

SDE, reversed-SDE를 활용

↳ perturbed data의 t에서 gradient(score)에만 의존적

Sample 생성 시, NN으로 score 추정 후, numerical SDE solver로 sample 생성.

Discretized reverse time SDE의 error 수치를 위한 Predictor-Corrector까지.

그리고 neural ODE로 정확한 likelihood 추정과 sampling 효율↑.

1. Introduction

Probabilistic generative model은 noise를 추가하고 복원하는 방법이 포함됨.

SMLD는 \rightarrow noise scale의 score를 추정하고, noise를 제거하는데

score로부터 likelihood가 높은 data로 sampling하는 LD(Langevin Dynamics)를 반복한다.

DDPM은 reverse의 functional form을 사용하여 reverse를 위한 model 훈련.

Continuous의 DDPM은 각 noise level의 대형 score를 암시적으로 추정.

→ SMLD, DDPM 모두 score-based model

→ SDE는 위 두 방법을 포함하는 framework.

Finite한 time step이 아닌 continuum하기 때문에

→ Data를 점진적으로 noise로 diffuse하고

data의 의존하지 않고 trainable param을 SDE의 의해 주어짐

Reverse에서 복드럽기 data를 생성 가능

본장으로, reverse process는 forward SDE와 시간의 합으로 marginal probability density의

score가 주어지면 forward SDE로부터 유도되는 reverse SDE를 만족함

따라서, NN으로 t의 대형 score를 추정하고, numerical SDE solver로 sample을 생성하여

sample을 생성하여 reverse SDE 근사.

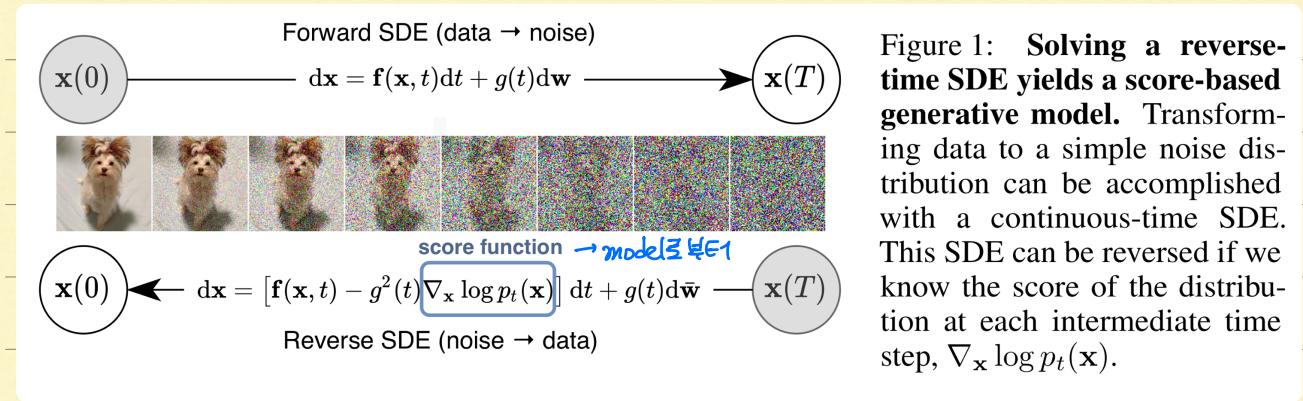


Figure 1: Solving a reverse-time SDE yields a score-based generative model. Transforming data to a simple noise distribution can be accomplished with a continuous-time SDE. This SDE can be reversed if we know the score of the distribution at each intermediate time step, $\nabla_x \log p_t(x)$.

Contribution.

1. flexible sampling & likelihood computation.

general purpose SDE solver를 적용할 수 있고, 고속화를 가능

i) Predictor - Corrector (PC)

numerical SDE solver와 score-based MCMC를 결합 (ex) Langevin MCMC, HMC

\rightarrow score-based sampling 방식 통합

ii) Probability flow ODE와 deterministic samplers.

\rightarrow black-box ODE solver로 빠른 sampling, 유연한 latent 조건, likelihood 계산 등을 가능하게 함

2. Controllable Generation.

unconditional score로 부터 conditional reverse-SDE는 효율적으로 계산될 수 있기 때문에

현재 외의 condition 정보로 부터 generation process를 조정한다.

\rightarrow retrain 때 conditional generation 가능.

3. Unified framework

SMLD와 DDPG는 SDE의 다른 discretized로 통합될 수 있다.

2. Background

-SMLD

$$P_\sigma(\tilde{x}|x) = \mathcal{N}(\tilde{x}; x, \sigma^2 I) \text{ 일 때, } P_\sigma(\tilde{x}) = \int P_{\text{data}}(x) P_\sigma(\tilde{x}|x) dx$$

$\sigma_{\min} \sim \sigma_{\max}$ 인 NCSN 모델.

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^N \sigma_i^2 \mathbb{E}_{p_{\text{data}}(x)} \mathbb{E}_{p_{\sigma_i}(\tilde{x}|x)} [\|\mathbf{s}_\theta(\tilde{x}, \sigma_i) - \nabla_{\tilde{x}} \log p_{\sigma_i}(\tilde{x} | x)\|_2^2]. \quad (1)$$

Sampling의 경우 각 $P_{\sigma_i}(x)$ 에서 M step 만큼 Langevin MCMC 수행

$$\mathbf{x}_i^m = \mathbf{x}_i^{m-1} + \epsilon_i \mathbf{s}_\theta^*(\mathbf{x}_i^{m-1}, \sigma_i) + \sqrt{2\epsilon_i} \mathbf{z}_i^m, \quad m = 1, 2, \dots, M, \quad (2)$$

\downarrow step size

$$\mathbf{x}_N^0 \sim \mathcal{N}(\mathbf{x}|0, \sigma_{\max}^2 I), \quad \mathbf{x}_i^0 = \mathbf{x}_{i+1}^M$$

설명 $M \rightarrow \infty, \epsilon_i \rightarrow 0$ 일 때, $\mathbf{x}_i^M \sim P_{\text{data}}(x)$

- DDPM.

$0 < \beta_i < 1$ 인 디스크리트 Markov chain $\mathbf{x}_0 \sim P_{\text{data}}(x)$ 을 수행

$$P(x_i | x_{i-1}) = \mathcal{N}(x_i; \sqrt{1-\beta_i} x_{i-1}, \beta_i I)$$

$$\rightarrow P(x_i | x_0) = \mathcal{N}(x_i; \sqrt{\alpha_i} x_0, (1-\alpha_i) I), \text{ where } \alpha_i = \prod_{j=1}^i (1-\beta_j)$$

Reverse의 경우,

$$P_\theta(x_{i-1} | x_i) = \mathcal{N}(x_{i-1}; \frac{1}{\sqrt{1-\beta_i}} (x_i + \beta_i \mathbf{s}_\theta(x_i, i)), \beta_i I)$$

ELBO:

$\hookrightarrow = \mathcal{L}_{\text{simple}}$

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^N (1 - \alpha_i) \mathbb{E}_{p_{\text{data}}(x)} \mathbb{E}_{p_{\alpha_i}(\tilde{x}|x)} [\|\mathbf{s}_\theta(\tilde{x}, i) - \nabla_{\tilde{x}} \log p_{\alpha_i}(\tilde{x} | x)\|_2^2]. \quad (3)$$

Reverse Markov:

\hookrightarrow ancestral sampling

$$\mathbf{x}_{i-1} = \frac{1}{\sqrt{1-\beta_i}} (\mathbf{x}_i + \beta_i \mathbf{s}_\theta^*(\mathbf{x}_i, i)) + \sqrt{\beta_i} \mathbf{z}_i, \quad i = N, N-1, \dots, 1. \quad (4)$$

(1) 과 (3)의 $\sigma_i^2, (1-\alpha_i) \propto 1/E[\|\nabla_x \log p_{\theta_i}(\tilde{x}|x)\|_2^2]$ 로 실질적으로 같은 form

3. Score-based generative modeling with SDEs.

이전 농법은 여러 noise scale을 data를 perturbing 하는 것이 핵심이다.

본 논문은 잡음이 써짐에 따라, perturbed data distribution의 SDE를 따라가도록 infinite number of noise scale로 일반화.

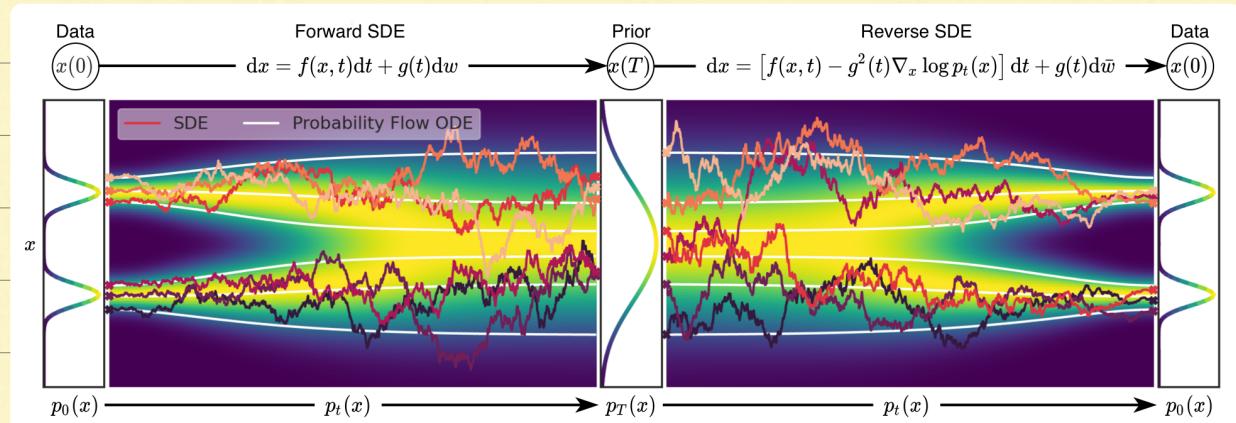


Figure 2: Overview of score-based generative modeling through SDEs. We can map data to a noise distribution (the prior) with an SDE (Section 3.1), and reverse this SDE for generative modeling (Section 3.2). We can also reverse the associated probability flow ODE (Section 4.3), which yields a deterministic process that samples from the same distribution as the SDE. Both the reverse-time SDE and probability flow ODE can be obtained by estimating the score $\nabla_x \log p_t(x)$ (Section 3.3).

- Perturbing data with SDEs.

Continuous time variable $t \in [0, T]$ 에 대해 $x(0) \sim P_0, x(T) \sim P_T$ 의 tractable sample를 효율적으로 생성하는 것이 목표다.

즉, P_0 는 data, P_T 는 prior.

Diffusion은 Itô SDE의 대한 solution으로 modeling 할 수 있다.

$$dx = f(x, t)dt + g(t)d\mathbf{w}, \quad \text{brownian motion.}$$

L_{drift}

$L_{\text{diffusion}}$

SDE는 drift가 Lipschitz 하다면 unique. strong solution.

$P_t(x)$ 는 $x(t)$ 의 probability density, $P_{st}(x(t)|x(s))$ 는 $0 \leq s < t \leq T$ 일 때,

$x(s)$ 와 $x(t)$ 를 transition kernel로 정의 한다.

일반적으로 P_t 는 P_0 의 정보를 담고 있지 않은 unstructured prior distribution.

고정된 P_t 의 분포로 SDE를 diffusion하는 방식을 사용. ex) SMID, DDPM.

- Generating Samples by reversing the SDE

reverse process도 또한 SDE이며, reverse-SDE로 부른다.

$$dx = [f(x, t) - g(t)^2 \nabla_x \log p_t(x)]dt + g(t)d\bar{w}, \quad (6)$$

\Rightarrow sample of marginal distribution (mean, std) of score ($\nabla_x \log P_t(x)$) +

계산되면, 우리는 P_0 로 simulate한 (6)으로부터 reverse diffusion process를 도출할 수 있다.

- Estimating scores for the SDE.

Distribution의 score는 SDE로 model을 훈련시켜 얻을 수 있다.

$\nabla_x \log P_t(x)$ 를 추정하기 위해 식 (1), (3)을 continuous로 일반화하여 $J_\theta(x, t)$ 훈련

$$\theta^* = \arg \min_{\theta} \mathbb{E}_t \left\{ \lambda(t) \mathbb{E}_{x(0)} \mathbb{E}_{x(t)|x(0)} \left[\|s_{\theta}(x(t), t) - \nabla_{x(t)} \log p_{0t}(x(t) | x(0))\|_2^2 \right] \right\}. \quad (7) \quad , t=[0, T]$$

weight func

* DSM은 perturbed score를 끌어는 것은 이용. $\rightarrow P_{0t}(x(t) | x(0))$

일반적으로 drift가 affine 일 때, transition kernel은 gaussian 이거나 closed-form이다.

affine or not closed-form인 경우 probabilistic flow ODE를 쓰거나

sliced score matching은 MIM, $\nabla_{x(t)} \log P_{0t}(x(t) | x(0))$ 의 계산을 피한다.

Appendix.A

- Examples : VE, VP SDE and beyond.

SMLD, DDPM에서 사용하는 noise perturbation은 각각 다른 SDE의 discretized
↳ VE(Variance Exploding) VP(Variance Preserving) \Rightarrow 자세한 Appendix.B.

Noise scale이 N 이상인 경우,

SMLD에서 x_i 에 대한 perturbation kernel $P_{t_i}(x|x_0)$ 은 다음 Markov Chain을 따른다.

$$x_i = x_{i-1} + \sqrt{\sigma_i^2 - \sigma_{i-1}^2} z_{i-1}, \quad i = 1, \dots, N, \quad (8)$$

Continuous form으로 보면, $t \in [0, 1]$ 이고 $\{X(t)\}_{t=0}^1$ 는 다음을 따른다.

$$dx = \sqrt{\frac{d[\sigma^2(t)]}{dt}} dw. \quad (9)$$

DDPM의 경우

$$x_i = \sqrt{1 - \beta_i} x_{i-1} + \sqrt{\beta_i} z_{i-1}, \quad i = 1, \dots, N. \quad (10)$$

는 continuum에서는 다음으로 누적한다.

$$dx = -\frac{1}{2} \beta(t) x dt + \sqrt{\beta(t)} dw. \quad (11)$$

식(9)은 $t \rightarrow \infty$ 일 때 variance exploding (VE) 일 때

식(11)은 초기 불변의 unit variance가 있을 때, fixed variance (VP)를 갖는다.

VP의 영향을 받아, 논문에서는 likelihood의 좋은 성능을 보이는 sub-VP SDE 제시.

$$dx = -\frac{1}{2} \beta(t) x dt + \sqrt{\beta(t)(1 - e^{-2 \int_0^t \beta(s) ds})} dw. \quad (12)$$

우리 모든 SDE는 affine이며 작용. \therefore 식(12)이 훈련적.

4. Solving the reverse SDE.

Score를 model로 훈련한 후, S_0 를 reverse-SDE를 구성하고,

sample을 생성하는 numerical approach를 simulate 하는 것에 사용한다.

- General-purpose numerical SDE solvers.

Numerical solver는 SDE의 근제를 근사한다.

Stochastic dynamics의 다른 양자화로 대응되는

Euler-maruyama, stochastic Runge-kutta 같은 많은 general-purpose numerical method가 존재

DDPM에서 사용된 ancestral-sampling은 reverse-time VPSDE의 양자화에 대응되지만,

non-trivial하다. $\Delta t \rightarrow 0$ 때면 \rightarrow Appendix E.

∴ Reverse Diffusion sampler를 제안한다.

- Predictor-Corrector Samplers

다른 generic SDE와 다르게 $S_{0x}(x,t) \approx \nabla_x \log P_t(x)$ 를 찾고 있기 때문에

Langevin MCMC, HMC 같은 P_t 를 바로 sample이 가능하고,

numerical solver의 solution을 맞추는 score-based MCMC를 쓸 수 있다.

Predictor: 다음 step의 sample을 estimate numerical solver

Corrector: estimate sample의 marginal distribution을 맞추는 score-based MCMC

• PC sampler. Appendix G.

PC는 SMLD: Predictor - Identity, Corrector - Langevin dynamics

DDPM : Predictor - ancestral sampling, Corrector - Identity

Table 1: Comparing different reverse-time SDE solvers on CIFAR-10. Shaded regions are obtained with the same computation (number of score function evaluations). Mean and standard deviation are reported over five sampling runs. “P1000” or “P2000”: predictor-only samplers using 1000 or 2000 steps. “C2000”: corrector-only samplers using 2000 steps. “PC1000”: Predictor-Corrector (PC) samplers using 1000 predictor and 1000 corrector steps.

FID ↓\n\\ Sampler \\ Predictor	Variance Exploding SDE (SMLD)				Variance Preserving SDE (DDPM)			
	P1000	P2000	C2000	PC1000	P1000	P2000	C2000	PC1000
ancestral sampling	$4.98 \pm .06$	$4.88 \pm .06$		$3.62 \pm .03$	$3.24 \pm .02$	$3.24 \pm .02$		$3.21 \pm .02$
reverse diffusion	$4.79 \pm .07$	$4.74 \pm .08$	$20.43 \pm .07$	$3.60 \pm .02$	$3.21 \pm .02$	$3.19 \pm .02$	$19.06 \pm .06$	$3.18 \pm .01$
probability flow	$15.41 \pm .15$	$10.54 \pm .08$		$3.51 \pm .04$	$3.59 \pm .04$	$3.23 \pm .03$		$3.06 \pm .03$

- Probabilistic flow and connection to neural ODEs.

Score-based SDE는 다른 numerical method를 가능하게 한다.

모든 diffusion 과정에서 drift는 share 한다.

이 deterministic한 과정은 ODE를 만족한다. → Appendix D.1.

$$dx = \left[f(x, t) - \frac{1}{2} g(t)^2 \nabla_x \log p_t(x) \right] dt, \quad (13)$$

↳ probability flow ODE ~ NODE.

* Exact likelihood Computation

ODE solver로, input data와 같은 정확한 likelihood를 얻는다. → Appendix D.2.

Table 2: NLLs and FIDs (ODE) on CIFAR-10.

Model	NLL Test ↓	FID ↓
RealNVP (Dinh et al., 2016)	3.49	-
iResNet (Behrmann et al., 2019)	3.45	-
Glow (Kingma & Dhariwal, 2018)	3.35	-
MintNet (Song et al., 2019b)	3.32	-
Residual Flow (Chen et al., 2019)	3.28	46.37
FFJORD (Grathwohl et al., 2018)	3.40	-
Flow++ (Ho et al., 2019)	3.29	-
DDPM (L) (Ho et al., 2020)	$\leq 3.70^*$	13.51
DDPM (L_{simple}) (Ho et al., 2020)	$\leq 3.75^*$	3.17
DDPM	3.28	3.37
DDPM cont. (VP)	3.21	3.69
DDPM cont. (sub-VP)	3.05	3.56
DDPM++ cont. (VP)	3.16	3.93
DDPM++ cont. (sub-VP)	3.02	3.16
DDPM++ cont. (deep, VP)	3.13	3.08
DDPM++ cont. (deep, sub-VP)	2.99	2.92

→ 다른 DDPM의 적용에서도 더 뛰어나고.

improved된 DDPM이라는

Sub-VPSDE의 경우 VPSDE보다 훨씬

* Manipulating latent representations.

NODE 및 NFE 같은 invertible model의 editing, interpolating 등을 위한 latent manipulating 가능

↳ Appendix D.4.



Figure 3: Probability flow ODE enables fast sampling with adaptive stepsizes as the numerical precision is varied (left), and reduces the number of score function evaluations (NFE) without harming quality (middle). The invertible mapping from latents to images allows for interpolations (right).

* Uniquely Identifiable encoding

SDE의 trainable parameter에 따른 충분한 data의 capacity 때문에 각각이 unique

↳ Appendix D.5.

* Efficient sampling

NODE의 경우 X로부터 좋은 샘플의 sample을 뽑을 수 있다.

Fixed discretized의 경우 correctors 같이 사용되면 성능↑

Black-box ODE solver의 경우, 고질의 성능과 효율성의 trade-off를 조절할 수 있다.

↳ Appendix D.4.

Table 3: CIFAR-10 sample quality.

Model	FID↓	IS↑
Conditional		
BigGAN (Brock et al., 2018)	14.73	9.22
StyleGAN2-ADA (Karras et al., 2020a)	2.42	10.14
Unconditional		
StyleGAN2-ADA (Karras et al., 2020a)	2.92	9.83
NCSN (Song & Ermon, 2019)	25.32	$8.87 \pm .12$
NCSNv2 (Song & Ermon, 2020)	10.87	$8.40 \pm .07$
DDPM (Ho et al., 2020)	3.17	$9.46 \pm .11$
DDPM++	2.78	9.64
DDPM++ cont. (VP)	2.55	9.58
DDPM++ cont. (sub-VP)	2.61	9.56
DDPM++ cont. (deep, VP)	2.41	9.68
DDPM++ cont. (deep, sub-VP)	2.41	9.57
NCSN++	2.45	9.73
NCSN++ cont. (VE)	2.38	9.83
NCSN++ cont. (deep, VE)	2.20	9.89

- Architecture Improvement.

↳ Appendix H.

5. Controllable generation.

Continuous의 경우 P_0 이 모두 P_t 로 변한 data를 sampling 할 수 있다.

$$dx = \{f(x, t) - g(t)^2 [\nabla_x \log p_t(x) + \nabla_x \log p_t(y | x)]\} dt + g(t) d\bar{w}. \quad (14)$$

→ Appendix I.4.

Class-conditional generation, image imputation and Colorization 가능.

→ Appendix I.

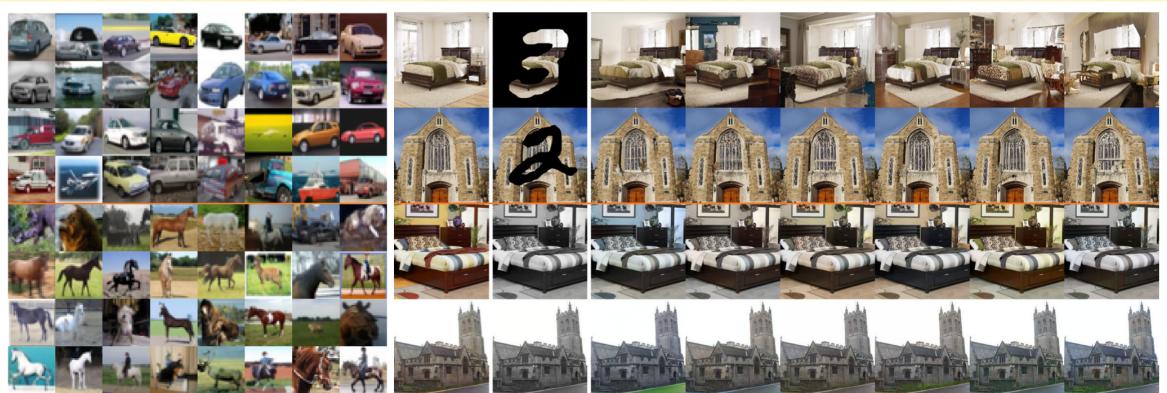


Figure 4: Left: Class-conditional samples on 32×32 CIFAR-10. Top four rows are automobiles and bottom four rows are horses. Right: Inpainting (top two rows) and colorization (bottom two rows) results on 256×256 LSUN. First column is the original image, second column is the masked/grayscale image, remaining columns are sampled image completions or colorizations.

6. Conclusion.

SDE 기반 Score-based Generative model framework 제시.

기존방식의 경직한 계산, 통합 계산 등을 이론화, condition을 추가함.

Appendix

A. The framework for more general SDEs

SDE: $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + \mathbf{G}(\mathbf{x}, t)d\mathbf{w}, \quad (15)$

$$\mathbf{f}(\cdot, t) : \mathbb{R}^d \rightarrow \mathbb{R}^d, \quad \mathbf{G}(\cdot, t) : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$$

Reverse SDE:

$$d\mathbf{x} = \{\mathbf{f}(\mathbf{x}, t) - \nabla \cdot [\mathbf{G}(\mathbf{x}, t)\mathbf{G}(\mathbf{x}, t)^T] - \mathbf{G}(\mathbf{x}, t)\mathbf{G}(\mathbf{x}, t)^T \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\}dt + \mathbf{G}(\mathbf{x}, t)d\bar{\mathbf{w}}, \quad (16)$$

Probability flow ODE:

→ Appendix D.

$$d\mathbf{x} = \left\{ \mathbf{f}(\mathbf{x}, t) - \frac{1}{2} \nabla \cdot [\mathbf{G}(\mathbf{x}, t)\mathbf{G}(\mathbf{x}, t)^T] - \frac{1}{2} \mathbf{G}(\mathbf{x}, t)\mathbf{G}(\mathbf{x}, t)^T \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \right\} dt. \quad (17)$$

Conditional reversed-time SDE

$$d\mathbf{x} = \{\mathbf{f}(\mathbf{x}, t) - \nabla \cdot [\mathbf{G}(\mathbf{x}, t)\mathbf{G}(\mathbf{x}, t)^T] - \mathbf{G}(\mathbf{x}, t)\mathbf{G}(\mathbf{x}, t)^T \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) - \mathbf{G}(\mathbf{x}, t)\mathbf{G}(\mathbf{x}, t)^T \nabla_{\mathbf{x}} \log p_t(\mathbf{y} | \mathbf{x})\}dt + \mathbf{G}(\mathbf{x}, t)d\bar{\mathbf{w}}. \quad (18)$$

Transition kernel or affine or other reverse kernel은 closed-form으로 드러나는 경우

∴ Sliced Score Matching

$$\theta^* = \arg \min_{\theta} \mathbb{E}_t \left\{ \lambda(t) \mathbb{E}_{\mathbf{x}(0)} \mathbb{E}_{\mathbf{x}(t)} \mathbb{E}_{\mathbf{v} \sim p_{\mathbf{v}}} \left[\frac{1}{2} \|\mathbf{s}_{\theta}(\mathbf{x}(t), t)\|_2^2 + \mathbf{v}^T \mathbf{s}_{\theta}(\mathbf{x}(t), t) \mathbf{v} \right] \right\}, \quad (19)$$

B. VE, VP and sub-VP SDEs.

SMLD \rightarrow VE

DDPM \rightarrow VP

SMLD의 경우, N noise scale 있으니,

$$\mathbf{x}_i = \mathbf{x}_{i-1} + \sqrt{\sigma_i^2 - \sigma_{i-1}^2} \mathbf{z}_{i-1}, \quad i = 1, \dots, N,$$

$N \rightarrow \infty$ 면, continuous.

$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \sqrt{\sigma^2(t + \Delta t) - \sigma^2(t)} \mathbf{z}(t) \approx \mathbf{x}(t) + \sqrt{\frac{d[\sigma^2(t)]}{dt} \Delta t} \mathbf{z}(t),$$

$t \rightarrow 0$ 이면,

$$d\mathbf{x} = \sqrt{\frac{d[\sigma^2(t)]}{dt}} dw, \quad \rightarrow \text{VESDE}$$

DDPM의 경우,

$$\mathbf{x}_i = \sqrt{1 - \beta_i} \mathbf{x}_{i-1} + \sqrt{\beta_i} \mathbf{z}_{i-1}, \quad i = 1, \dots, N,$$

$N \rightarrow \infty$ 면,

$$\mathbf{x}_i = \sqrt{1 - \frac{\beta_i}{N}} \mathbf{x}_{i-1} + \sqrt{\frac{\beta_i}{N}} \mathbf{z}_{i-1}, \quad i = 1, \dots, N.$$

$\Delta t \rightarrow \frac{t}{N}$ 이면,

$$\begin{aligned} \mathbf{x}(t + \Delta t) &= \sqrt{1 - \beta(t + \Delta t) \Delta t} \mathbf{x}(t) + \sqrt{\beta(t + \Delta t) \Delta t} \mathbf{z}(t) \\ &\approx \mathbf{x}(t) - \frac{1}{2} \beta(t + \Delta t) \Delta t \mathbf{x}(t) + \sqrt{\beta(t + \Delta t) \Delta t} \mathbf{z}(t) \\ &\approx \mathbf{x}(t) - \frac{1}{2} \beta(t) \Delta t \mathbf{x}(t) + \sqrt{\beta(t) \Delta t} \mathbf{z}(t), \end{aligned} \quad (24)$$

$\Delta t \rightarrow 0$ 이면,

$$d\mathbf{x} = -\frac{1}{2} \beta(t) \mathbf{x} dt + \sqrt{\beta(t)} dw. \quad \rightarrow \text{VPSDE}$$

VPOIM Variance를 minimize하는 ODE를 의미함

$$\frac{d\Sigma_{VP}(t)}{dt} = \beta(t)(\mathbf{I} - \Sigma_{VP}(t)),$$

ODE를 풀면,

$$\Sigma_{VP}(t) = \mathbf{I} + e^{\int_0^t -\beta(s)ds}(\Sigma_{VP}(0) - \mathbf{I}),$$