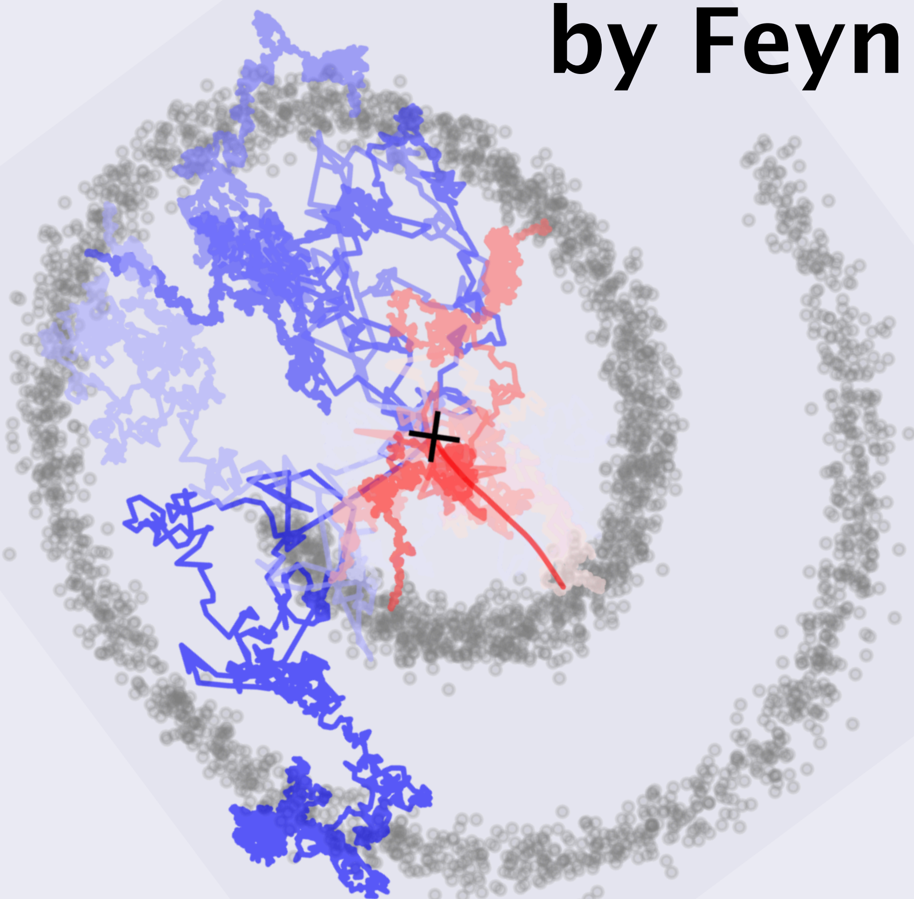
Understanding Diffusion Models by Feynman's Path Integral



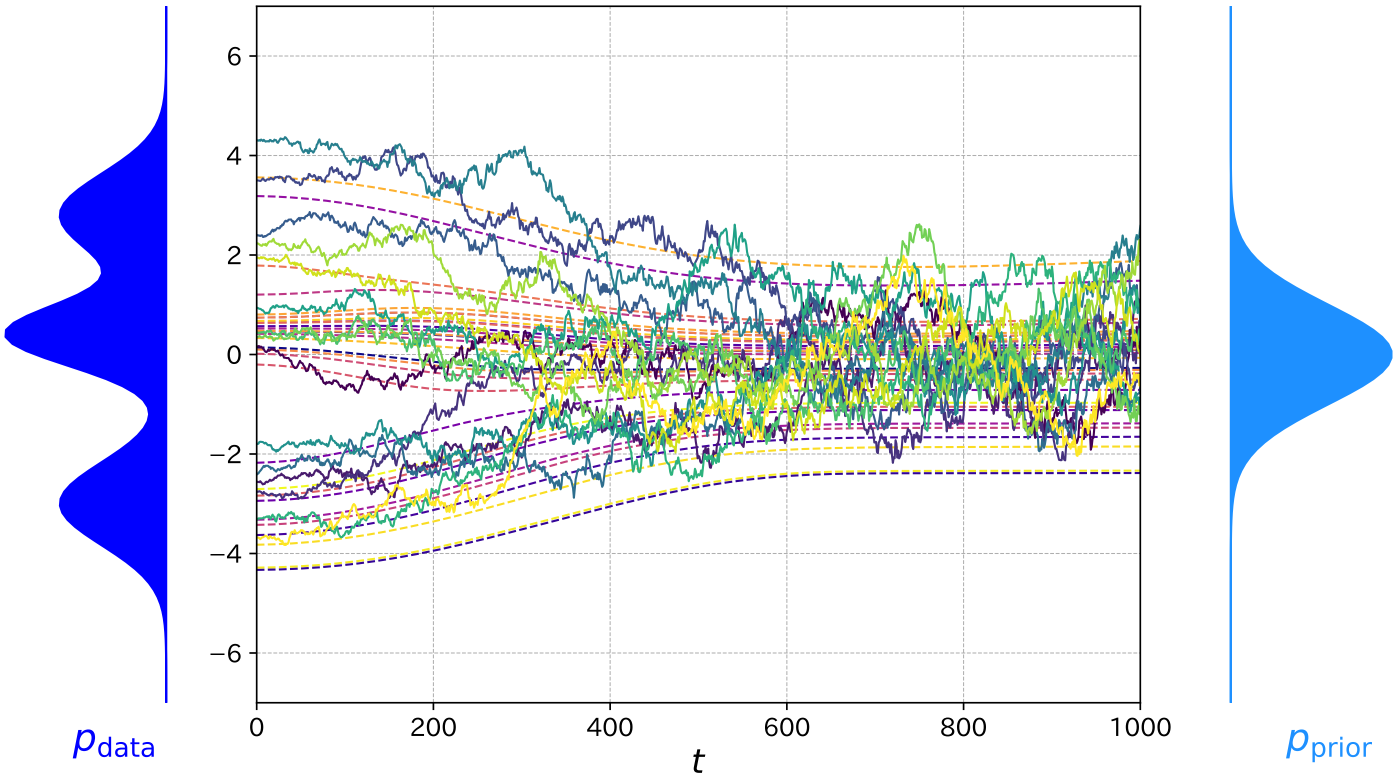
Yuji Hirono (Univ. of Tsukuba)

Collaborators: A. Tanaka (RIKEN), K. Fukushima(U. Tokyo)

Generative diffusion models

- **Diffusion models**: Generative AI for images, movies, and texts
  - Text-to-image generation, inpainting, hyper-resolution, Large Language Model, …
- We reformulate diffusion models via the path-integral method
  - Understanding ML models via physics methods

[Hirono-Tanaka-Fukushima, ICML'24]



Examples of generated images

"A family of lions in a cozy ramen shop"

"A set of sushi that look like dogs"

Examples of generated movies

"There is a family of lions in a cozy ramen shop. They eat noodles with chopsticks"

Examples of generated movies

“A family of **real majestic** lions in a cozy ramen shop. They use chopsticks to eat noodles. ”

Understanding diffusion models via path integral

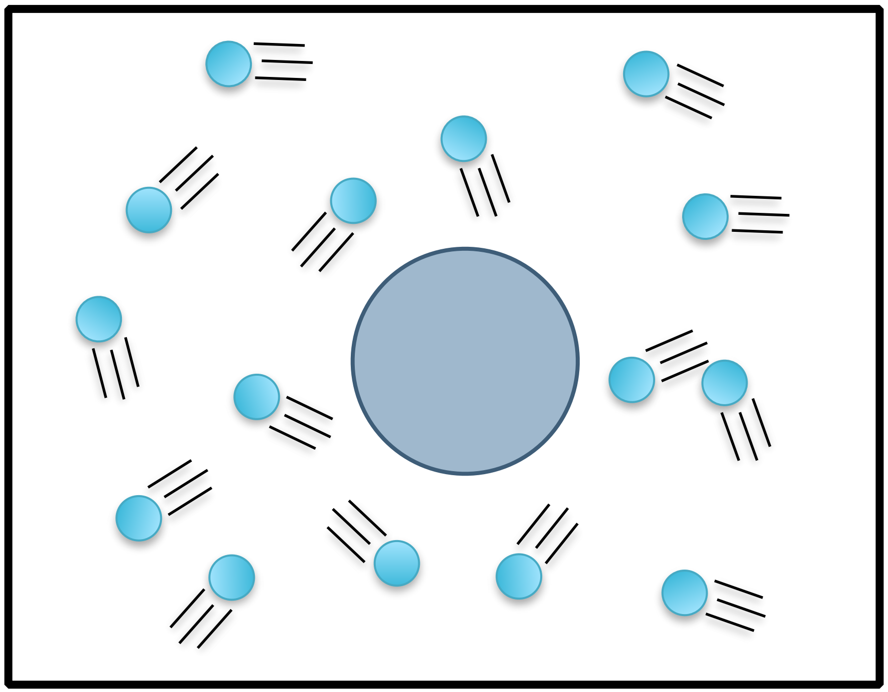
- **How diffusion models work**
- **Path-integral formulation of diffusion models**
- **“Classical limit” and beyond**

How diffusion models work

Diffusion

https://www.youtube.com/watch?v=_Owb7Nbhhkg

Brownian motion



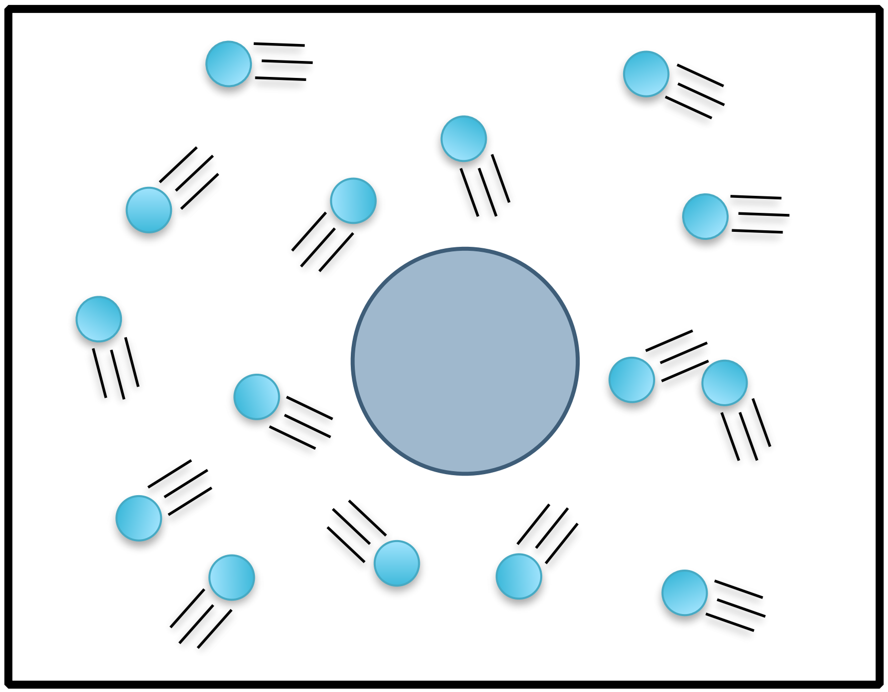
- EOM of an ink molecule in a fluid

$$\mathbf{x}_{t+\mathrm{d}t} = \mathbf{x}_t + \underline{\mathrm{d}\mathbf{w}_t}$$

Random force

Langevin equation

Brownian motion and discovery of atoms



Estimation of the Avogadro constant N_A

$$\langle(\mathbf{x}_t - \mathbf{x}_0)^2\rangle = 2Dt$$

Mean squared displacement

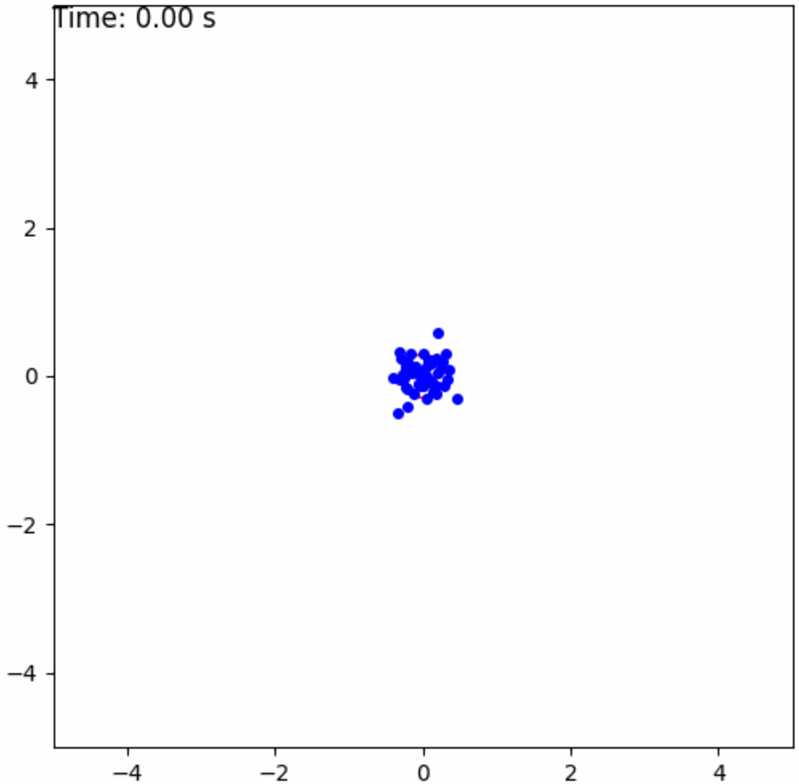
Diffusion constant $$D = \frac{RT}{N_\mathrm{A}\gamma}$$

$$\gamma = 6\pi a \eta$$

R : Gas constant
T : Temperature

η : Viscosity
a : Radius

Brownian motion and discovery of atoms



Estimation of the Avogadro constant N_A

$$\langle (\mathbf{x}_t - \mathbf{x}_0)^2 \rangle = 2Dt$$

Mean squared displacement

Diffusion constant $$D = \frac{RT}{N_\mathrm{A}\gamma}$$

$$\gamma = 6\pi a \eta$$

R : Gas constant
T : Temperature

η : Viscosity
a : Radius

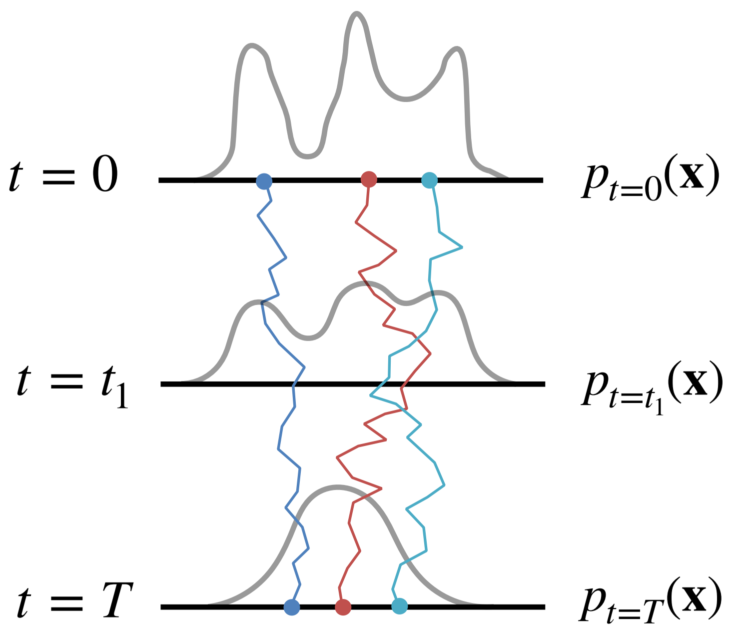
Stochastic Differential Equations

- We consider the stochastic differential equation (SDE) of the form

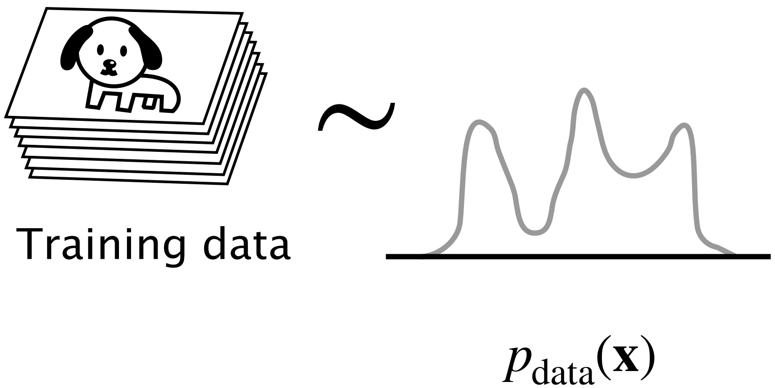
$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t$$ where $\mathbf{w}_t$ is a Wiener process

- This process is equivalent to the following Fokker–Planck equation

$$\partial_t p_t(\mathbf{x}) = -\nabla \cdot \left[\mathbf{f}(\mathbf{x}, t) p_t(\mathbf{x}) - \frac{g^2(t)}{2} \nabla p_t(\mathbf{x}) \right]$$

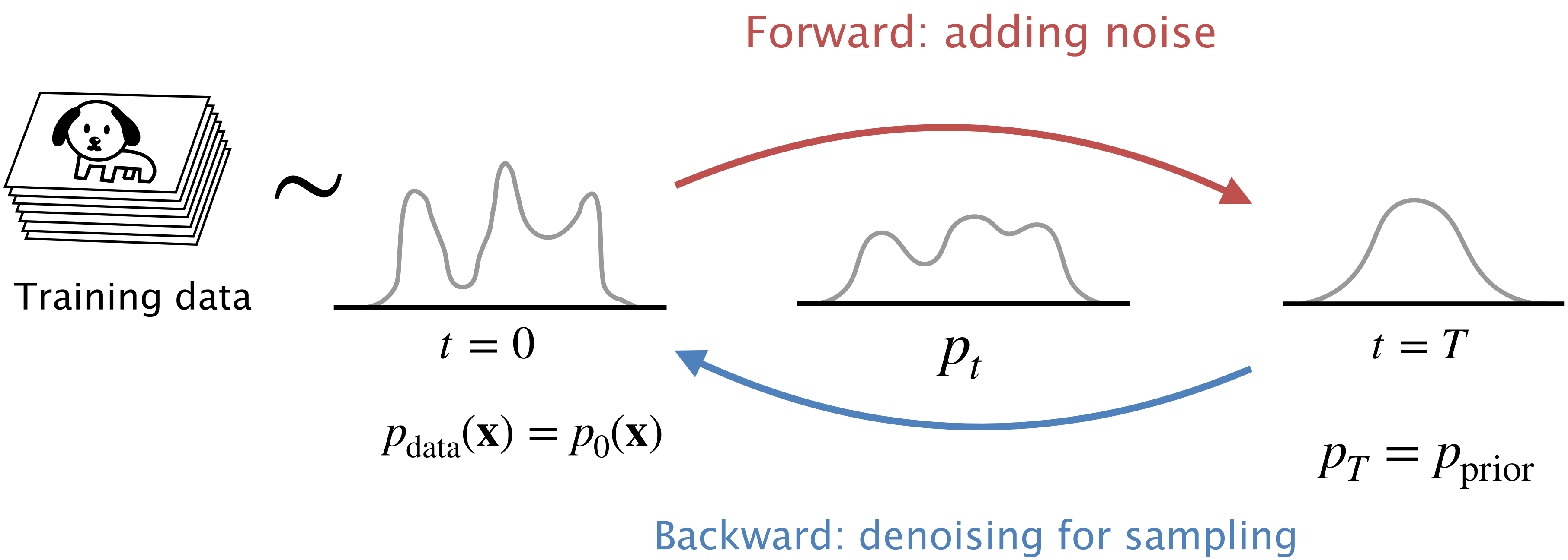


Diffusion models via SDEs [Song et. al., ICML'21]



- We'd like to sample from p_{data}
  - p_{data} is unknown
  - Even if we know p_{data}, sampling via Markov Chain Monte-Carlo (MCMC) is inefficient

Diffusion models via SDEs [Song et. al., ICML'21]



- For the forward process, one can employ, for example,

$$\partial_t p_t(\mathbf{x}) = -\nabla \cdot \left[\mathbf{f}(\mathbf{x}, t) p_t(\mathbf{x}) - \frac{g^2(t)}{2} \nabla p_t(\mathbf{x}) \right] \quad \text{with} \quad \mathbf{f}(\mathbf{x}, t) = -\beta \mathbf{x}, \quad g(t) = \sqrt{\beta}$$

The stationary distribution is $p_{\text{ss}}(\mathbf{x}) = \mathcal{N}(\mathbf{x} \,|\, \mathbf{0}, \mathbf{1})$

Diffusion models via SDEs [Song et. al., ICML'21]

- The forward process is governed by

$$\partial_t p_t(\mathbf{x}) = -\nabla \cdot \left[\mathbf{f}(\mathbf{x}, t) p_t(\mathbf{x}) - \frac{g^2(t)}{2} \nabla p_t(\mathbf{x}) \right] = \frac{g^2(t)}{2} \nabla^2 p_t(\mathbf{x}) + \cdots$$

- Since p_{data} is unknown, this process is performed for samples by

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t$$

- What is the SDE corresponding to the time-reversed FP equation?

$$\partial_t p_t(\mathbf{x}) = -\nabla \cdot \left[\mathbf{f}(\mathbf{x}, t) p_t(\mathbf{x}) - \frac{g^2(t)}{2} \nabla p_t(\mathbf{x}) - \frac{g^2(t)}{2} \nabla p_t(\mathbf{x}) + \frac{g^2(t)}{2} \nabla p_t(\mathbf{x}) \right]$$

$$= -\nabla \cdot \left[\left(\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla \log p_t(\mathbf{x}) \right) p_t(\mathbf{x}) + \frac{g^2(t)}{2} \nabla p_t(\mathbf{x}) \right]$$

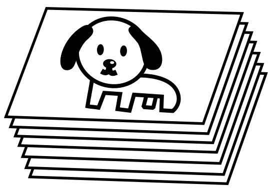
$$\Updownarrow$$

$$\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - g(t)^2 \nabla \log p_t(\mathbf{x}_t) \right] \mathrm{d}t + g(t)\mathrm{d}\tilde{\mathbf{w}}_t$$

SDE for the backward process

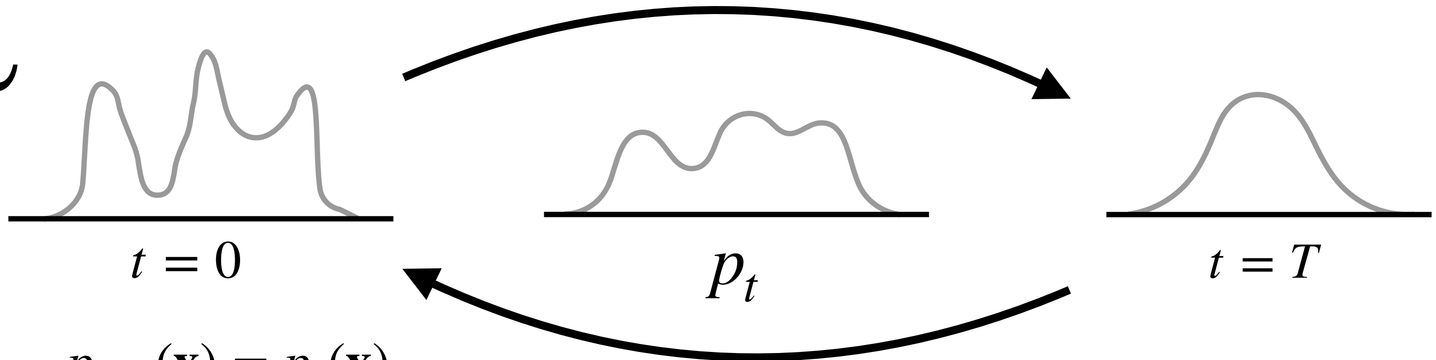
Diffusion models via SDEs [Song et. al., ICML’21]

Forward: noising $\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t$



Training data

$\sim$



$p_{\mathrm{data}}(\mathbf{x}) = p_0(\mathbf{x})$

$p_T = p_{\mathrm{prior}}$

$q_0 \simeq p_{\mathrm{data}} \longleftarrow q_T = p_{\mathrm{prior}}$

Score function

Backward: sampling $\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(t, \mathbf{x}_t) - g(t)^2 \underbrace{\nabla \log p_t(\mathbf{x}_t)}_{\simeq \mathbf{s}_\theta(\mathbf{x}_t, t)}\right]\mathrm{d}t + g(t)\mathrm{d}\tilde{\mathbf{w}}_t$

Training objective

- $\mathbf{s}_\theta(\mathbf{x}, t)$ is learned to minimize the following loss function

$$\mathscr{L}(\theta) = \int_0^T \frac{g(t)^2}{2} \mathbb{E}_{\mathbf{x} \sim p_t} \left[\left\| \mathbf{s}_\theta(\mathbf{x}_t, t) - \nabla \log p_t(\mathbf{x}_t) \right\|^2 \right] \mathrm{d}t$$

- This loss function gives the upper bound of $D_{\mathrm{KL}}(p_0 \parallel q_0)$,

$$D_{\mathrm{KL}}(p_0 \parallel q_0) \leq D_{\mathrm{KL}}(p_T \parallel q_T) + \mathscr{L}(\theta)$$

Path-integral formulation of diffusion models

Path integral in quantum mechanics

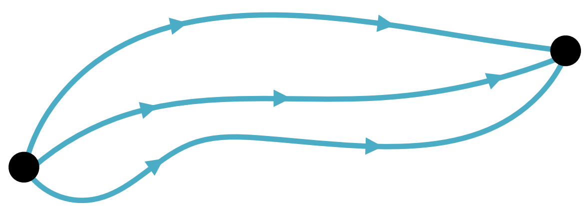
- Path integral: a formulation of quantum mechanics / QFTs

[Feynman, Rev. Mod. Phys. 20 (1948)]

- Expectation value of an observable $\mathcal{O}(\{\mathbf{x}_t\})$ is represented as

$$\langle \mathcal{O}(\mathbf{x}_t) \rangle = N \sum_{\text{paths}} e^{i\mathscr{A}[\{\mathbf{x}_t\}]/\hbar} \mathcal{O}(\mathbf{x}_t)$$

$\mathscr{A}[\{\mathbf{x}_t\}_{t \in [0,T]}]$: "Action"



- Classical mechanics: a path with least action $\delta \mathscr{A}[\{\mathbf{x}_t\}] = 0$

Path integral formulation of diffusion models

- The expectation value of $\mathcal{O}(\{\mathbf{x}_t\})$ where $\mathbf{x}_t$ obeys the SDE

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t$$

can be represented in the path-integral form as

$$\mathbb{E}\left[\mathcal{O}(\{\mathbf{x}_t\})\right] = \int [D\mathbf{x}_t]\, \mathcal{O}(\{\mathbf{x}_t\}) \underline{p_0(\mathbf{x}_0) e^{-\mathscr{A}}}$$

$=: P(\{\mathbf{x}_t\}_{t\in[0,T]})$ Path probability

"Action" $$\mathscr{A} := \int_0^T \frac{1}{2g(t)^2} \left\| \dot{\mathbf{x}}_t - \mathbf{f}(\mathbf{x}_t, t) \right\|^2 \mathrm{d}t$$

Known as Onsager-Machlup function

[Onsager-Machlup, Phys. Rev., 1953]

Backward process in path integral

- Backward SDE: $\mathrm{d}\mathbf{x}_t = \tilde{\mathbf{f}}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\tilde{\mathbf{w}}_t$

  where $\tilde{\mathbf{f}}(\mathbf{x}, t) := \mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla \log p_t(\mathbf{x})$

- Backward SDE can be derived naturally in path integral:

  The path probability can be written as

$$P(\{\mathbf{x}_t\}_{t\in[0,T]}) = p_0(\mathbf{x}_0)\, e^{-\mathscr{A}} = e^{-\tilde{\mathscr{A}}} p_T(\mathbf{x}_T)$$

  Action for backward process $\tilde{\mathscr{A}} = \int_0^T \frac{1}{2g(t)^2} \left\| \dot{\mathbf{x}}_t - \tilde{\mathbf{f}}(\mathbf{x}_t, t) \right\|^2 \mathrm{d}t$

“Classical limit” and beyond

Sampling processes of diffusion models

- Stochastic: $\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}, t) - g(t)^2 \mathbf{s}_\theta(\mathbf{x}, t)\right] \mathrm{d}t + g(t)\mathrm{d}\tilde{\mathbf{w}}_t$

- Deterministic: **Probability Flow (PF) ODE** [Song et. al., ICML'21]

$$\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \mathbf{s}_\theta(\mathbf{x}, t)\right] \mathrm{d}t$$

- These sampling methods are equivalent if the learned score is perfect, $\mathbf{s}_\theta(\mathbf{x}, t) = \nabla \log p_t(\mathbf{x})$.

Sampling processes of diffusion models

- Introducing a parameter $\mathfrak{h}$, FP equation can be written as

$$\partial_t p_t(\mathbf{x}) = -\nabla \cdot \left[\mathbf{f}(\mathbf{x},t)p_t(\mathbf{x}) - \frac{g^2(t)}{2}\nabla p_t(\mathbf{x}) - \frac{\mathfrak{h}g^2(t)}{2}\nabla p_t(\mathbf{x}) + \frac{\mathfrak{h}g^2(t)}{2}\nabla p_t(\mathbf{x})\right]$$

$$= -\nabla \cdot \left[\left(\mathbf{f}(\mathbf{x},t) - \frac{1+\mathfrak{h}}{2}g(t)^2\nabla \log p_t(\mathbf{x})\right)p_t(\mathbf{x}) + \frac{\mathfrak{h}g^2(t)}{2}\nabla p_t(\mathbf{x})\right]$$

$$\Updownarrow$$

$$\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t,t) - \frac{1+\mathfrak{h}}{2}g(t)^2\nabla \log p_t(\mathbf{x})\right]\mathrm{d}t + \sqrt{\mathfrak{h}}g(t)\mathrm{d}\tilde{\mathbf{w}}_t$$

$\mathfrak{h} = 1$ Stochastic sampling

$\mathfrak{h} = 0$ Deterministic sampling (PF-ODE)

Sampling processes of diffusion models

- Stochastic: $\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - g(t)^2 \mathbf{s}_\theta(\mathbf{x}_t, t)\right] \mathrm{d}t + g(t)\mathrm{d}\tilde{\mathbf{w}}_t$
- Deterministic: **Probability Flow (PF) ODE** [Song et. al., ICML'21]

$$\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - \frac{1}{2} g(t)^2 \mathbf{s}_\theta(\mathbf{x}_t, t)\right] \mathrm{d}t$$

  - Pro: one-to-one correspondence:
    - Faster sampling
    - The log-likelihood of an image $\mathbf{x}_0$ can be evaluated by

$$\log p_0(\mathbf{x}_0) = \log p_T(\mathbf{x}_T) + \int_0^T \nabla \cdot \left[\mathbf{f}(\mathbf{x}_t, t) - \frac{1}{2} g(t)^2 \mathbf{s}_\theta(\mathbf{x}_t, t)\right] \mathrm{d}t$$

using the solution $\{\mathbf{x}_t\}_{t \in [0,T]}$ of PF ODE

  - Con: worse sample quality

When the score is imperfect, the two sampling processes are inequivalent

"Classical limit" and beyond

- Stochastic & deterministic samplings are continuously interpolated:

$$d\mathbf{x}_t = \tilde{\mathbf{F}}_{\theta,\mathfrak{h}}(\mathbf{x}_t, t)dt + \sqrt{\mathfrak{h}}\, g(t) d\tilde{\mathbf{w}}_t$$

$\mathfrak{h} = 1$ Stochastic

$$\tilde{\mathbf{F}}_{\theta,\mathfrak{h}}(\mathbf{x}, t) := \mathbf{f}(\mathbf{x}, t) - \frac{1+\mathfrak{h}}{2} g(t)^2 \mathbf{s}_\theta(\mathbf{x}, t)$$

$\mathfrak{h} = 0$ Deterministic

- Path probability of a model is

$$Q_{\mathfrak{h}}(\{\mathbf{x}_t\}_{t\in[0,T]}) = e^{-\frac{1}{\mathfrak{h}}\tilde{\mathscr{A}}_{\theta,\mathfrak{h}}} q_T(\mathbf{x}_T) \longleftrightarrow e^{i\frac{1}{\hbar}\mathscr{A}} \quad \text{in QM}$$

$$\tilde{\mathscr{A}}_{\theta,\mathfrak{h}} := \int_0^T \frac{1}{2g(t)^2} \left\| \dot{\mathbf{x}}_t - \tilde{\mathbf{F}}_{\theta,\mathfrak{h}}(\mathbf{x}, t) \right\|^2 dt$$

- **Parameter $\mathfrak{h}$ is the counterpart of Planck's constant $\hbar$**

"Classical limit" and beyond

- Deterministic sampling appears as the "classical limit" $\mathfrak{h} \to 0$

$$Q_{\mathfrak{h}}(\{\mathbf{x}_t\}_{t\in[0,T]}) = e^{-\frac{1}{\mathfrak{h}}\tilde{\mathscr{A}}_{\theta,\mathfrak{h}}} q_T(\mathbf{x}_T) \qquad \tilde{\mathscr{A}}_{\theta,\mathfrak{h}} := \int_0^T \frac{1}{2g(t)^2} \left\| \dot{\mathbf{x}}_t - \tilde{\mathbf{F}}_{\theta,\mathfrak{h}}(\mathbf{x}, t) \right\|^2 \mathrm{d}t$$

$$\overset{\mathfrak{h}\to 0}{\to} \prod_{t\in[0,T]} \delta\left(\dot{\mathbf{x}}_t - \tilde{\mathbf{f}}_\theta^{\mathrm{PF}}(\mathbf{x}_t, t)\right) q_T(\mathbf{x}_T) \qquad \tilde{\mathbf{f}}_\theta^{\mathrm{PF}}(\mathbf{x}, t) := \mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \mathbf{s}_\theta(\mathbf{x}, t)$$

- Likelihood computation for $\mathfrak{h} \neq 0$ via **WKB expansion**
  - To the first order in $\mathfrak{h}$,

$$\log q_0^{\mathfrak{h}}(\mathbf{x}_0) = \log q_0^{\mathfrak{h}=0}(\mathbf{x}_0) + \mathfrak{h}\underbrace{\left[\delta\mathbf{x}_T \cdot \nabla \log q_T^{\mathfrak{h}=0}(\mathbf{x}_T) + \int_0^T \nabla \cdot \delta\tilde{\mathbf{f}}_\theta^{\mathrm{PF}}(\mathbf{x}_t, \delta\mathbf{x}_t, t)\mathrm{d}t\right]}_{\text{First-order correction to log-likelihood}}$$

First-order correction to log-likelihood

where $\{\mathbf{x}_t, \delta\mathbf{x}_t\}$ are the solution of the following ODE with $\mathbf{x}_{t=0} = \mathbf{x}_0$, $\delta\mathbf{x}_{t=0} = \mathbf{0}$

$$\begin{cases} \dot{\mathbf{x}}_t = \tilde{\mathbf{f}}_\theta^{\mathrm{PF}}(\mathbf{x}_t, t) \\ \dot{\delta\mathbf{x}}_t = \delta\tilde{\mathbf{f}}_\theta^{\mathrm{PF}}(\mathbf{x}_t, \delta\mathbf{x}_t, t) \end{cases} \qquad \delta\tilde{\mathbf{f}}_\theta^{\mathrm{PF}}(\mathbf{x}_t, \delta\mathbf{x}_t, t) := (\delta\mathbf{x}_t \cdot \nabla)\tilde{\mathbf{f}}_\theta^{\mathrm{PF}}(\mathbf{x}_t, t) - \frac{\mathfrak{g}(t)^2}{2}[\mathbf{s}_\theta(\mathbf{x}_t, t) - \nabla \log q_t^{\mathfrak{h}=0}(\mathbf{x}_t)]$$

Noise improves sample quality

Swiss-roll

SDE (NLL)	tol	1st-corr	errors
simple (1.39± 0.05)	1e−3	-0.31±0.21	0.13±0.00
	1e−5	-0.44±0.38	0.13±0.00
cosine (1.42± 0.02)	1e−3	-1.59±0.57	0.35±0.00
	1e−5	-3.27±1.11	0.37±0.02

25-gaussian

SDE (NLL)	tol	1st-corr	errors
simple (-1.22± 0.01)	1e−3	-3.64±0.49	0.31±0.00
	1e−5	-3.61±0.64	0.32±0.01
cosine (-1.71± 0.02)	1e−3	-17.57±5.56	0.70±0.01
	1e−5	-19.65±17.46	0.67±0.03

$O(h^1)$ correction to Negative Log Likelihood (NLL) $\mathbb{E}\left[-\log q_0\right]$

- Negative correction → noise improves sample quality

Path integral formulation of diffusion models

- Useful for physicists in understanding various aspects of diffusion models: backward process, training objective

- Deterministic sampling by PF ODE appears as "classical limit"

$$Q_{\mathfrak{h}}(\{\mathbf{x}_t\}_{t\in[0,T]}) = e^{-\frac{1}{\mathfrak{h}}\tilde{\mathscr{A}}_{\theta,\mathfrak{h}}} q_T(\mathbf{x}_T) \longleftrightarrow e^{i\frac{1}{\hbar}\mathscr{A}} \quad \text{in QM}$$

- Likelihood evaluation via WKB expansion

- Physics methods for analyzing generative AI models

Quantum × AI in the next 100 years?

- Discrete diffusion models are used to build Large Language models
  - Mercury Coder (Feb. 2025 –)
    Gemini Diffusion (May. 2025 –)
- Based on “classical” stochastic dynamics
- A possible new mechanism of LLMs based on quantum dynamics?