Manifold Learning and Artificial Intelligence
Lecture 6

Generative AI (4)

Stochastic Differential Equation-based Generative Models

Momiao Xiong, University of Texas School of Public Health

- Time: 9:00 pm, US East Time, 12/03/2022
- 10:00 am, Beijing Time. 12/04/2022
- Zoom
  https://uwmadison.zoom.us/j/93316139423?pwd=Q0NVWFYvRFg5RmxCNkwxMmYrbW41dz09
- Meeting ID: 933 1613 9423
- Passcode: 416262

Github Address: https://ai2healthcare.github.io/

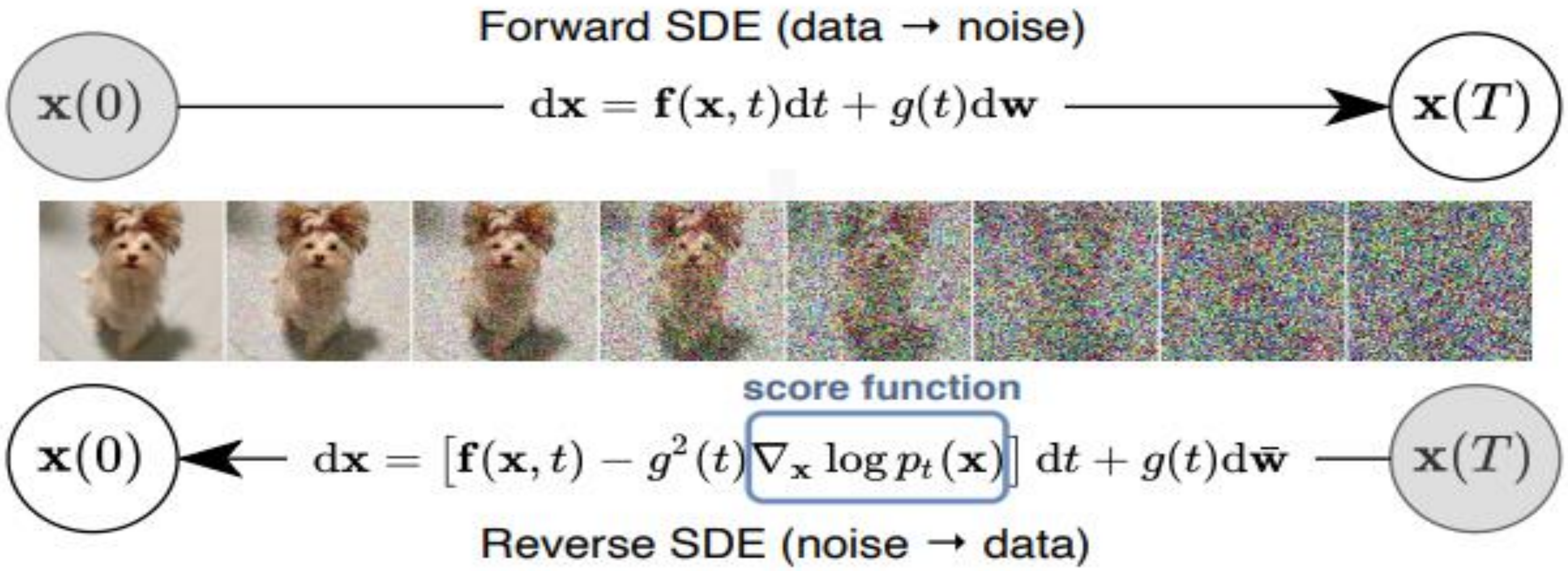
Song et al. 2021. SCORE-BASED GENERATIVE MODELING THROUGH STOCHASTIC DIFFERENTIAL EQUATIONS

Dabral, 2021, Stochastic Differential Equations and Diffusion Models

Karlin and Taylor (1981) A second Course in Stochastic Process

Bernt ØKsendal 2003, Stochastic Differential Equations.

1.6. Generative Model through Stochastic Differential Equation (SDE)



Song et al. 2021. SCORE-BASED GENERATIVE MODELING THROUGH STOCHASTIC DIFFERENTIAL EQUATIONS

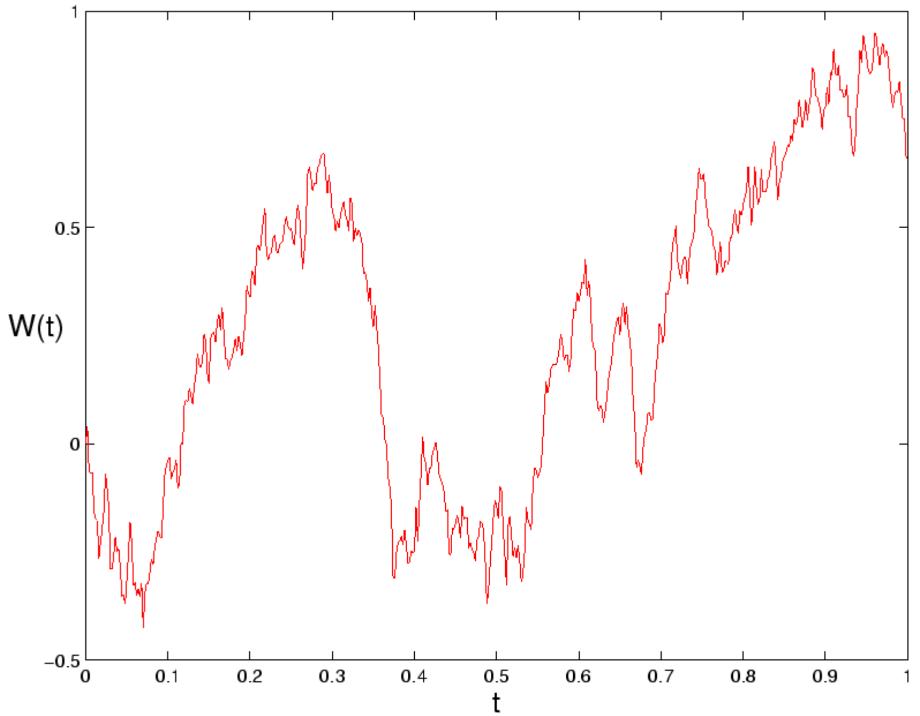
1.6.1. Basics of SDE

- **Browning Motion (Wiener Process)**

Browning motion is a regular diffusion process on the interval $(-\infty, +\infty)$ with $\mu(x) = 0, \sigma^2(x) = \sigma^2$, constant for all x

- $\mathrm{W}(0) = 0$
- $\mathrm{W}(t) - W(s) \sim N(0, t - s)$
- $Var(\Delta W(t)) = \Delta t$

$\Delta \mathrm{W}(\mathrm{t}) = \mathrm{W}(\mathrm{t} + \Delta t) - W(t)$

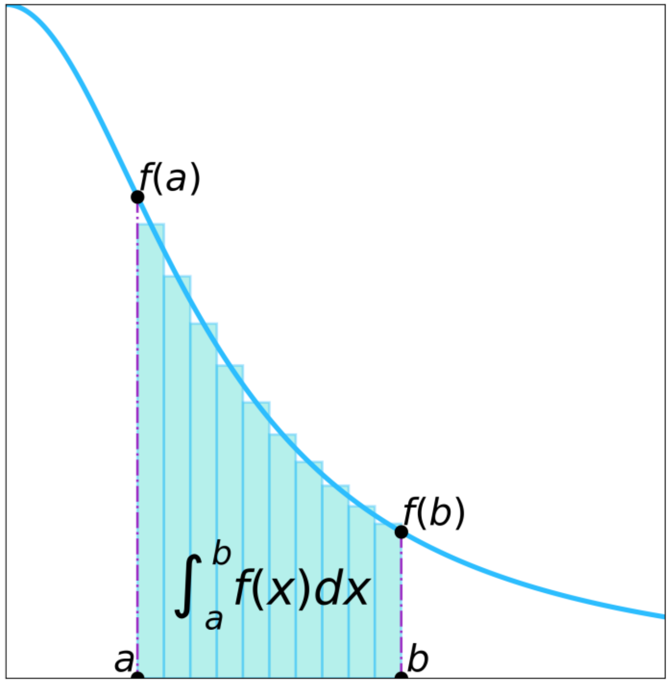


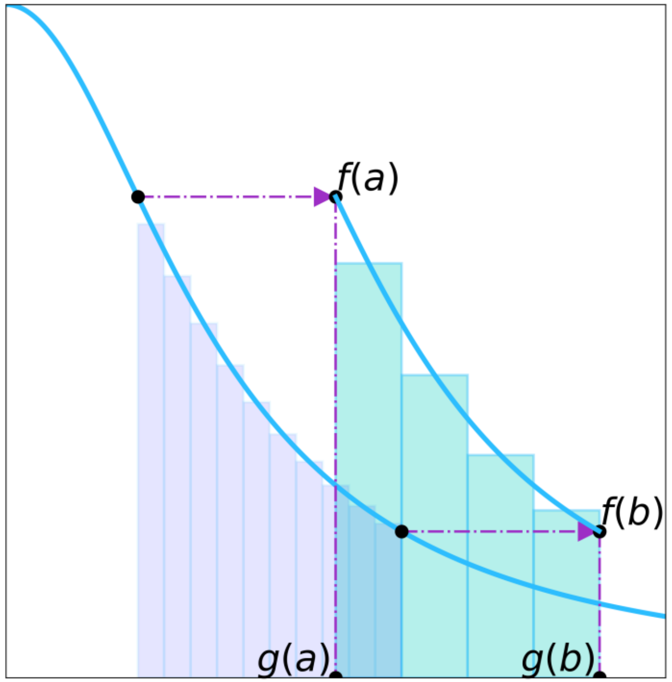
- **Three Types of Integrals**

- Riemann-Stieltjes Integral

$$\int_a^b X(t)dg(t) = \sum_{k=1}^{n} X(u_k)[g(t_k) - g(t_{k-1})]$$

$$a = t_0 < t_1 < \cdots < t_n = b$$
$$t_{k-1} \le u_k \le t_k$$





by Vladimir Ilievski

- **Orthogonal Increment**

$$E[X(t)] = 0$$

$$a \le t_1 < t_2 \le t_3 < t_4 \le b$$

$$E\left[(X(t_2) - X(t_1))\overline{(X(t_4) - X(t_3))}\right] = 0$$

$$X(a) = 0$$

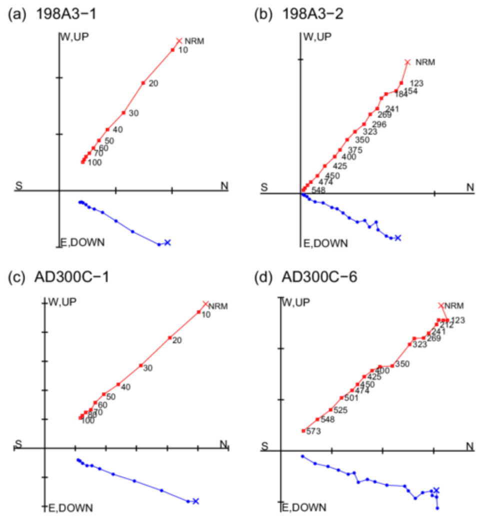
$$F(s) = E\left[X(s)\overline{X(s)}\right] = E\left[|X(s)|^2\right]$$

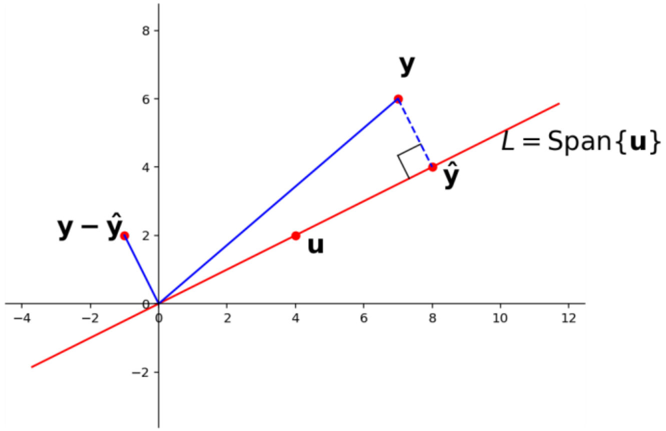
$$\Gamma(s,t) = E\left[X(s)\overline{X(t)}\right] \qquad s < t$$

$$= E\left[X(s)\overline{(X(s) + X(t) - X(s))}\right]$$

$$= E\left[X(s)\overline{X(s)}\right] = F(s) = F(\min(s,t))\Big]$$

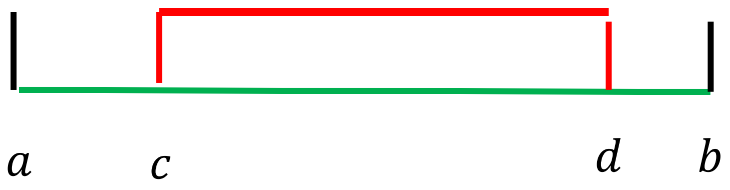
Inner Product





- **Integral with Orthogonal Increment**

(1) $a \le c < d \le b,\ \mathrm{f(t)} = \mathcal{X}_{[c,d)}(\mathrm{t})$



$$\int_a^b \mathcal{X}_{[c,d)}(\mathrm{t})dX(t) = X(d) - X(c)$$

(2) $f(t) = \sum_{i=1}^{n} k_i \mathcal{X}_{[c_i,d_i)}$

$$\int_a^b f(t)dX(t) = \sum_{i=1}^{n} k_i[X(d_i) - X(c_i)]$$

(3)

$$E\left[\int_a^b f_1(t)dX(t)\overline{\int_a^b f_2(t)dX(t)}\right] = \int_a^b f_1(t)\overline{f_2(t)}\,dF(t)$$

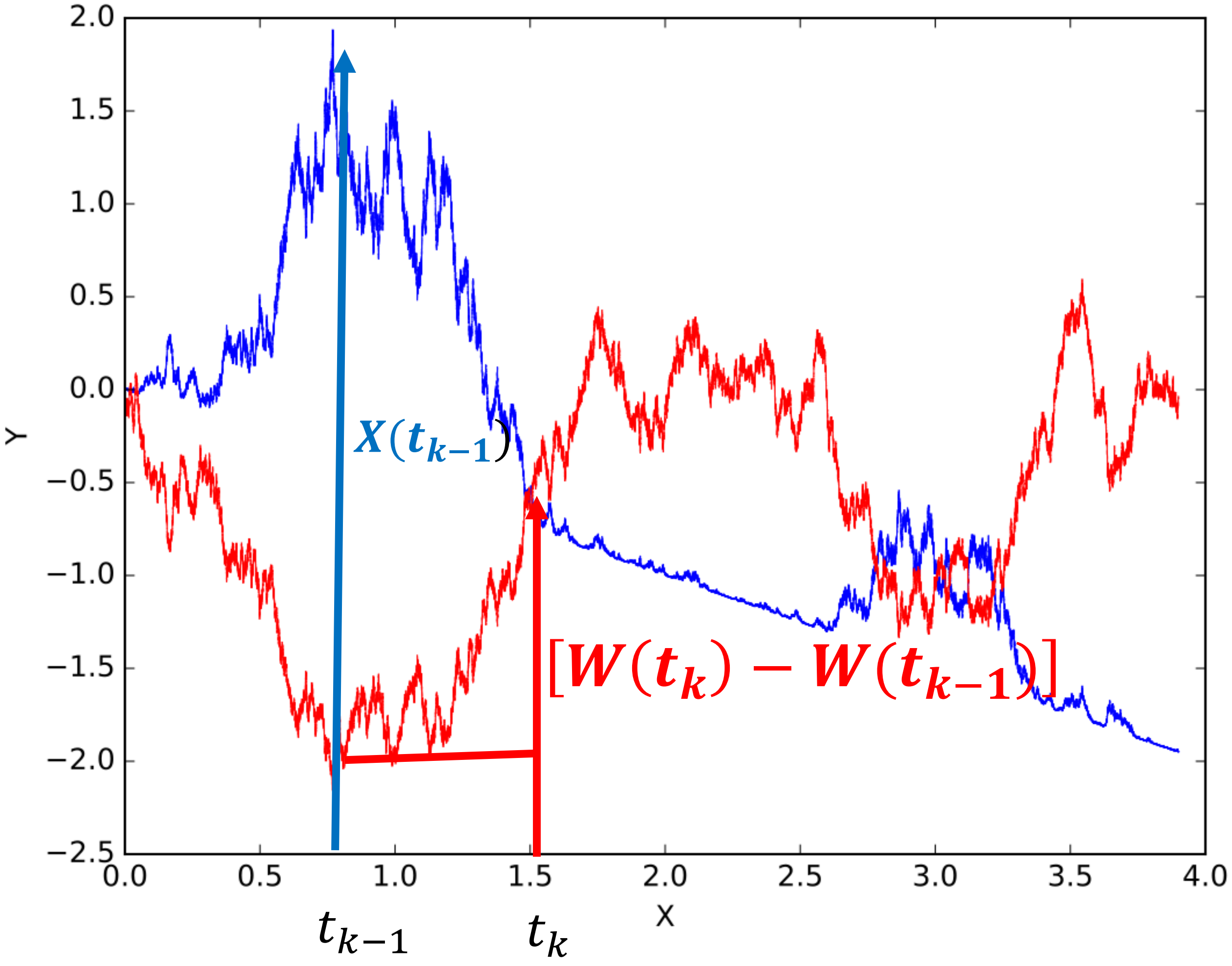
- ***Itô* Integral**

$$a = t_0 < t_1 < \cdots < t_n = b$$

$$\Delta_n = \max_{1 \le k \le n} (t_k - t_{k-1})$$

$$I_n = \sum_{k=1}^{n} X(t_{k-1})[W(t_k) - W(t_{k-1})]$$

$$\int_a^b X(t)dW(t) = \lim_{\Delta_n \to 0} I_n$$



$$\int_a^b W(t)dW(t) = \frac{1}{2}[W^2(b) - W^2(a)] - \frac{1}{2}(b - a)$$

1.6.2. Concept of SDE

- **Drift and Diffusion Coefficient**

$$X(t+\Delta t)-X(t)\approx \mu(x,t)\Delta t+\sigma(x,t)\Delta W(t) \qquad \textbf{(1)}$$

$$\lim_{\Delta t\downarrow 0}\frac{E[\Delta X]}{\Delta t}=\lim_{\Delta t\downarrow 0}\frac{1}{\Delta t}E[\mu(x,t)\Delta t+\sigma(x,t)\Delta W(t)]=\mu(x,t) \qquad \textbf{(2)}$$

$$\lim_{\Delta t\downarrow 0}\frac{1}{\Delta t}E\left[(\Delta X)^2\right]=\lim_{\Delta t\downarrow 0}\frac{1}{\Delta t}\left\{Var(\Delta X)+(E[\Delta X])^2\right\}$$

$$=\lim_{\Delta t\downarrow 0}\frac{1}{\Delta t}\{Var(\sigma(x,t)\Delta W(t))\}+\lim_{\Delta t\downarrow 0}\frac{1}{\Delta t}(\mu(x,t)\Delta t)^2$$

$$=\lim_{\Delta t\downarrow 0}\frac{1}{\Delta t}\sigma^2(x,t)\Delta t=\sigma^2(x,t) \qquad \textbf{(3)}$$

- **Summation of Diffusion**

Let $0 = t_0 < t_1 < \cdots < t_n = t$

$$X_{k+1} - X_k = \mu(t_k, X_k)\Delta t_k + \sigma(t_k, X_k)\Delta W_k \qquad\qquad X_k = X(t_k)$$

Therefore, summarizing increment in the whole region, we obtain

$$X_k = X_0 + \sum_{k=1}^{n} \mu(t_{k-1}, X_{k-1})\Delta t_k + \sum_{k=1}^{n} \sigma(t_{k-1}, X_{k-1})\Delta W_k$$

$$= X_0 + \int_0^t \mu(s, X_s)ds + \int_0^t \sigma(s, X_s)dW_s \qquad \textbf{(4)}$$

- **One dimensional SDE**

$$dX(t) = \mu(X(t), t)dt + \sigma(X(t), t)dW(t) \tag{5}$$

Drift **Diffusion coefficient**

- **M-dimensional SDE**

$$\boldsymbol{X}(\boldsymbol{t}) \in R^m, \boldsymbol{\mu}(\boldsymbol{X}(\boldsymbol{t}), \boldsymbol{t}) \in R^m, \boldsymbol{\sigma}(\boldsymbol{X}(\boldsymbol{t}), \boldsymbol{t}) \in \boldsymbol{R}^{\boldsymbol{m}\times\boldsymbol{m}}, \boldsymbol{W}(\boldsymbol{t}) \in \boldsymbol{R}^{\boldsymbol{m}}$$

$$\boldsymbol{dX}(\boldsymbol{t}) = \boldsymbol{\mu}(\boldsymbol{X}(\boldsymbol{t}), \boldsymbol{t})\,\mathrm{dt} + \boldsymbol{\sigma}(\boldsymbol{X}(\boldsymbol{t}), \boldsymbol{t})\mathrm{d}\,\boldsymbol{W}(\boldsymbol{t})$$

Example

$$\begin{bmatrix} dx_1(t) \\ dx_2(t) \end{bmatrix} = \begin{bmatrix} \mu_1(\boldsymbol{X}(t), t) \\ \mu_2(\boldsymbol{X}(t), t) \end{bmatrix} \mathrm{dt} + \begin{bmatrix} \sigma_{11}(\boldsymbol{X}(t), t) & \sigma_{12}(\boldsymbol{X}(t), t) \\ \sigma_{21}(\boldsymbol{X}(t), t) & \sigma_{22}(\boldsymbol{X}(t), t) \end{bmatrix} \begin{bmatrix} dW_1(t) \\ dW_{2(t)} \end{bmatrix}$$

1.6.3. Transformation Law for the Ito Stochastic Differential

- **Taylor Expansion**

Define $Y(t) = f(X(t), t)$.

$$dY(t) = df(X(t), t) = f_x(X(t), t)dX(t) + f_t(X(t), t)dt$$

$$+\frac{1}{2} f_{xx}(X(t), t)[dX(t)]^2 + f_{x,t}(X(t), t)dXdt + \frac{1}{2} f_{tt}(X(t), t)(dt)^2 \qquad \textbf{(6)}$$

Substituting equation (5) into equation (6) and applying $[dW(t)]^2 \approx dt$ yields

$$dY(t) = \left[f_x(X(t), t)\mu(X(t), t) + f_t(X(t), t) + \frac{1}{2} f_{xx}(X(t), t)\sigma^2(X(t), t) \right] \mathrm{dt}$$

$$+f_x(X(t), t)\sigma(X(t), t)dW(t) \qquad \textbf{(7)}$$

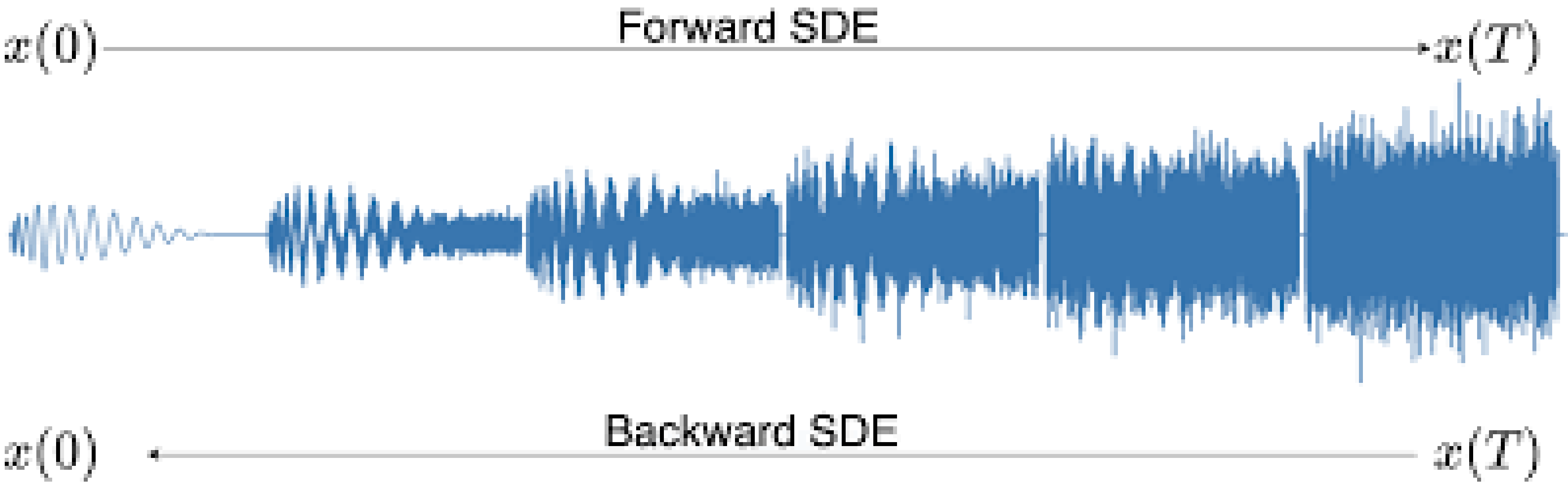
1.6.4. Diffusion Models as SDEs

- **Forward SDE**

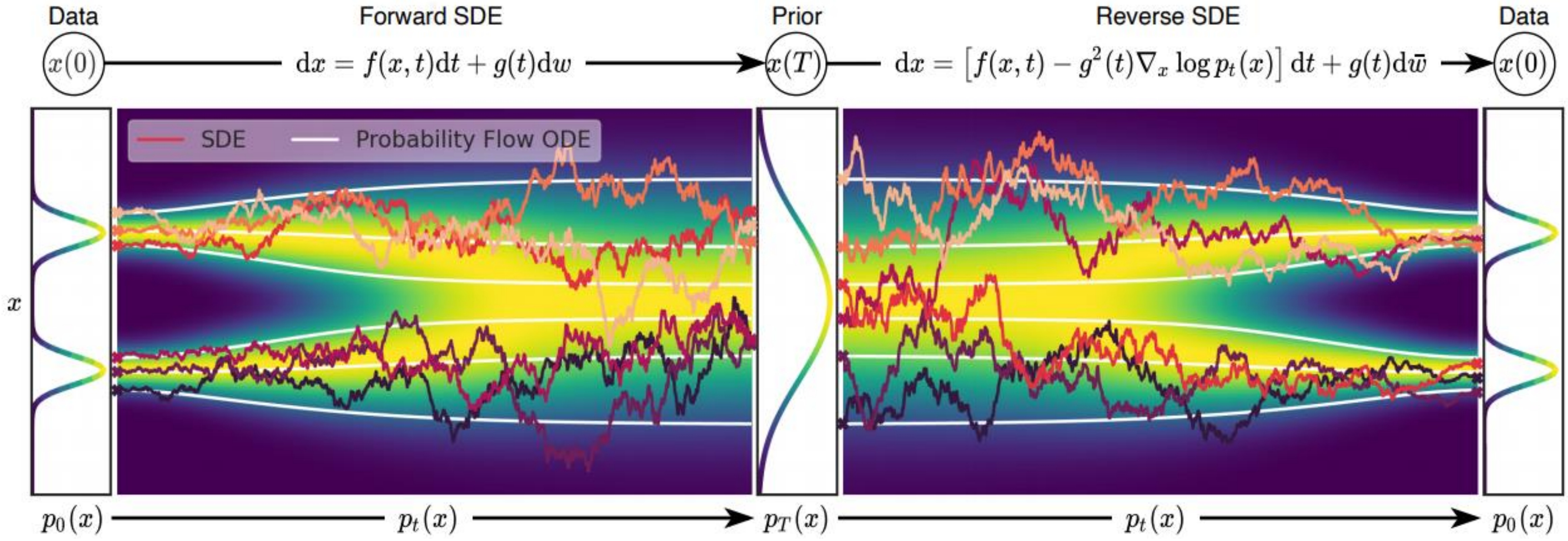
$$dX(t) = \mu(X,t)dt + \sigma(t)dW \qquad (8)$$

- **Reverse-Time SDE**

$$dX(t) = \left[\mu(X,t) - \sigma^2(t)\nabla_x \log P(X,t)\right]dt + \sigma(t)dW \qquad (9)$$



Forward SDE
$x(0)$
$x(T)$
Backward SDE
$x(0)$
$x(T)$



Data
Forward SDE
Prior
Reverse SDE
Data
$x(0)$
$\mathrm{d}x = f(x,t)\mathrm{d}t + g(t)\mathrm{d}w$
$x(T)$
$\mathrm{d}x = \left[f(x,t) - g^2(t)\nabla_x \log p_t(x)\right]\mathrm{d}t + g(t)\mathrm{d}\bar{w}$
$x(0)$
SDE
Probability Flow ODE
x
$p_0(x)$
$p_t(x)$
$p_T(x)$
$p_t(x)$
$p_0(x)$

- **Learn Score**

All we need to learn to formulate the reverse process is the term $\boldsymbol{\nabla_x \log P(X_t; t)}$

We want to learn a neural network –based parametrized function $S(X;\theta)$ that predict $\boldsymbol{\nabla_x \log P(X_t; t)}$. To achieve this goal, we minimize

$$\mathcal{L}_t(\theta) = E_{P(X_t,t)}\left[\left\|S(X_t,t;\theta) - \boldsymbol{\nabla_x \log P(X_t;t)}\right\|^2\right]$$

$$= E_{P(X_t,t)}\left[\|S(X_t,t;\theta)\|^2 - 2S(X_t,t;\theta)^T \boldsymbol{\nabla_x \log P(X_t;t)} + \left\|\boldsymbol{\nabla_x \log P(X_t;t)}\right\|^2\right]$$

$$\approx E_{P(X_t,t)}\left[\|S(X_t,t;\theta)\|^2 - 2E_{P(X_t,t)}\left[S(X_t,t;\theta)^T \boldsymbol{\nabla_x \log P(X_t;t)}\right]\right] \qquad \textbf{(10)}$$

The third term does not involve θ and hence can be ignored.

Next we calculate $E_{P(X_t,t)}\left[S(X_t,t;\theta)^T \boldsymbol{\nabla_x \log P(X_t;t)}\right]$

$$E_{P(X_t,t)}\left[S(X_t,t;\theta)^T \boldsymbol{\nabla_x \log P(X_t;t)}\right] = \int P(X_t,t)\, S(X_t,t;\theta)^T \boldsymbol{\nabla_x \log P(X_t;t) dX_t}$$

$$= \int\int P(X_0,0)P(X_t,t|X_0)\ \ S(X_t,t;\theta)^T\ \boldsymbol{\nabla_x \log P(X_0,0)P(X_t,t|X_0) dX_t dX_0}$$

$$\boldsymbol{= \nabla_x \log P(X_0,0) + \nabla_x \log P(X_t,t|X_0)}$$

$$\boldsymbol{= 0}$$

$$= \int P(X_0,0)P(X_t,t|X_0)\ \ S(X_t,t;\theta)^T \nabla_x \log \boldsymbol{P(X_t,t|X_0) dX_t dX_0}$$

$$= E_{P(x_0;0)} E_{P(X_t;t|X_0,0)}\left[S(X_t,t;\theta)^T \nabla_x \log \boldsymbol{P(X_t,t|X_0)}\ \right] \qquad \textbf{(!1)}$$

Dabral, 2021, Stochastic Differential Equations and Diffusion Models

Substituting equation (11) into equation (10), we obtain

$$\mathcal{L}_t(\theta) = E_{P(x_0;0)} E_{P(X_t;t|X_0,0)} \left[\|S(X_t,t;\theta)\|^2 - 2S(X_t,t;\theta)^T \nabla_x \log \boldsymbol{P}(\boldsymbol{X_t}, \boldsymbol{t}|\boldsymbol{X_0}) \right] \qquad \textbf{(12)}$$

Since $\nabla_x \log \boldsymbol{P}(\boldsymbol{X_t}, \boldsymbol{t}|\boldsymbol{X_0})$ does not involve θ, adding $E_{P(x_0;0)} E_{P(X_t;t|X_0,0)} \left[\left\| \nabla_x \log \boldsymbol{P}(\boldsymbol{X_t}, \boldsymbol{t}|\boldsymbol{X_0}) \right\|^2 \right]$ will not affect estimation of parameters in $\mathcal{L}_t(\theta)$

Therefore, adding $E_{P(x_0;0)} E_{P(X_t;t|X_0,0)} \left[\left\| \nabla_x \log \boldsymbol{P}(\boldsymbol{X_t}, \boldsymbol{t}|\boldsymbol{X_0}) \right\|^2 \right]$ in equation (12) yields

$$\mathcal{L}_t(\theta) = E_{P(x_0;0)} E_{P(X_t;t|X_0,0)} \left[\left\| S(X_t,t;\theta) - \nabla_x \log \boldsymbol{P}(\boldsymbol{X_t}, \boldsymbol{t}|\boldsymbol{X_0}) \right\|^2 \right] \qquad \textbf{(13)}$$

Dabral, 2021, Stochastic Differential Equations and Diffusion Models

Using arguments in Dabral (2021), we can rewrite above integral as an expectation over a uniform distribution, and also add a positive weighting function $\lambda(t)$ if we want to focus on certain time instants more than the others. This finally gets us to the loss function mentioned in (Song et al., 2020):

$$\mathcal{L}(\theta) = E_{t\sim U[0,T]} E_{P(X_0;0)} \left[\lambda(t) \left\| S(X_t, t; \theta) - \nabla_x \log \boldsymbol{P(X_t, t|X_0)} \right\|^2 \right] \qquad \textbf{(14)}$$

- ## Methods for Calculating Score

Denoising DDPM:

$$\log \boldsymbol{P(X_t, t|X_0)} = \boldsymbol{N(x_t; \sqrt{\overline{\alpha}_t} x_0,\ (1 - \overline{\alpha}_t) I)}$$

Sliced Score Matching

$$\theta^* = \underset{\theta}{\operatorname{argmin}} E_t \left[\lambda(t) E_{(X_0,0)} E_{X(t)} E_{V\sim P_V} \left[\frac{1}{2}\right] \|S(X_t, t; \theta)\|_2^2 + V^T S(X_t, t; \theta) V \right], P_V \sim N(0, I)$$

Dabral, 2021, Stochastic Differential Equations and Diffusion Models

- **Example 1**

DENOISING SCORE MATCHING WITH LANGEVIN DYNAMICS (SMLD)

Recursive formula for X_t

$$X_t = X_{t-1} + \sqrt{\sigma_t^2 - \sigma_{t-1}^2} Z_{t-1}, Z_{t-1} \sim N(0, I)$$

$$\Delta X_t = \sqrt{\frac{d[\sigma_t^2]}{dt}} Z_{t-1} \Delta t$$

Thus,

$$\boldsymbol{dx = \sqrt{\frac{d[\sigma_t^2]}{dt}} dW} \qquad \textbf{(15)}$$

- **Example 2 (DDPM)**

$$X_t = \sqrt{1-\beta_t} X_{t-1} + \sqrt{\beta_t} Z_{t-1}$$

$$\approx \left(1 - \tfrac{1}{2}\beta_t\right) X_{t-1} + \sqrt{\beta_t}\, Z_{t-1}$$

Taylor Expansion

$$\sqrt{1-\beta_t} \approx \left(1 - \frac{1}{2}\beta_t\right)$$

$$\Delta X_t = -\frac{1}{2}\beta_t X_{t-1}\Delta t + \sqrt{\beta_t}\, Z_{t-1}\Delta t$$

which implies the following forward SDE

$$\boldsymbol{dX = -\frac{1}{2}\beta_t X_{t-1} dt + \sqrt{\beta_t} dW}$$

Reverse-Time SDE

$$\boldsymbol{dX = \left(-\tfrac{1}{2}\beta_t X_{t-1} - \beta_t \nabla_x \log P(X,t)\right)\mathrm{dt}+ \sqrt{\beta_t} dW}$$

1.6.5. More General SDE

- **Forward SDE**

$$dX(t) = \mu(X,t)dt + \sigma(X,t)dW$$

- **Reverse-Time SDE**

$$dX(t) = \left(\mu(X,t) - \frac{\boldsymbol{\partial\sigma^2(x,t)}}{\boldsymbol{\partial x}} - \sigma^2(x,t)\nabla_x \log P(X,t)\right)\mathrm{dt} + \sigma(X,t)dW$$

- **Forward SDE**

$$d\boldsymbol{X}(t) = \boldsymbol{\mu}(\boldsymbol{X},\boldsymbol{t})dt + \sigma(t)\boldsymbol{dW} \qquad \boldsymbol{X}(t), \boldsymbol{\mu}(\boldsymbol{X},\boldsymbol{t}), \boldsymbol{W}(t) \in R^d$$

- **Reverse-Time SDE**

$$\boldsymbol{dX(t)} = \left(\boldsymbol{\mu}(\boldsymbol{X},\boldsymbol{t}) - \sigma^2(t)\nabla_x \log P(X,t)\right)\mathrm{dt}+ \sigma(t)\boldsymbol{dW}$$

- **Forward SDE**

$$dX(t) = \boldsymbol{\mu}(\boldsymbol{X}, \boldsymbol{t})dt + \sigma(x, t)\boldsymbol{dW}$$

- **Reverse-Time SDE**

$$\boldsymbol{dX}(\boldsymbol{t}) = \left(\boldsymbol{\mu}(\boldsymbol{X}, \boldsymbol{t}) - \nabla_x \sigma^2(x, t) - \sigma^2(x, t)\nabla_x \log P(X, t)\right)\mathrm{dt} + \sigma(X, t)\boldsymbol{dW}$$