

Deep Generative Models

Lecture 12

Roman Isachenko

Moscow Institute of Physics and Technology
Yandex School of Data Analysis

2024, Autumn

Recap of previous lecture

Continuous-in-time dynamics

Consider Ordinary Differential Equation (ODE)

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{f}_{\theta}(\mathbf{x}(t), t); \quad \text{with initial condition } \mathbf{x}(0) = \mathbf{x}_0.$$

$$\mathbf{x}(1) = \int_0^1 \mathbf{f}_{\theta}(\mathbf{x}(t), t) dt + \mathbf{x}_0$$

Here $\mathbf{f}_{\theta} : \mathbb{R}^m \times [0, 1] \rightarrow \mathbb{R}^m$ is a vector field.

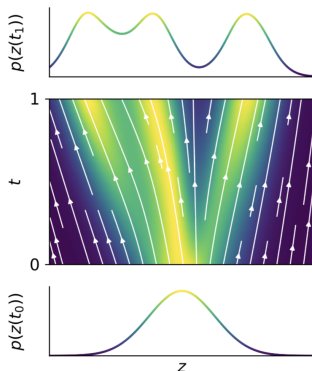
Euler update step

$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \Delta t \cdot \mathbf{f}_{\theta}(\mathbf{x}(t), t)$$

- ▶ Euler method is the simplest version of the ODEsolve that is unstable in practice.
- ▶ It is possible to use more sophisticated numerical methods instead of Euler (e.x. Runge-Kutta methods).

Recap of previous lecture

- ▶ $\mathbf{x}(0) \sim p(\mathbf{x}(0))$.
- ▶ $\mathbf{x}(1) \sim p(\mathbf{x}(1))$.
- ▶ $p_t(\mathbf{x}) = p(\mathbf{x}, t)$ is the **probability path** between $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$.
- ▶ $p_0(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$ is the base distribution and $p_1(\mathbf{x}) = \pi(\mathbf{x})$ is the data distribution.



Theorem (Picard)

If \mathbf{f} is uniformly Lipschitz continuous in \mathbf{x} and continuous in t , then the ODE has a **unique** solution.

$$\mathbf{x}(1) = \mathbf{x}(0) + \int_0^1 \mathbf{f}_\theta(\mathbf{x}(t), t) dt; \quad \mathbf{x}(0) = \mathbf{x}(1) + \int_1^0 \mathbf{f}_\theta(\mathbf{x}(t), t) dt$$

Recap of previous lecture

Theorem (continuity equation)

If \mathbf{f} is uniformly Lipschitz continuous in \mathbf{x} and continuous in t , then

$$\frac{d \log p_t(\mathbf{x}(t))}{dt} = -\text{tr} \left(\frac{\partial \mathbf{f}(\mathbf{x}(t), t)}{\partial \mathbf{x}(t)} \right)$$

$$\log p_1(\mathbf{x}(1)) = \log p_0(\mathbf{x}(0)) - \int_0^1 \text{tr} \left(\frac{\partial \mathbf{f}(\mathbf{x}(t), t)}{\partial \mathbf{x}(t)} \right) dt.$$

- ▶ **Discrete-in-time NF**: evaluation of determinant of the Jacobian costs $O(m^3)$ (we need invertible \mathbf{f}).
- ▶ **Continuous-in-time NF**: getting the trace of the Jacobian costs $O(m^2)$ (we need smooth \mathbf{f}).

Hutchinson's trace estimator

$$\log p_1(\mathbf{x}(1)) = \log p_0(\mathbf{x}(0)) - \mathbb{E}_{p(\epsilon)} \int_0^1 \left[\epsilon^T \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \epsilon \right] dt.$$

Recap of previous lecture

Forward pass (Loss function)

$$L(\mathbf{x}) = -\log p_1(\mathbf{x}(1)|\boldsymbol{\theta}) = -\log p_0(\mathbf{x}(0)) + \int_0^1 \text{tr} \left(\frac{\partial \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}(t), t)}{\partial \mathbf{x}(t)} \right) dt$$

Adjoint functions

$$\mathbf{a}_{\mathbf{x}}(t) = \frac{\partial L}{\partial \mathbf{x}(t)}; \quad \mathbf{a}_{\boldsymbol{\theta}}(t) = \frac{\partial L}{\partial \boldsymbol{\theta}(t)}.$$

Theorem (Pontryagin)

$$\frac{d\mathbf{a}_{\mathbf{x}}(t)}{dt} = -\mathbf{a}_{\mathbf{x}}(t)^T \cdot \frac{\partial \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}(t), t)}{\partial \mathbf{x}}; \quad \frac{d\mathbf{a}_{\boldsymbol{\theta}}(t)}{dt} = -\mathbf{a}_{\mathbf{x}}(t)^T \cdot \frac{\partial \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}(t), t)}{\partial \boldsymbol{\theta}}.$$

$$\frac{\partial L}{\partial \boldsymbol{\theta}(0)} = \mathbf{a}_{\boldsymbol{\theta}}(0) = - \int_1^0 \mathbf{a}_{\mathbf{x}}(t)^T \frac{\partial \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}(t), t)}{\partial \boldsymbol{\theta}(t)} dt + 0$$

$$\frac{\partial L}{\partial \mathbf{x}(0)} = \mathbf{a}_{\mathbf{x}}(0) = - \int_1^0 \mathbf{a}_{\mathbf{x}}(t)^T \frac{\partial \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}(t), t)}{\partial \mathbf{x}(t)} dt + \frac{\partial L}{\partial \mathbf{x}(1)}$$

Recap of previous lecture

Forward pass

$$\mathbf{x}(1) = \mathbf{x}(0) + \int_0^1 \mathbf{f}_{\theta}(\mathbf{x}(t), t) dt \Rightarrow \text{ODE Solver}$$

Backward pass

$$\left. \begin{aligned} \frac{\partial L}{\partial \theta(0)} = \mathbf{a}_{\theta}(0) &= - \int_1^0 \mathbf{a}_{\mathbf{x}}(t)^T \frac{\partial \mathbf{f}_{\theta}(\mathbf{x}(t), t)}{\partial \theta(t)} dt + 0 \\ \frac{\partial L}{\partial \mathbf{x}(0)} = \mathbf{a}_{\mathbf{x}}(0) &= - \int_1^0 \mathbf{a}_{\mathbf{x}}(t)^T \frac{\partial \mathbf{f}_{\theta}(\mathbf{x}(t), t)}{\partial \mathbf{x}(t)} dt + \frac{\partial L}{\partial \mathbf{x}(1)} \\ \mathbf{x}(0) &= - \int_0^1 \mathbf{f}_{\theta}(\mathbf{x}(t), t) dt + \mathbf{x}(1). \end{aligned} \right\} \Rightarrow \text{ODE Solver}$$

Note: These scary formulas are the standard backprop in the discrete case.

Outline

1. SDE basics
2. Probability flow ODE
3. Reverse SDE
4. Diffusion and Score matching SDEs
5. Score-based generative models through SDEs

Outline

1. SDE basics
2. Probability flow ODE
3. Reverse SDE
4. Diffusion and Score matching SDEs
5. Score-based generative models through SDEs

Stochastic differential equation (SDE)

Let define stochastic process $\mathbf{x}(t)$ with initial condition $\mathbf{x}(0) \sim p_0(\mathbf{x}) = \pi(\mathbf{x})$:

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

- ▶ $\mathbf{f}(\mathbf{x}, t) : \mathbb{R}^m \times [0, 1] \rightarrow \mathbb{R}^m$ is the **drift** function of $\mathbf{x}(t)$.
- ▶ $g(t) : \mathbb{R} \rightarrow \mathbb{R}$ is the **diffusion** function of $\mathbf{x}(t)$.
- ▶ $\mathbf{w}(t)$ is the standard Wiener process (Brownian motion):
 1. $\mathbf{w}(0) = 0$ (almost surely);
 2. $\mathbf{w}(t)$ has independent increments;
 3. $\mathbf{w}(t) - \mathbf{w}(s) \sim \mathcal{N}(0, (t - s)\mathbf{I})$, for $t > s$.
- ▶ $d\mathbf{w} = \mathbf{w}(t + dt) - \mathbf{w}(t) = \mathcal{N}(0, \mathbf{I} \cdot dt) = \epsilon \cdot \sqrt{dt}$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.
- ▶ If $g(t) = 0$ we get standard ODE.

Stochastic differential equation (SDE)

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

- ▶ In contrast to ODE, initial condition $\mathbf{x}(0)$ does not uniquely determine the process trajectory.
- ▶ We have two sources of randomness: initial distribution $p_0(\mathbf{x})$ and Wiener process $\mathbf{w}(t)$.

Discretization of SDE (Euler method)

$$\mathbf{x}(t + dt) = \mathbf{x}(t) + \mathbf{f}(\mathbf{x}(t), t) \cdot dt + g(t) \cdot \epsilon \cdot \sqrt{dt}$$

If $dt = 1$, then

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{f}(\mathbf{x}_t, t) + g(t) \cdot \epsilon$$

- ▶ At each moment t we have the density $p_t(\mathbf{x}) = p(\mathbf{x}, t)$.
- ▶ $p : \mathbb{R}^m \times [0, 1] \rightarrow \mathbb{R}_+$ is a **probability path** between $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$.
- ▶ How to get the distribution path $p_t(\mathbf{x})$ for $\mathbf{x}(t)$?

Stochastic differential equation (SDE)

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}, \quad d\mathbf{w} = \boldsymbol{\epsilon} \cdot \sqrt{dt}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}).$$

Theorem (Kolmogorov-Fokker-Planck)

Evolution of the distribution $p_t(\mathbf{x})$ is given by the following equation:

$$\frac{\partial p_t(\mathbf{x})}{\partial t} = -\operatorname{div}(\mathbf{f}(\mathbf{x}, t)p_t(\mathbf{x})) + \frac{1}{2}g^2(t)\Delta_{\mathbf{x}}p_t(\mathbf{x})$$

Here

$$\operatorname{div}(\mathbf{v}) = \sum_{i=1}^m \frac{\partial v_i(\mathbf{x})}{\partial x_i} = \operatorname{tr} \left(\frac{\partial \mathbf{v}(\mathbf{x})}{\partial \mathbf{x}} \right)$$

$$\Delta_{\mathbf{x}}p_t(\mathbf{x}) = \sum_{i=1}^m \frac{\partial^2 p_t(\mathbf{x})}{\partial x_i^2} = \operatorname{tr} \left(\frac{\partial^2 p_t(\mathbf{x})}{\partial \mathbf{x}^2} \right)$$

$$\frac{\partial p_t(\mathbf{x})}{\partial t} = \operatorname{tr} \left(-\frac{\partial}{\partial \mathbf{x}} [\mathbf{f}(\mathbf{x}, t)p_t(\mathbf{x})] + \frac{1}{2}g^2(t) \frac{\partial^2 p_t(\mathbf{x})}{\partial \mathbf{x}^2} \right)$$

Stochastic differential equation (SDE)

Theorem (Kolmogorov-Fokker-Planck)

$$\frac{\partial p_t(\mathbf{x})}{\partial t} = \text{tr} \left(-\frac{\partial}{\partial \mathbf{x}} [\mathbf{f}(\mathbf{x}, t) p_t(\mathbf{x})] + \frac{1}{2} g^2(t) \frac{\partial^2 p_t(\mathbf{x})}{\partial \mathbf{x}^2} \right)$$

- ▶ KFP theorem does not define the SDE uniquely in general case.
- ▶ This is the generalization of continuity equation that we used in continuous-in-time NF:

$$\frac{d \log p_t(\mathbf{x}(t))}{dt} = -\text{tr} \left(\frac{\partial \mathbf{f}(\mathbf{x}, t)}{\partial \mathbf{x}} \right).$$

Langevin SDE (special case)

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

$$d\mathbf{x} = \frac{1}{2} \frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) dt + \mathbf{1} \cdot d\mathbf{w}$$

Let apply KFP theorem to this SDE.

Langevin SDE (special case)

$$d\mathbf{x} = \frac{1}{2} \frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) dt + 1 \cdot d\mathbf{w}$$

$$\begin{aligned} \frac{\partial p_t(\mathbf{x})}{\partial t} &= \text{tr} \left(-\frac{\partial}{\partial \mathbf{x}} \left[p_t(\mathbf{x}) \frac{1}{2} \frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right] + \frac{1}{2} \frac{\partial^2 p_t(\mathbf{x})}{\partial \mathbf{x}^2} \right) = \\ &= \text{tr} \left(-\frac{\partial}{\partial \mathbf{x}} \left[\frac{1}{2} \frac{\partial}{\partial \mathbf{x}} p_t(\mathbf{x}) \right] + \frac{1}{2} \frac{\partial^2 p_t(\mathbf{x})}{\partial \mathbf{x}^2} \right) = 0 \end{aligned}$$

The density $p_t(\mathbf{x}) = \text{const}(t)$!

If $\mathbf{x}(0) \sim p_0(\mathbf{x})$, then $\mathbf{x}(t) \sim p_0(\mathbf{x})$.

Discretized Langevin SDE

$$\mathbf{x}_{t+1} - \mathbf{x}_t = \frac{\eta}{2} \cdot \frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) + \sqrt{\eta} \cdot \epsilon, \quad \eta \approx dt.$$

Langevin dynamic

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \frac{\eta}{2} \cdot \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta) + \sqrt{\eta} \cdot \epsilon, \quad \eta \approx dt.$$

Outline

1. SDE basics
2. Probability flow ODE
3. Reverse SDE
4. Diffusion and Score matching SDEs
5. Score-based generative models through SDEs

Probability flow ODE

ODE and continuity equation

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt$$

$$\frac{d \log p_t(\mathbf{x}(t))}{dt} = -\text{tr} \left(\frac{\partial \mathbf{f}_\theta(\mathbf{x}, t)}{\partial \mathbf{x}} \right) \Leftrightarrow \frac{\partial p_t(\mathbf{x})}{\partial t} = -\text{div}(\mathbf{f}(\mathbf{x}, t)p_t(\mathbf{x}))$$

The only source of stochasticity is the distribution $p_0(\mathbf{x})$.

SDE and KFP equation

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

$$\frac{\partial p_t(\mathbf{x})}{\partial t} = -\text{div}(\mathbf{f}(\mathbf{x}, t)p_t(\mathbf{x})) + \frac{1}{2}g^2(t)\Delta_{\mathbf{x}}p_t(\mathbf{x})$$

We have two sources of randomness: initial distribution $p_0(\mathbf{x})$ and Wiener process $\mathbf{w}(t)$.

Probability flow ODE

Theorem

Assume SDE $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$ induces the probability path $p_t(\mathbf{x})$. Then there exists ODE with identical probability path $p_t(\mathbf{x})$ of the form

$$d\mathbf{x} = \left(\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g^2(t)\frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right) dt$$

Proof

$$\begin{aligned} \frac{\partial p_t(\mathbf{x})}{\partial t} &= \text{tr} \left(-\frac{\partial}{\partial \mathbf{x}} [\mathbf{f}(\mathbf{x}, t)p_t(\mathbf{x})] + \frac{1}{2}g^2(t)\frac{\partial^2 p_t(\mathbf{x})}{\partial \mathbf{x}^2} \right) = \\ &= \text{tr} \left(-\frac{\partial}{\partial \mathbf{x}} \left[\mathbf{f}(\mathbf{x}, t)p_t(\mathbf{x}) - \frac{1}{2}g^2(t)\frac{\partial p_t(\mathbf{x})}{\partial \mathbf{x}} \right] \right) = \\ &= \text{tr} \left(-\frac{\partial}{\partial \mathbf{x}} \left[\mathbf{f}(\mathbf{x}, t)p_t(\mathbf{x}) - \frac{1}{2}g^2(t)p_t(\mathbf{x})\frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right] \right) = \\ &= \text{tr} \left(-\frac{\partial}{\partial \mathbf{x}} \left[\left(\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g^2(t)\frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right) p_t(\mathbf{x}) \right] \right) \end{aligned}$$

Probability flow ODE

Theorem

Assume SDE $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$ induces the probability path $p_t(\mathbf{x})$. Then there exists ODE with identical probabilities distribution $p_t(\mathbf{x})$ of the form

$$d\mathbf{x} = \left(\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g^2(t)\frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right) dt$$

Proof (continued)

$$\begin{aligned} \frac{\partial p_t(\mathbf{x})}{\partial t} &= \text{tr} \left(-\frac{\partial}{\partial \mathbf{x}} \left[\left(\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g^2(t)\frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right) p_t(\mathbf{x}) \right] \right) = \\ &= \text{tr} \left(-\frac{\partial}{\partial \mathbf{x}} \left[\tilde{\mathbf{f}}(\mathbf{x}, t) p_t(\mathbf{x}) \right] \right) \end{aligned}$$

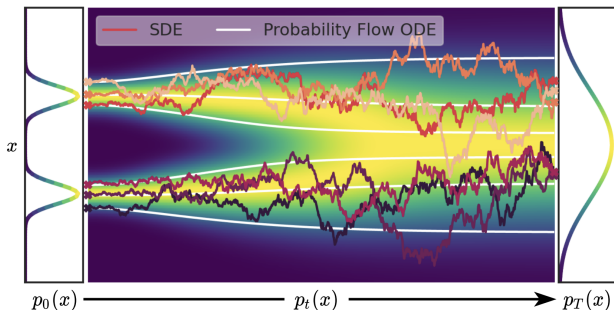
$$d\mathbf{x} = \tilde{\mathbf{f}}(\mathbf{x}, t)dt + 0 \cdot d\mathbf{w} = \left(\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g^2(t)\frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right) dt$$

Probability flow ODE

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w} - \text{SDE}$$

$$d\mathbf{x} = \left(\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g^2(t)\frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right) dt - \text{probability flow ODE}$$

- ▶ The term $\mathbf{s}(\mathbf{x}, t) = \frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x})$ is a score function for continuous time.
- ▶ ODE has more stable trajectories.



Outline

1. SDE basics
2. Probability flow ODE
3. Reverse SDE
4. Diffusion and Score matching SDEs
5. Score-based generative models through SDEs

Reverse SDE

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt, \quad \mathbf{x}(t + dt) = \mathbf{x}(t) + \mathbf{f}(\mathbf{x}, t)dt$$

Here dt could be > 0 or < 0 .

Reverse ODE

Let $\tau = 1 - t$ ($d\tau = -dt$).

$$d\mathbf{x} = -\mathbf{f}(\mathbf{x}, 1 - \tau)d\tau$$

- ▶ How to revert SDE $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$?
- ▶ Wiener process gives the randomness that we have to revert.

Theorem

There exists the reverse SDE for the SDE $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$ that has the following form

$$d\mathbf{x} = \left(\mathbf{f}(\mathbf{x}, t) - g^2(t) \frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right) dt + g(t)d\mathbf{w}$$

with $dt < 0$.

Reverse SDE

Theorem

There exists the reverse SDE for the SDE $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$ that has the following form

$$d\mathbf{x} = \left(\mathbf{f}(\mathbf{x}, t) - g^2(t) \frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right) dt + g(t)d\mathbf{w}$$

with $dt < 0$.

Note: Here we also see the score function $\mathbf{s}(\mathbf{x}, t) = \frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x})$.

Sketch of the proof

- ▶ Convert initial SDE to probability flow ODE.
- ▶ Revert probability flow ODE.
- ▶ Convert reverse probability flow ODE to reverse SDE.

Reverse SDE

Proof

- Convert initial SDE to probability flow ODE

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

$$d\mathbf{x} = \left(\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g^2(t)\frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right) dt$$

- Revert probability flow ODE

$$d\mathbf{x} = \left(\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g^2(t)\frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right) dt$$

$$d\mathbf{x} = \left(-\mathbf{f}(\mathbf{x}, 1 - \tau) + \frac{1}{2}g^2(1 - \tau)\frac{\partial}{\partial \mathbf{x}} \log p_{1-\tau}(\mathbf{x}) \right) d\tau$$

- Convert reverse probability flow ODE to reverse SDE

$$d\mathbf{x} = \left(-\mathbf{f}(\mathbf{x}, 1 - \tau) + \frac{1}{2}g^2(1 - \tau)\frac{\partial}{\partial \mathbf{x}} \log p_{1-\tau}(\mathbf{x}) \right) d\tau$$

$$d\mathbf{x} = \left(-\mathbf{f}(\mathbf{x}, 1 - \tau) + g^2(1 - \tau)\frac{\partial}{\partial \mathbf{x}} \log p_{1-\tau}(\mathbf{x}) \right) d\tau + g(1 - \tau)d\mathbf{w}$$

Reverse SDE

Theorem

There exists the reverse SDE for the SDE $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$ that has the following form

$$d\mathbf{x} = \left(\mathbf{f}(\mathbf{x}, t) - g^2(t) \frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right) dt + g(t)d\mathbf{w}$$

with $dt < 0$.

Proof (continued)

$$d\mathbf{x} = \left(-\mathbf{f}(\mathbf{x}, 1 - \tau) + g^2(1 - \tau) \frac{\partial}{\partial \mathbf{x}} \log p_{1-\tau}(\mathbf{x}) \right) d\tau + g(1 - \tau)d\mathbf{w}$$

$$d\mathbf{x} = \left(\mathbf{f}(\mathbf{x}, t) - g^2(t) \frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right) dt + g(t)d\mathbf{w}$$

Here $d\tau > 0$ and $dt < 0$.

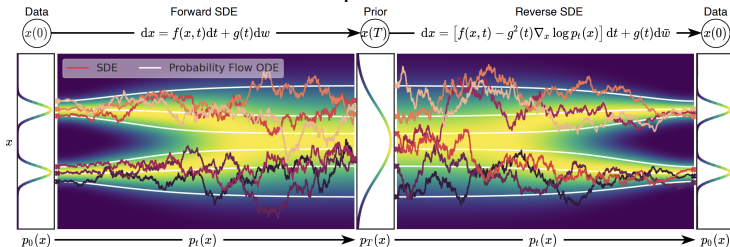
Reverse SDE

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w} - \text{SDE}$$

$$d\mathbf{x} = \left(\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g^2(t)\frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right) dt - \text{probability flow ODE}$$

$$d\mathbf{x} = \left(\mathbf{f}(\mathbf{x}, t) - g^2(t)\frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right) dt + g(t)d\mathbf{w} - \text{reverse SDE}$$

- ▶ We got the way to transform one distribution to another via SDE with some probability path $p_t(\mathbf{x})$.
- ▶ We are able to revert this process with the score function.



Outline

1. SDE basics
2. Probability flow ODE
3. Reverse SDE
4. Diffusion and Score matching SDEs
5. Score-based generative models through SDEs

Score matching SDE

Denoising score matching

$$\mathbf{x}_t = \mathbf{x} + \sigma_t \cdot \epsilon_t, \quad q(\mathbf{x}_t | \mathbf{x}) = \mathcal{N}(\mathbf{x}, \sigma_t^2 \cdot \mathbf{I})$$

$$\mathbf{x}_{t-1} = \mathbf{x} + \sigma_{t-1} \cdot \epsilon_{t-1}, \quad q(\mathbf{x}_{t-1} | \mathbf{x}) = \mathcal{N}(\mathbf{x}, \sigma_{t-1}^2 \cdot \mathbf{I})$$

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \sqrt{\sigma_t^2 - \sigma_{t-1}^2} \cdot \epsilon, \quad q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_{t-1}, (\sigma_t^2 - \sigma_{t-1}^2) \cdot \mathbf{I})$$

Let turn this Markov chain to the continuous stochastic process $\mathbf{x}(t)$ taking $T \rightarrow \infty$:

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{x}(t - dt) + \sqrt{\sigma^2(t) - \sigma^2(t - dt)} \cdot \epsilon \\ &= \mathbf{x}(t - dt) + \sqrt{\frac{\sigma^2(t) - \sigma^2(t - dt)}{dt}} dt \cdot \epsilon \\ &= \mathbf{x}(t - dt) + \sqrt{\frac{d[\sigma^2(t)]}{dt}} \cdot d\mathbf{w} \end{aligned}$$

Score matching SDE

$$\mathbf{x}(t) = \mathbf{x}(t - dt) + \sqrt{\frac{d[\sigma^2(t)]}{dt}} \cdot d\mathbf{w}$$

Variance Exploding SDE

$$d\mathbf{x} = \sqrt{\frac{d[\sigma^2(t)]}{dt}} \cdot d\mathbf{w}$$

$\sigma(t)$ is a monotonically increasing function.

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}, \quad \mathbf{f}(\mathbf{x}, t) = 0, \quad g(t) = \sqrt{\frac{d[\sigma^2(t)]}{dt}}$$

$$d\mathbf{x} = \left(-\frac{1}{2} \frac{d[\sigma^2(t)]}{dt} \frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right) dt - \text{probability flow ODE}$$

$$d\mathbf{x} = \left(-\frac{d[\sigma^2(t)]}{dt} \frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}) \right) dt + \sqrt{\frac{d[\sigma^2(t)]}{dt}} d\mathbf{w} - \text{reverse SDE}$$

Diffusion SDE

Denoising Diffusion

$$\mathbf{x}_t = \sqrt{1 - \beta_t} \cdot \mathbf{x}_{t-1} + \sqrt{\beta_t} \cdot \epsilon, \quad q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1 - \beta_t} \cdot \mathbf{x}_{t-1}, \beta_t \cdot \mathbf{I})$$

Let turn this Markov chain to the continuous stochastic process taking $T \rightarrow \infty$ and taking $\beta(\frac{t}{T}) = \beta_t \cdot T$

$$\begin{aligned} \mathbf{x}(t) &= \sqrt{1 - \beta(t)dt} \cdot \mathbf{x}(t - dt) + \sqrt{\beta(t)dt} \cdot \epsilon \approx \\ &\approx (1 - \frac{1}{2}\beta(t)dt) \cdot \mathbf{x}(t - dt) + \sqrt{\beta(t)dt} \cdot \epsilon = \\ &= \mathbf{x}(t - dt) - \frac{1}{2}\beta(t)\mathbf{x}(t - dt)dt + \sqrt{\beta(t)} \cdot d\mathbf{w} \end{aligned}$$

Variance Preserving SDE

$$d\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}(t)dt + \sqrt{\beta(t)} \cdot d\mathbf{w}$$

Diffusion SDE

Variance Preserving SDE

$$d\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}(t)dt + \sqrt{\beta(t)} \cdot d\mathbf{w}$$

$$\mathbf{f}(\mathbf{x}, t) = -\frac{1}{2}\beta(t)\mathbf{x}(t), \quad g(t) = \sqrt{\beta(t)}$$

Variance is preserved if $\mathbf{x}(0)$ has a unit variance.

$$d\mathbf{x} = \left(-\frac{1}{2}\beta(t)\mathbf{x}(t) - \frac{1}{2}\beta(t)\frac{\partial}{\partial\mathbf{x}}\log p_t(\mathbf{x}) \right) dt - \text{probability flow ODE}$$

$$d\mathbf{x} = \left(-\frac{1}{2}\beta(t)\mathbf{x}(t) - \beta(t)\frac{\partial}{\partial\mathbf{x}}\log p_t(\mathbf{x}) \right) dt + \sqrt{\beta(t)}d\mathbf{w} - \text{reverse SDE}$$

Diffusion SDE

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

Variance Exploding SDE (NCSN)

$$d\mathbf{x} = \sqrt{\frac{d[\sigma^2(t)]}{dt}} \cdot d\mathbf{w}, \quad \mathbf{f}(\mathbf{x}, t) = 0, \quad g(t) = \sqrt{\frac{d[\sigma^2(t)]}{dt}}$$

Variance Preserving SDE (DDPM)

$$d\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}(t)dt + \sqrt{\beta(t)} \cdot d\mathbf{w}$$

$$\mathbf{f}(\mathbf{x}, t) = -\frac{1}{2}\beta(t)\mathbf{x}(t), \quad g(t) = \sqrt{\beta(t)}$$

Is it possible to train score-based generative model (DDPM or NCSN) in continuous time?

Song Y., et al. Score-Based Generative Modeling through Stochastic Differential Equations, 2020

Outline

1. SDE basics
2. Probability flow ODE
3. Reverse SDE
4. Diffusion and Score matching SDEs
5. Score-based generative models through SDEs

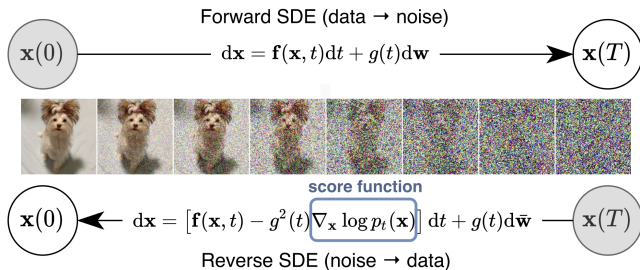
Score-based generative models through SDEs

Discrete-in-time objective

$$\mathbb{E}_{\pi(\mathbf{x}_0)} \mathbb{E}_{t \sim U\{1, T\}} \mathbb{E}_{q(\mathbf{x}_t | \mathbf{x}_0)} \left\| \mathbf{s}_{\theta, t}(\mathbf{x}_t) - \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t | \mathbf{x}_0) \right\|_2^2$$

Continuous-in-time objective

$$\mathbb{E}_{\pi(\mathbf{x}(0))} \mathbb{E}_{t \sim U[0, 1]} \mathbb{E}_{q(\mathbf{x}(t) | \mathbf{x}(0))} \left\| \mathbf{s}_{\theta}(\mathbf{x}(t), t) - \nabla_{\mathbf{x}(t)} \log q(\mathbf{x}(t) | \mathbf{x}(0)) \right\|_2^2$$



Score-based generative models through SDEs

Continuous-in-time objective

$$\mathbb{E}_{\pi(\mathbf{x}(0))} \mathbb{E}_{t \sim U[0,1]} \mathbb{E}_{q(\mathbf{x}(t)|\mathbf{x}(0))} \left\| \mathbf{s}_{\theta}(\mathbf{x}(t), t) - \nabla_{\mathbf{x}(t)} \log q(\mathbf{x}(t)|\mathbf{x}(0)) \right\|_2^2$$

$$q(\mathbf{x}(t)|\mathbf{x}(0)) = \mathcal{N}\left(\boldsymbol{\mu}(\mathbf{x}(t), \mathbf{x}(0)), \boldsymbol{\Sigma}(\mathbf{x}(t), \mathbf{x}(0))\right)$$

Theorem

Moments of the SDE $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$ satisfies the equations

$$\frac{d\boldsymbol{\mu}(\mathbf{x}(t), \mathbf{x}(0))}{dt} = \mathbb{E}[\mathbf{f}(\mathbf{x}(t), t)|\mathbf{x}(0)]$$

$$\frac{d\boldsymbol{\Sigma}(\mathbf{x}(t), \mathbf{x}(0))}{dt} = \mathbb{E}\left[\mathbf{f} \cdot (\mathbf{x}(t) - \boldsymbol{\mu})^T + (\mathbf{x}(t) - \boldsymbol{\mu}) \cdot \mathbf{f}^T | \mathbf{x}(0)\right] + g^2(t) \cdot \mathbf{I}$$

Let prove the first one.

Score-based generative models through SDEs

Theorem

$$\frac{d\mu(\mathbf{x}(t), \mathbf{x}(0))}{dt} = \mathbb{E} [\mathbf{f}(\mathbf{x}(t), t) | \mathbf{x}(0)]$$

Proof

$$\begin{aligned}\mathbb{E} [d\mathbf{x} | \mathbf{x}(0)] &= \mathbb{E} [\mathbf{f}(\mathbf{x}, t) dt | \mathbf{x}(0)] + \mathbb{E} [g(t) d\mathbf{w} | \mathbf{x}(0)] \\ &= \mathbb{E} [\mathbf{f}(\mathbf{x}, t) | \mathbf{x}(0)] dt + g(t) \mathbb{E} [d\mathbf{w} | \mathbf{x}(0)] \\ &= \mathbb{E} [\mathbf{f}(\mathbf{x}, t) | \mathbf{x}(0)] dt\end{aligned}$$

$$\frac{d\mathbb{E} [\mathbf{x} | \mathbf{x}(0)]}{dt} = \frac{d\mu(\mathbf{x}(t), \mathbf{x}(0))}{dt} = \mathbb{E} [\mathbf{f}(\mathbf{x}, t) | \mathbf{x}(0)]$$

Examples

$$\text{NCSN: } \mathbf{f}(\mathbf{x}, t) = 0 \quad \Rightarrow \quad \mu = \mathbf{x}(0)$$

$$\text{DDPM: } \mathbf{f}(\mathbf{x}, t) = -\frac{1}{2}\beta(t)\mathbf{x}(t) \quad \Rightarrow \quad \mu = \mathbf{x}(0) \exp\left(-\frac{1}{2} \int_0^t \beta(s) ds\right)$$

Score-based generative models through SDEs

Training

$$\mathbb{E}_{\pi(\mathbf{x}(0))} \mathbb{E}_{t \sim U[0,1]} \mathbb{E}_{q(\mathbf{x}(t)|\mathbf{x}(0))} \left\| \mathbf{s}_{\theta}(\mathbf{x}(t), t) - \nabla_{\mathbf{x}(t)} \log q(\mathbf{x}(t)|\mathbf{x}(0)) \right\|_2^2$$

$$q(\mathbf{x}(t)|\mathbf{x}(0)) = \mathcal{N}\left(\boldsymbol{\mu}(\mathbf{x}(t), \mathbf{x}(0)), \boldsymbol{\Sigma}(\mathbf{x}(t), \mathbf{x}(0))\right)$$

NCSN

$$q(\mathbf{x}(t)|\mathbf{x}(0)) = \mathcal{N}\left(\mathbf{x}(0), [\sigma^2(t) - \sigma^2(0)] \cdot \mathbf{I}\right)$$

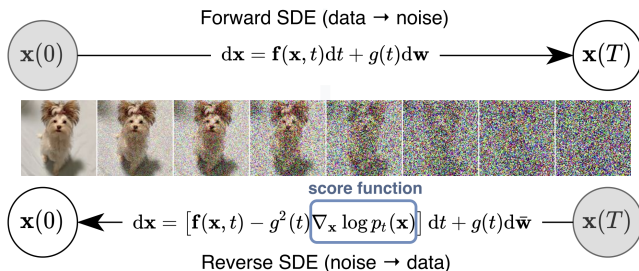
DDPM

$$q(\mathbf{x}(t)|\mathbf{x}(0)) = \mathcal{N}\left(\mathbf{x}(0)e^{-\frac{1}{2} \int_0^t \beta(s) ds}, \left(1 - e^{-\int_0^t \beta(s) ds}\right) \cdot \mathbf{I}\right)$$

Score-based generative models through SDEs

Sampling

Solve reverse SDE using numerical solvers (ODESolve).



- ▶ Discretization of the reverse SDE gives us the ancestral sampling.
- ▶ If we use probability flow instead of SDE than the reverse ODE gives us the DDIM sampling.

Summary

- ▶ SDE defines a stochastic process with drift and diffusion terms. ODEs are the special case of SDEs.
- ▶ KFP equation defines the dynamic of the probability function for the SDE.
- ▶ Langevin SDE has constant probability path.
- ▶ There exists special probability flow ODE for each SDE that gives the same probability path.
- ▶ It is possible to revert SDE using the score function.
- ▶ Score matching (NCSN) and diffusion models (DDPM) are the discretizations of the SDEs (variance exploding and variance preserving).
- ▶ It is possible to train the continuous-in-time score-based generative models through forward and reverse SDEs.