

Deep Generative Models

Lecture 11

Roman Isachenko

Moscow Institute of Physics and Technology
Yandex School of Data Analysis

2025, Autumn

Recap of Previous Lecture

DDPM Objective

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

In practice, **this coefficient** is usually omitted.

NCSN Objective

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

Note: The objectives of DDPM and NCSN are almost identical; however, their sampling procedures differ:

- ▶ NCSN utilizes annealed Langevin dynamics,
- ▶ DDPM employs ancestral sampling.

Recap of Previous Lecture

Unconditional Generation

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}} \cdot \mathbf{x}_t + \frac{\beta_t}{\sqrt{1-\beta_t}} \cdot \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\boldsymbol{\theta}) + \sigma_t \cdot \epsilon$$

Conditional Generation

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}} \cdot \mathbf{x}_t + \frac{\beta_t}{\sqrt{1-\beta_t}} \cdot \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{y}, \boldsymbol{\theta}) + \sigma_t \cdot \epsilon$$

Conditional Distribution

$$\begin{aligned} \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{y}, \boldsymbol{\theta}) &= \nabla_{\mathbf{x}_t} \log p(\mathbf{y}|\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\boldsymbol{\theta}) \\ &= \nabla_{\mathbf{x}_t} \log p(\mathbf{y}|\mathbf{x}_t) - \frac{\epsilon_{\theta,t}(\mathbf{x}_t)}{\sqrt{1-\bar{\alpha}_t}} \end{aligned}$$

Here, $p(\mathbf{y}|\mathbf{x}_t)$ denotes a classifier operating on noisy samples (which must be trained separately).

Recap of Previous Lecture

Classifier-Corrected Noise Prediction

$$\epsilon_{\theta,t}(\mathbf{x}_t, \mathbf{y}) = \epsilon_{\theta,t}(\mathbf{x}_t) - \sqrt{1 - \bar{\alpha}_t} \cdot \nabla_{\mathbf{x}_t} \log p(\mathbf{y}|\mathbf{x}_t)$$

Guidance Scale

$$\epsilon_{\theta,t}(\mathbf{x}_t, \mathbf{y}) = \epsilon_{\theta,t}(\mathbf{x}_t) - \gamma \cdot \sqrt{1 - \bar{\alpha}_t} \cdot \nabla_{\mathbf{x}_t} \log p(\mathbf{y}|\mathbf{x}_t)$$

- ▶ Train DDPM as usual.
- ▶ Train a separate classifier $p(\mathbf{y}|\mathbf{x}_t)$ on noisy samples \mathbf{x}_t .

Guided Sampling

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \cdot \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t(1 - \bar{\alpha}_t)}} \cdot \epsilon_{\theta,t}(\mathbf{x}_t, \mathbf{y}) + \sigma_t \cdot \epsilon$$

Note: The guidance scale γ sharpens the distribution $p(\mathbf{y}|\mathbf{x}_t)$.

Recap of Previous Lecture

The previous method requires an additional classifier $p(\mathbf{y}|\mathbf{x}_t)$ trained on noisy data. Let's try to avoid this requirement.

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{y}|\mathbf{x}_t) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{y}, \boldsymbol{\theta}) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\boldsymbol{\theta})$$

$$\begin{aligned} \nabla_{\mathbf{x}_t}^\gamma \log p(\mathbf{x}_t|\mathbf{y}, \boldsymbol{\theta}) &= \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\boldsymbol{\theta}) + \gamma \cdot \nabla_{\mathbf{x}_t} \log p(\mathbf{y}|\mathbf{x}_t) = \\ &= (1 - \gamma) \cdot \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\boldsymbol{\theta}) + \gamma \cdot \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{y}, \boldsymbol{\theta}) \end{aligned}$$

Classifier-Free-Corrected Noise Prediction

$$\hat{\epsilon}_{\boldsymbol{\theta}, t}(\mathbf{x}_t, \mathbf{y}) = \gamma \cdot \epsilon_{\boldsymbol{\theta}, t}(\mathbf{x}_t, \mathbf{y}) + (1 - \gamma) \cdot \epsilon_{\boldsymbol{\theta}, t}(\mathbf{x}_t)$$

- ▶ Train one model $\epsilon_{\boldsymbol{\theta}, t}(\mathbf{x}_t, \mathbf{y})$ on **supervised** data, alternating between true conditioning \mathbf{y} and empty conditioning $\mathbf{y} = \emptyset$.
- ▶ During inference, apply this model twice.

Recap of Previous Lecture

Continuous-Time Dynamics (Neural ODE)

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{f}_{\theta}(\mathbf{x}(t), t); \quad \text{where } \mathbf{x}(t_0) = \mathbf{x}_0.$$

$$\mathbf{x}(t_1) = \int_{t_0}^{t_1} \mathbf{f}_{\theta}(\mathbf{x}(t), t) dt + \mathbf{x}_0 \approx \text{ODESolve}_f(\mathbf{x}_0, \theta, t_0, t_1).$$

Here, $\mathbf{f}_{\theta} : \mathbb{R}^m \times [t_0, t_1] \rightarrow \mathbb{R}^m$ is a vector field.

Euler Update Step (ODESolve)

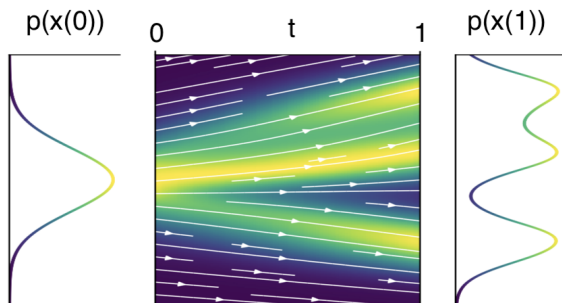
$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \Delta t \cdot \mathbf{f}_{\theta}(\mathbf{x}(t), t)$$

- ▶ The Euler method is the simplest version of ODESolve, but it's unstable in practice.
- ▶ More advanced numerical methods (such as Runge-Kutta) are often used instead.

Recap of Previous Lecture

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{f}_\theta(\mathbf{x}(t), t); \quad \mathbf{x}(t_0) = \mathbf{x}_0$$

- ▶ Suppose $\mathbf{x}(0)$ is a random variable with density $p_0(\mathbf{x})$. Then, $\mathbf{x}(t)$ is a random variable with density $p_t(\mathbf{x})$.
- ▶ $p_t(\mathbf{x}) = p(\mathbf{x}, t)$ describes the **probability path** between $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$.



Recap of Previous Lecture

Theorem (Picard)

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

This means the ODE can be **uniquely inverted**:

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

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

Note: Unlike discrete-time NF, \mathbf{f} need not be invertible (uniqueness ensures bijectivity).

How can we compute $p_t(\mathbf{x})$ for any t ?

Outline

1. Continuity Equation for NF Log-Likelihood
2. SDE Basics
3. Probability Flow ODE
4. Reverse SDE

Outline

1. Continuity Equation for NF Log-Likelihood
2. SDE Basics
3. Probability Flow ODE
4. Reverse SDE

Continuous-Time NF

Theorem (Continuity Equation)

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

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

This result states: given $\mathbf{x}_0 = \mathbf{x}(0)$, the solution to the continuity equation gives the density $p_1(\mathbf{x}(1))$.

Solution of the Continuity Equation

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

- ▶ This provides the density along the trajectory (not the total probability path).
- ▶ However, **the latter term** is difficult to estimate efficiently.

Outline

1. Continuity Equation for NF Log-Likelihood
2. SDE Basics
3. Probability Flow ODE
4. Reverse SDE

Stochastic Differential Equation (SDE)

Let's define a 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.
- ▶ $g(t) : \mathbb{R} \rightarrow \mathbb{R}$ is the **diffusion** function.
- ▶ $\mathbf{w}(t)$ is the standard Wiener process (Brownian motion), defined by:
 1. $\mathbf{w}(0) = 0$ (almost surely);
 2. $\mathbf{w}(t)$ has independent increments;
 3. $\mathbf{w}(t)$ trajectories are continuous;
 4. $\mathbf{w}(t) - \mathbf{w}(s) \sim \mathcal{N}(0, (t - s)\mathbf{I})$ 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 recover the standard ODE.

Stochastic Differential Equation (SDE)

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

- ▶ Unlike ODEs, the initial condition $\mathbf{x}(0)$ doesn't uniquely determine the trajectory.
- ▶ There are two sources of randomness: the initial distribution $p_0(\mathbf{x})$ and the Wiener process $\mathbf{w}(t)$.

Discretizing the SDE (Euler Method) – SDEsolve

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

If $dt = 1$, then

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

- ▶ At any time t , the process has density $p_t(\mathbf{x}) = p(\mathbf{x}, t)$.
- ▶ $p : \mathbb{R}^m \times [0, 1] \rightarrow \mathbb{R}_+$ specifies a **probability path** from $p_0(\mathbf{x})$ to $p_1(\mathbf{x})$.
- ▶ How can we obtain the probability 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)

The evolution of $p_t(\mathbf{x})$ is governed by

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

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)

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

- ▶ The KFP theorem uniquely determines $p_t(\mathbf{x})$.
- ▶ This generalizes the continuity equation for continuous-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)

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

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

Let's now apply the 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 $p_t(\mathbf{x})$ is constant in t ; i.e., 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. Continuity Equation for NF Log-Likelihood
2. SDE Basics
3. Probability Flow ODE
4. Reverse SDE

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 randomness is the initial distribution $p_0(\mathbf{x})$.

SDE and KFP Equation

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

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

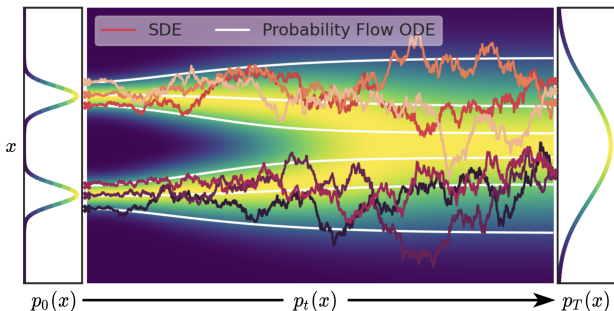
Now there are two sources of randomness: the initial distribution $p_0(\mathbf{x})$ and the Wiener process $\mathbf{w}(t)$.

Probability Flow ODE

Theorem

Suppose the 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 an ODE with the same probability path $p_t(\mathbf{x})$, given by

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



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

Probability Flow ODE

Theorem

Suppose the 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 an ODE with the same probability path $p_t(\mathbf{x})$, given by

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

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

Proof (Continued)

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

$$\tilde{\mathbf{f}}(\mathbf{x}, t) = \mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g^2(t) \frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x}); \quad \tilde{g}(t) = 0$$

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

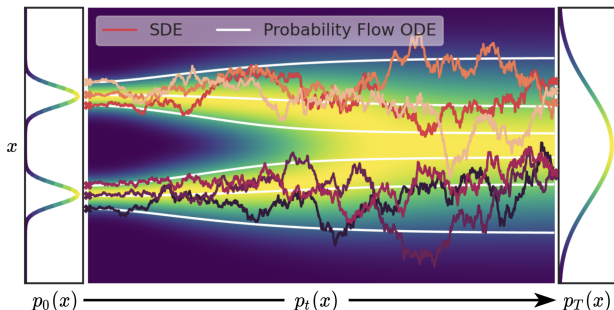
$$\frac{\partial p_t(\mathbf{x})}{\partial t} = -\text{div}(\tilde{\mathbf{f}}(\mathbf{x}, t) p_t(\mathbf{x})) + \frac{1}{2} \tilde{g}^2(t) \Delta_{\mathbf{x}} p_t(\mathbf{x})$$

Probability Flow ODE

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

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

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



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

Outline

1. Continuity Equation for NF Log-Likelihood
2. SDE Basics
3. Probability Flow ODE
4. Reverse SDE

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 can 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 do we reverse the SDE $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$?
- ▶ The Wiener process introduces randomness that must be reversed.

Theorem

There exists a reverse SDE for $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$, given by:

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

where $dt < 0$.

Reverse SDE

Theorem

There exists a reverse SDE for $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$, given by:

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

where $dt < 0$.

Note: Again, the score function appears: $\mathbf{s}(\mathbf{x}, t) = \frac{\partial}{\partial \mathbf{x}} \log p_t(\mathbf{x})$.

Proof Sketch

- ▶ Convert the initial SDE to a probability flow ODE.
- ▶ Reverse the probability flow ODE.
- ▶ Convert the reversed probability flow ODE back to an SDE.

Reverse SDE

Proof

- Convert the initial SDE to a probability flow ODE:

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

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

- Reverse the 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 the reversed probability flow ODE back to an 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 a reverse SDE for $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$, given by:

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

where $dt < 0$.

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 $dt < 0$.

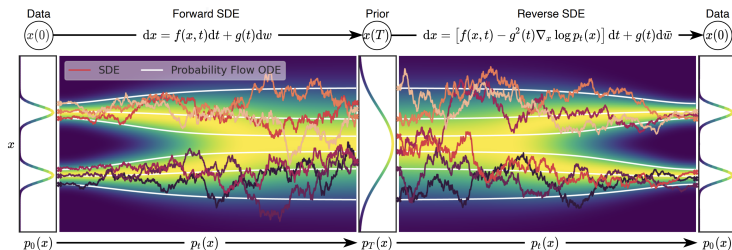
Reverse SDE

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

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

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

- ▶ This framework allows us to transform one distribution into another via an SDE with a prescribed probability path $p_t(\mathbf{x})$.
- ▶ We can invert this process using the score function.



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

Summary

- ▶ The continuity equation allows us to compute $\log p(\mathbf{x}, t)$ at any time t .
- ▶ An SDE defines a stochastic process with drift and diffusion terms; ODEs are a special case of SDEs.
- ▶ The KFP equation describes the probability dynamics of an SDE.
- ▶ The Langevin SDE preserves a constant probability path.
- ▶ Every SDE admits a corresponding probability flow ODE following the same probability path.
- ▶ SDEs can be reversed using the score function.