb{R}^d)$ - A third glance

Definition of the Wasserstein Space

Let $\mathcal{P}_2(\mathbb{R}^d)$ the space of probability measures on $\mathbb{R}^d$ with finite second moments

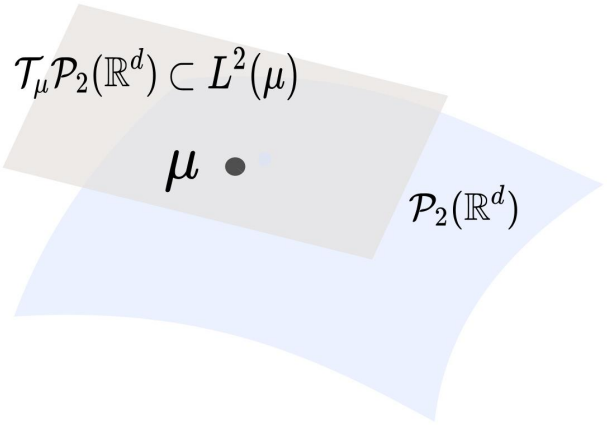
$$\mathcal{P}_2(\mathbb{R}^d) = \left\{\mu \in \mathcal{P}(\mathbb{R}^d), \int \|x\|^2 d\mu(x) < \infty\right\}$$

$\mathcal{P}_2(\mathbb{R}^d)$ is endowed with the Wasserstein-2 distance from optimal transport: $\forall \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathcal{W}_2^2(\mu, \nu) = \inf_{s \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, ds(x, y)$$

where $\Gamma(\mu, \nu)$ is the set of possible couplings μ and ν. The metric space $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ is called the Wasserstein space.

Riemannian structure of $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ and L^2 spaces



Denote by $L^2(\mu) = \left\{f\colon \mathbb{R}^d \to \mathbb{R}^d, \int_{\mathbb{R}^d} \|f(x)\|^2 d\mu(x) < \infty\right\}$ the space of vector-valued, square-integrable functions w.r.t μ.

Geometry of $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$

Optimization over $\mathcal{P}_2(\mathbb{R}^d)$ - A third glance

Brener's Theorem [Brenier, 1991]

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and that $\mu \ll dx$ (i.e., μ is absolutely continuous with respect to the Lebesgue measure). Then there exists a **unique** and **optimal** transport map T_μ^ν such that

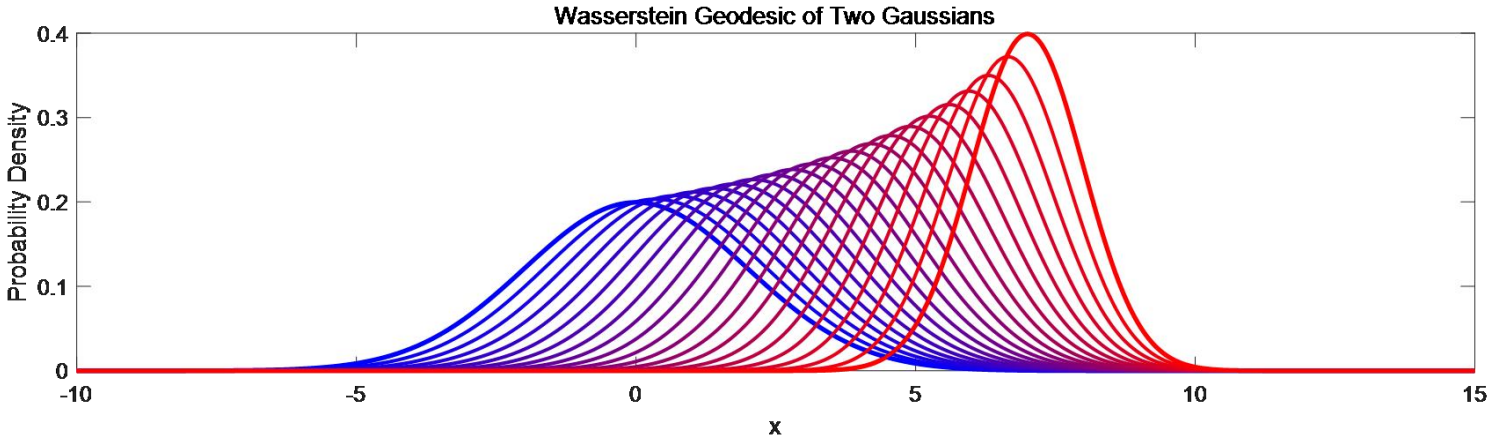
$$(Id \times T_\mu^\nu)_\# = \inf_{s \in \Gamma(\mu,\nu)} \mathcal{W}_2^2(\mu,\nu) = \inf_{s \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^2 \, ds(x,y)$$

Wasserstein Geodesic

Definition. Given the optimal transport map T_μ^ν, the wasserstein geodesic between μ and ν is defined as

$$\rho_t = ((1-t)Id + t\, T_\mu^\nu)_\# \mu, \quad t \in [0,1]$$

Example.



Geometry of $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$

Optimization over $\mathcal{P}_2(\mathbb{R}^d)$ - A third glance

Velocity Field of μ_t

Definition. Let ψ be a smooth test function, the velocity field $v_t(x)$ is defined as

$$\frac{d}{dt}\int \psi(x)\,\mu_t(x)dx = \int \nabla\psi(x)\,\cdot\, v_t(x)\,\mu_t(x)dx$$

Continuity Equation

Definition. Let $\mu_t = (\boldsymbol{X}(t))_{\#}\mu_0$ denotes the evolving probabilistic measure under the stochastic process $\boldsymbol{X}(t)$ and v_t represents the velocity field of μ_t. The continuity equation describes the relationship between flow field and probabilistic measure:

$$\frac{\partial \mu_t}{\partial t} + \nabla\cdot(v_t\mu_t) = 0$$

$\nabla\cdot()$ is the divergence operator e.g.

$$\nabla\cdot(p(x_1,x_2,x_3)) = \frac{\partial p}{\partial x_1} + \frac{\partial p}{\partial x_2} + \frac{\partial p}{\partial x_3}$$

Proof. Let ψ be a smooth test function, and consider the function $t \to \int \psi(x)\,\mu_t(x)dx$, we have

$$\int \psi(x)\frac{\partial \mu_t(x)}{\partial t}dx = \frac{d}{dt}\int \psi(x)\,\mu_t(x)dx = \frac{d}{dt}\int \psi(\boldsymbol{X}(x,t))\,\mu_0(x)dx$$

$$= \int \nabla\psi(\boldsymbol{X}(x,t))\,\cdot\, v_t(\boldsymbol{X}(x,t))\,\mu_0(x)dx$$

$$= \int \nabla\psi(x)\,\cdot\, v_t(x)\,\mu_t(x)dx = -\int \psi(x)\,\nabla\cdot(v_t(x)\,\mu_t(x))\,dx$$

■

Wasserstein Gradient Flow

Optimization over $\mathcal{P}_2(\mathbb{R}^d)$ - A third glance

Variational Optimization

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu), \;\; \mathcal{F}(\mu) = KL(\mu|\pi)$$

Wasserstein Gradient $\nabla_{\mathcal{W}_2}\mathcal{F}(\mu)$

Definition.

$$d\,\mathcal{F}(\mu)[v] = \lim_{t\to 0} \frac{\mathcal{F}((Id + tv)_{\#}\mu) - \mathcal{F}(\mu)}{t} = \langle \nabla_{\mathcal{W}_2}\mathcal{F}(\mu), v\rangle_{\mu}, \forall v \in L^2(\mu)$$

Wasserstein Gradient Flow $\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)$

Definition.

$$\dot{x}_t = v_t = -\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)$$

equivalently:

$$\frac{\partial \mu_t}{\partial t} = \nabla \cdot \left(\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)\,\mu_t\right)$$

specifically

$$\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t) = \nabla\left(\frac{\delta \mathcal{F}(\mu_t)}{\delta \mu_t}\right)$$

Wasserstein Gradient Flow

Optimization over $\mathcal{P}_2(\mathbb{R}^d)$ - A third glance

Wasserstein Gradient Flow $\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t) = \nabla\left(\frac{\delta\mathcal{F}(\mu_t)}{\delta\mu_t}\right)$

Proof.

$$d\,\mathcal{F}(\mu)[v] = \lim_{t\to 0}\frac{\mathcal{F}((Id+tv)_{\#}\mu) - \mathcal{F}(\mu)}{t}$$

we expand $\mu_t = (Id+tv)_{\#}\mu$

$$\mu_t = \mu + t - \nabla\cdot(\mu v)$$

then we expand $\mathcal{F}((Id+tv)_{\#}\mu)$

$$\mathcal{F}((Id+tv)_{\#}\mu) = \mathcal{F}(\mu + t - \nabla\cdot(\mu v)) = \mathcal{F}(\mu) - t\int\frac{\delta\mathcal{F}(\mu_t)}{\delta\mu_t}(\nabla\cdot(\mu v))dx$$

then by integration-by-parts

$$\mathcal{F}((Id+tv)_{\#}\mu) = \mathcal{F}(\mu) + t\int\nabla\frac{\delta\mathcal{F}(\mu_t)}{\delta\mu_t}\cdot v dx$$

finally

$$\langle\nabla_{\mathcal{W}_2}\mathcal{F}(\mu), v\rangle_{\mu} = d\,\mathcal{F}(\mu)[v] = \int\nabla\frac{\delta\mathcal{F}(\mu_t)}{\delta\mu_t}\cdot v\,\mu_t dx$$

$$\nabla_{\mathcal{W}_2}\mathcal{F}(\mu) = \nabla\frac{\delta\mathcal{F}(\mu_t)}{\delta\mu_t}$$

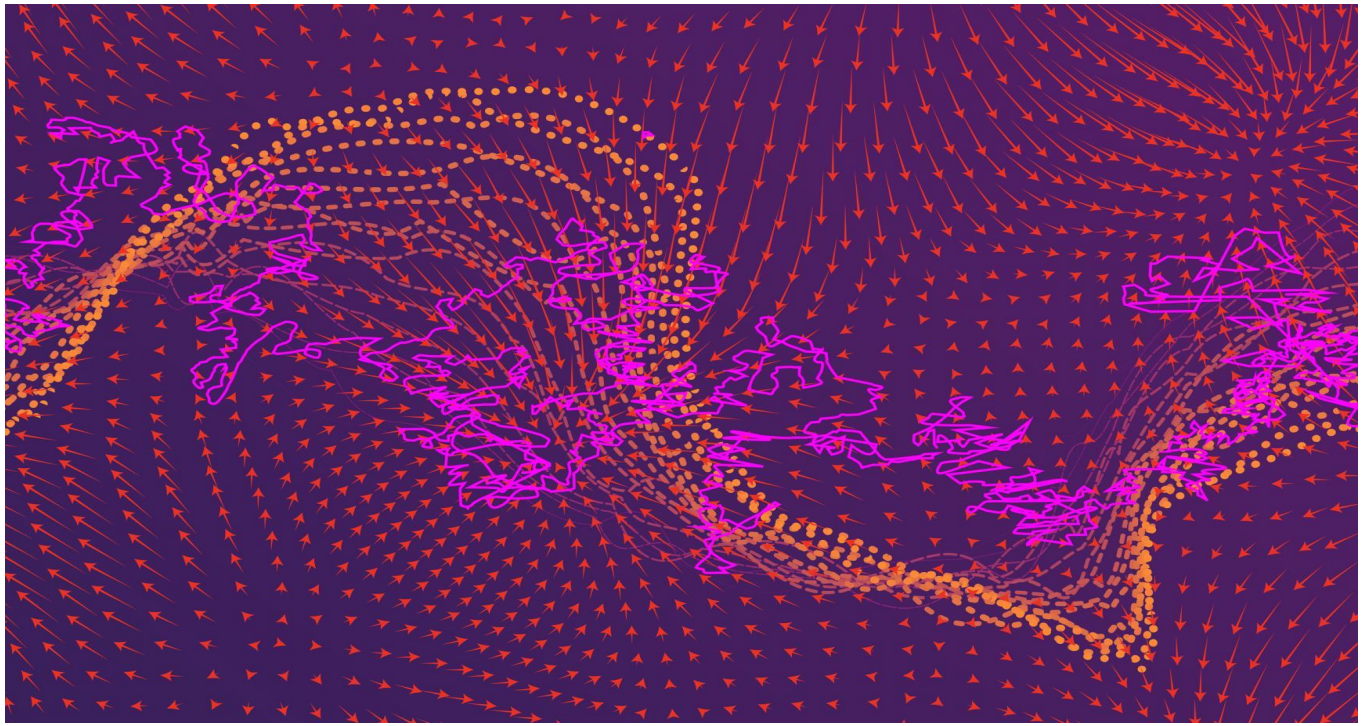
■

Wasserstein Gradient Flow

Optimization over $\mathcal{P}_2(\mathbb{R}^d)$ - A third glance

Wasserstein Gradient Flow Example

Variational Optimization Problem	$\mathcal{F}(\mu_t)$	$\frac{\delta \mathcal{F}(\mu_t)}{\delta \mu_t}$	$\nabla_{W_2} \mathcal{F}(\mu_t) = \nabla \frac{\delta \mathcal{F}(\mu_t)}{\delta \mu_t}$	$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (v_t \mu_t) = 0$
Potential Energy	$\int V(x)\, d\mu_t(x)$	V	∇V	$\frac{\partial \mu_t}{\partial t} = \nabla \cdot (\nabla V\, \mu_t)$
Negative Entropy	$\frac{1}{2} \int log(\mu_t(x))\, d\mu_t(x)$	$\frac{1}{2} log(\mu_t) + \frac{1}{2}$	$\frac{1}{2} \nabla log(\mu_t)$	$\frac{\partial \mu_t}{\partial t} = \frac{1}{2} \Delta \mu_t$ heat equation
KL Divergence	$\int V(x) + log(\mu_t(x)) d\mu_t(x)$	$V + log(\mu_t) + 1$	$\nabla log\left(\frac{\mu_t}{\pi}\right)$	$\frac{\partial \mu_t}{\partial t} = \Delta \mu_t + \nabla \cdot (\nabla V\, \mu_t)$ Fokker-Planck Equation

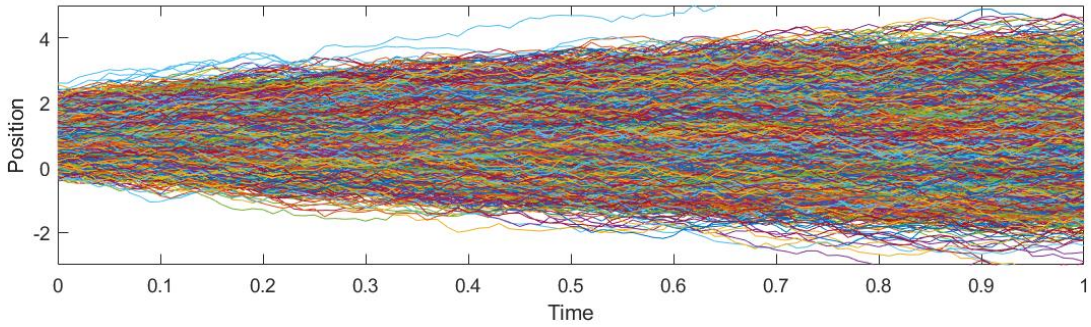


A unified framework - Conclusion

Brownian Motion
A unified framework - Conclusion

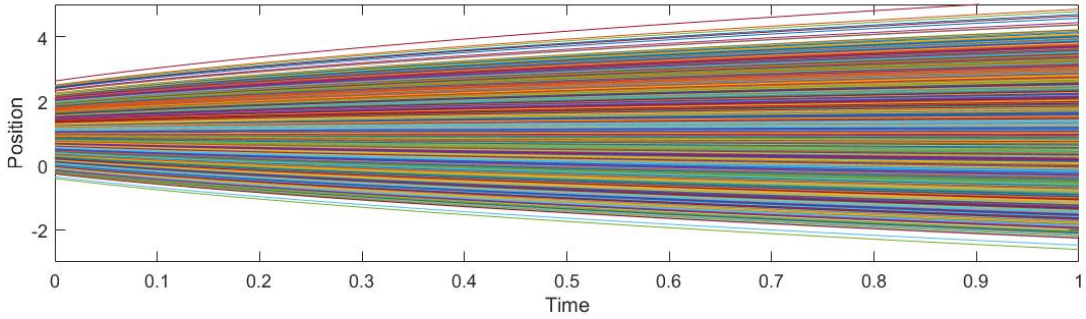
Brownian Motion

$$d\boldsymbol{X} = d\boldsymbol{W}$$



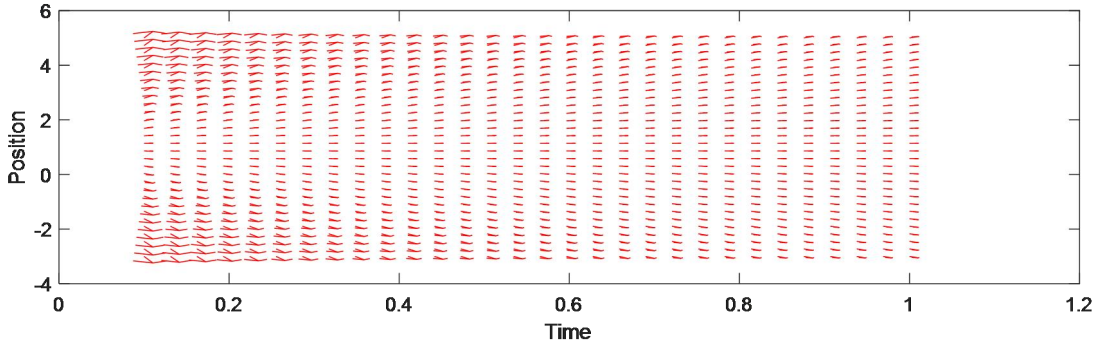
Probabilistic ODE

$$d\boldsymbol{X} = -\frac{1}{2}\nabla log(\mu_t)\, dt$$



Wasserstein Gradient Flow

$$\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t) = -\frac{1}{2}\nabla log(\mu_t)$$

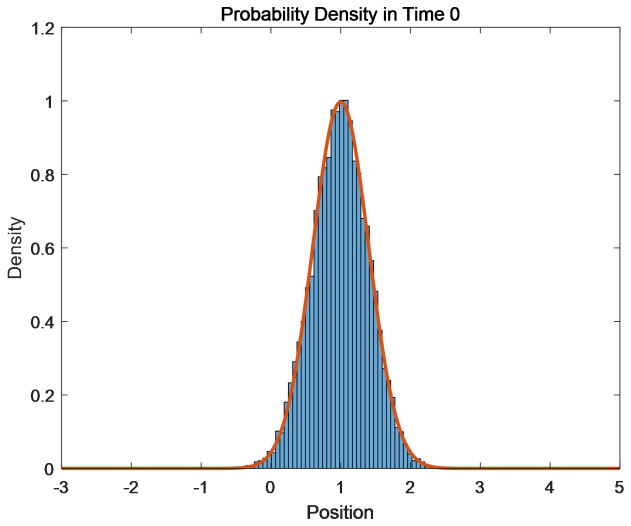


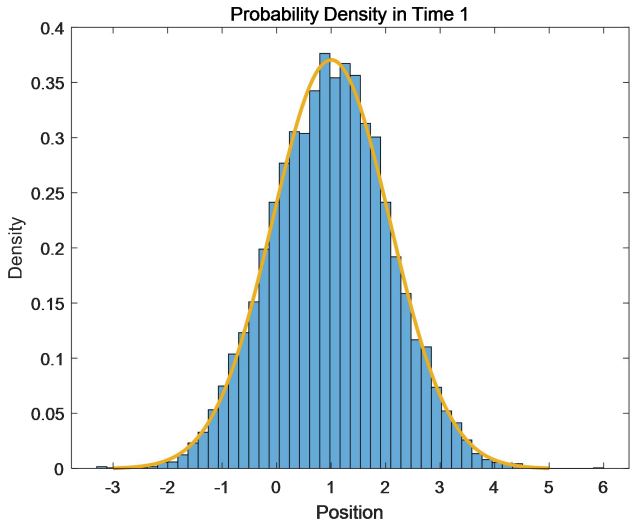
Brownian Motion

A unified framework - Conclusion

Heat Equation

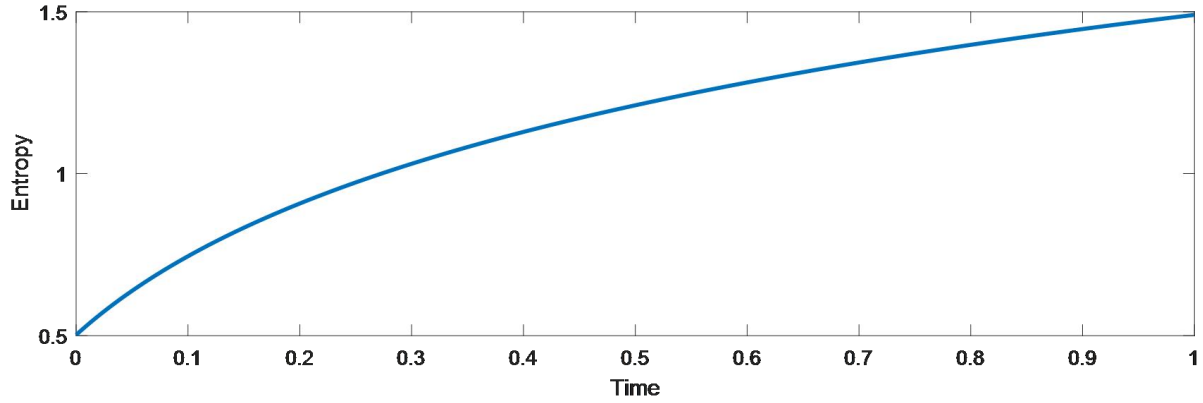
$$\frac{\partial p}{\partial t} = \frac{1}{2}\Delta p$$





Entropy

$$H(\mu_t) = -\frac{1}{2}\int \mu_t(x) log \mu_t(x) dx$$



Further Readings

1. Sampling as First-Order Optimization over a space of probability measures
2. Optimal Transport in Learning, Control, and Dynamical Systems
3. Computational Optimal Transport
4. An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows