DDPM은 markov를 기반으로 함

하지만 속도가 느림

∴ x_0를 condition으로. markov로 끊음.

1. Introduction.

더 빠르고 좋은 sample quality

sample의 consistency

↳ interpolation 가능.

2. Background.

$$p_\theta(\boldsymbol{x}_0) = \int p_\theta(\boldsymbol{x}_{0:T}) \mathrm{d}\boldsymbol{x}_{1:T}, \quad \text{where} \quad p_\theta(\boldsymbol{x}_{0:T}) := p_\theta(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \tag{1}$$

$$\max_\theta \mathbb{E}_{q(\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0)] \leq \max_\theta \mathbb{E}_{q(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)} [\log p_\theta(\boldsymbol{x}_{0:T}) - \log q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)] \tag{2}$$

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) := \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \text{where } q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := \mathcal{N}\left(\sqrt{\frac{\alpha_t}{\alpha_{t-1}}}\boldsymbol{x}_{t-1}, \left(1 - \frac{\alpha_t}{\alpha_{t-1}}\right)\boldsymbol{I}\right) \tag{3}$$

- forward process

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) := \int q(\boldsymbol{x}_{1:t}|\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}_{1:(t-1)} = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_0, (1-\alpha_t)\boldsymbol{I});$$

eq 2는 x_T가 gaussian이고, trainable mean, fixed variance 일 때,

$$L_\gamma(\epsilon_\theta) := \sum_{t=1}^{T} \gamma_t \mathbb{E}_{\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), \epsilon_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})} \left[\left\| \epsilon_\theta^{(t)}(\sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\epsilon_t) - \epsilon_t \right\|_2^2 \right] \tag{5}$$

↳ = DDPM

T는 hyperparam인데, T가 클수록 gaussian에 가까워지기 때문에 좋음! → 하지만 느림.

3. Variational Inference for non-markovian Forward process

Reverse는 inference의 근사이므로 iter를 줄이기 위해 infer 과정을 생각.

DDPM은 $p(x_t|x_0)$에 대한 objective func를 사용하고,

$p(x_{1:T}|x_0)$의 joint dist를 직접적으로 사용하지 않음

따라서 반복적인 marginal이 계산되고, 저자는 non-markovian을 탐색

Figure 1: Graphical models for diffusion (left) and non-Markovian (right) inference models.

DDPM과 같은 objective를 사용가능

3.1 Non-markovian forward process

Q를 inference family라고 하고, index $\sigma \in \mathbb{R}^T_{\geq 0}$ 일 때,

$$q_\sigma(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) := q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0)\prod_{t=2}^{T} q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \tag{6}$$

ㄴ $q_\sigma(x_T|x_0) = \mathcal{N}(\sqrt{\alpha_T}x_0, (1-\alpha_T)I)$

?

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\frac{\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\boldsymbol{I}\right). \tag{7}$$

ㄴ $\alpha_T = \prod_{s=1}^{t} 1-\beta_s$

ㄴ $q_\sigma(x_t|x_0) = \mathcal{N}(\sqrt{\alpha_t}x_0, (1-\alpha_t)I)$

→ 증명은 Appendix A.B

따라서 joint inference distribution의 marginal이 된

Forward process는

$$q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0) = \frac{q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}, \tag{8}$$

이고, x_t가 x_{t-1}, x_0 두개에 의존하므로, markovian이 아니다.

σ가 충분히 작으면, x_0, x_t에 대해 x_{t-1}는 fix가 됨

3.2 Generative process and Unified Variational Inference Objective.

$q_\sigma(x_{t-1}|x_t, x_0)$를 알 때, $p_\theta(x_{0:T})$를 학습

직관적으로, x_t가 주어지면 x_0를 예측하고, reverse $q_\sigma(x_{t-1}|x_t, x_0)$를 이용하여 x_{t-1} 얻음

x_t가 주어지면, $\epsilon_\theta^t(x_t)$는 x_t에서 ϵ_t를 추론하고,

x_0의 denoised prediction은

$$f_\theta^{(t)}(\boldsymbol{x}_t) := (\boldsymbol{x}_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta^{(t)}(\boldsymbol{x}_t))/\sqrt{\alpha_t}. \tag{9}$$

따라서 generative process는

$$p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \begin{cases} \mathcal{N}(f_\theta^{(1)}(\boldsymbol{x}_1), \sigma_1^2 \boldsymbol{I}) & \text{if } t = 1 \\ q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, f_\theta^{(t)}(\boldsymbol{x}_t)) & \text{otherwise,} \end{cases} \tag{10}$$

θ의 optimize는

$$J_\sigma(\epsilon_\theta) := \mathbb{E}_{\boldsymbol{x}_{0:T} \sim q_\sigma(\boldsymbol{x}_{0:T})}[\log q_\sigma(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) - \log p_\theta(\boldsymbol{x}_{0:T})] \tag{11}$$

$$= \mathbb{E}_{\boldsymbol{x}_{0:T} \sim q_\sigma(\boldsymbol{x}_{0:T})} \left[q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0) + \sum_{t=2}^{T} \log q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) - \sum_{t=1}^{T} \log p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) - \log p_\theta(\boldsymbol{x}_T) \right]$$

Lemma 1. $J_\sigma(\epsilon_\theta)$는 $\sigma > 0$ 일 때, $L_\gamma + C$와 같다. ↳ t를 random sampling하기 때문인듯

L_γ은 각 t에서 θ를 공유하지 않으면, optimal 이 γ에 의존 x.

이점 1. DDPM에서 γ을 1로 고정하는 것에 대한 정당성

2. J_σ의 optimal도 L_1과 같다.

4. Sampling from Generalized Generative processes

DDIM에서는 generative뿐 아니라 σ에 의해 매개변수화된 forward도 배운다.

따라서 pre-trained DDPM을 사용하여 σ를 탐색

4.1 DDIM.

Eq.10에 의해 x_t에서 x_{t-1}은

$$x_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\left(\frac{x_t - \sqrt{1-\alpha_t}\,\epsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}} \right)}_{\text{“ predicted } x_0 \text{”}} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_\theta^{(t)}(x_t)}_{\text{“direction pointing to } x_t \text{”}} + \underbrace{\sigma_t \epsilon_t}_{\text{random noise}} \qquad (12)$$

σ의 선택은 같은 ϵ_θ에서도 다른 생성 결과

모든 t에서, $\sigma_t = \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}} \sqrt{\frac{1-\alpha_t}{\alpha_{t-1}}}$ 면, DDPM과 같다. (markovian)

모든 t에서, $\sigma_t = 0$ 일때, 모든 forward는 x_{t-1}, x_0에 대해 deterministic ($t \neq 1$)

→ 생성 process에서 ϵ_t의 계수는 0

→ Implicit probabilistic model (x_T에서 x_0로 정해진 길을 따라감)

→ DDIM (forward가 diffusion이 아니어도 같은 objective function)

↳ generate 시에만 $\sigma_t=0$ 인듯...? → 코드 확인!

4.2 Accelerated generation processes

$q_\sigma(x_t|x_0)$를 바로 구할수 있기 때문에 step을 건너 뛸 수 있음

따라서, 재학습 없이 건너 뛰는 방법 소개.

$\{x_{\tau_1}, \cdots, x_{\tau_S}\}$를 $x_{1:T}$의 subset이라고 하면

$q(x_{\tau_i}|x_0) = \mathcal{N}(\sqrt{\alpha_{\tau_i}}\, x_0, (1-\alpha_{\tau_i})I)$ 이며, 다음과 같다.

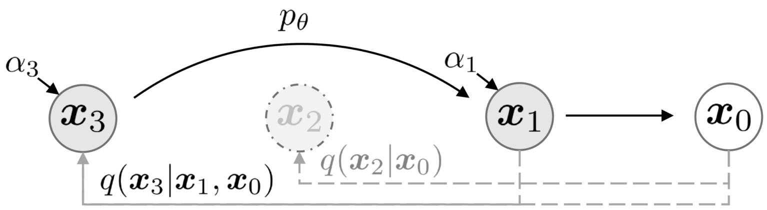

Figure 2: Graphical model for accelerated generation, where $\tau = [1, 3]$.

reverse(τ)에 따라 sampling하면, T보다 τ가 작을 때, 계산 효율성 ↑.

즉, eq.12와 같이 update 방식만 바꿔서, 새롭고, 빠른 sampling 가능

→ Appendix C.

즉, 훈련은 T만큼 하더라도, sampling은 그 중 일부만 사용 가능

그리고 연속적 t에 대해 sampling 가능

4.3 Relevance to Neural ODEs.

DDIM의 eq.12를 다시 쓰면,

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\left(\frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\,\epsilon_\theta^{(t)}(\boldsymbol{x}_t)}{\sqrt{\alpha_t}} \right)}_{\text{“ predicted } \boldsymbol{x}_0\text{”}} + \underbrace{\sqrt{1 - \alpha_{t-1} - \cancelto{0}{\sigma_t^2}} \cdot \epsilon_\theta^{(t)}(\boldsymbol{x}_t)}_{\text{“direction pointing to } \boldsymbol{x}_t\text{”}} + \underbrace{\cancelto{0}{\sigma_t \epsilon_t}}_{\text{random noise}} \quad (12)$$

ODE를 위한, Euler-integration과 비교

$$\frac{\boldsymbol{x}_{t-\Delta t}}{\sqrt{\alpha_{t-\Delta t}}} = \frac{\boldsymbol{x}_t}{\sqrt{\alpha_t}} + \left(\sqrt{\frac{1-\alpha_{t-\Delta t}}{\alpha_{t-\Delta t}}} - \sqrt{\frac{1-\alpha_t}{\alpha_t}} \right) \epsilon_\theta^{(t)}(\boldsymbol{x}_t) \quad (13)$$

$\sigma = \frac{\sqrt{1-\alpha}}{\alpha}$, $\bar{x} = \frac{x}{\sqrt{\alpha}}$라 하면, continuous에서, σ, x는 t의 함수고, σ가 continuous일 때, eq.13은

$$\mathrm{d}\bar{\boldsymbol{x}}(t) = \epsilon_\theta^{(t)} \left(\frac{\bar{\boldsymbol{x}}(t)}{\sqrt{\sigma^2+1}} \right) \mathrm{d}\sigma(t), \quad (14)$$

init cond: $x(T) \sim N(0, \sigma(T))$ for a very large $\sigma(T)$

→ 충분한 step으로 x_0, x_T를 discretize하면, x_0에서 x_T를 encoding 하고

eq.14의 ODE를 reverse해서 simulate 가능

→ 즉, DDPM과 달리, DDIM은 image ↔ latent 1:1 matching이 가능

↳ SDE-score의 probabilistic flow ODE와 비슷

→ DDIM은 continuous DDPM의 special case

Proposition 1. optimal $\epsilon_\theta^{(t)}$와 eq.14의 ODE는 Score-SDE의 VE-SDE에 해당하는 probability flow ODE를 갖는다.

→ Appendix B.

ODE는 같지만, sampling은 다름.

probability flow ODE의 Euler 방법의 update는

$$\frac{\boldsymbol{x}_{t-\Delta t}}{\sqrt{\alpha_{t-\Delta t}}} = \frac{\boldsymbol{x}_t}{\sqrt{\alpha_t}} + \frac{1}{2}\left(\frac{1-\alpha_{t-\Delta t}}{\alpha_{t-\Delta t}} - \frac{1-\alpha_t}{\alpha_t}\right) \cdot \sqrt{\frac{\alpha_t}{1-\alpha_t}} \cdot \epsilon_\theta^{(t)}(\boldsymbol{x}_t) \tag{15}$$

하지만 step 선택에서 다름

Proposed: $d\sigma(t)$

Score-ODE: dt

5. Experiments.

5.1 Sample quality and efficiency

Table 1: CIFAR10 and CelebA image generation measured in FID. $\eta = 1.0$ and $\hat{\sigma}$ are cases of DDPM (although Ho et al. (2020) only considered $T = 1000$ steps, and $S < T$ can be seen as simulating DDPMs trained with S steps), and $\eta = 0.0$ indicates DDIM.

	S	CIFAR10 (32 × 32) 10	20	50	100	1000	CelebA (64 × 64) 10	20	50	100	1000
η	0.0	**13.36**	**6.84**	**4.67**	**4.16**	4.04	**17.33**	**13.73**	**9.17**	**6.53**	3.51
	0.2	14.04	7.11	4.77	4.25	4.09	17.66	14.11	9.51	6.79	3.64
	0.5	16.66	8.35	5.25	4.46	4.29	19.86	16.06	11.01	8.09	4.28
	1.0	41.07	18.36	8.01	5.78	4.73	33.12	26.03	18.48	13.93	5.98
$\hat{\sigma}$		367.43	133.37	32.72	9.99	**3.17**	299.71	183.83	71.71	45.20	**3.26**

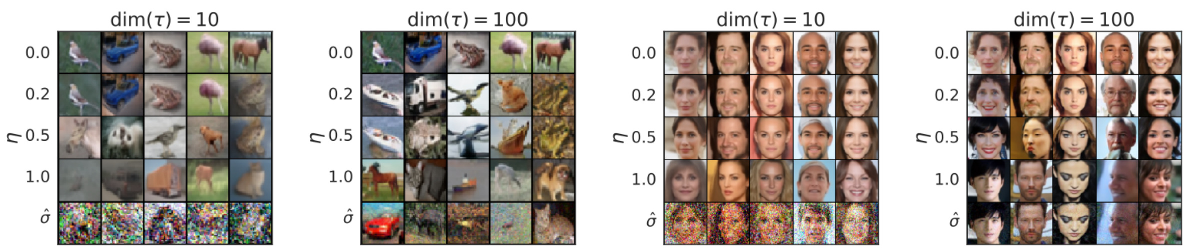

Figure 3: CIFAR10 and CelebA samples with $\dim(\tau) = 10$ and $\dim(\tau) = 100$.

5.2 Sample Consistency in DDIM.

DDIM에서 x_T와 x_0는 1:1 matching

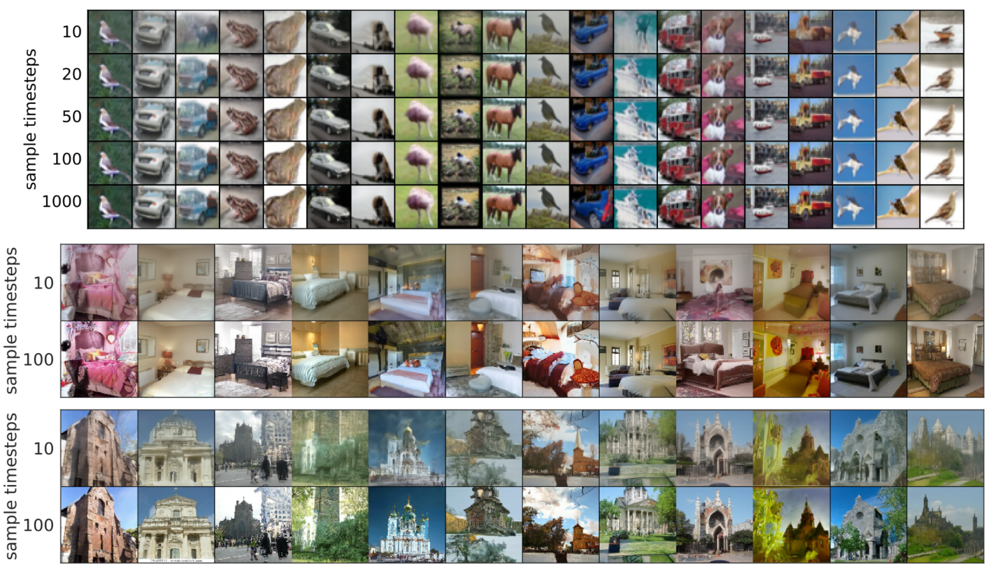

Figure 5: Samples from DDIM with the same random x_T and different number of steps.

1000 step 중, 20 step부터 전체윤곽이 정해짐

→ x_T만 informative latent.

5.4 Interpolation in deterministic Generative process

→ 가능함

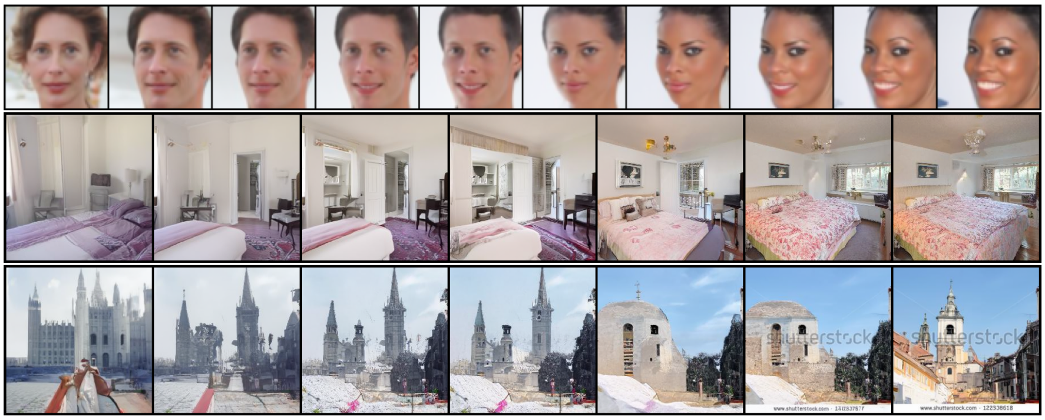

Figure 6: Interpolation of samples from DDIM with $\dim(\tau) = 50$.

5.4. Reconstruction from Latent Space

DDIM이 encoder로 사용될 수 있는지.

→ 특정한 ODE에 대한 Euler Integration이기 때문

Table 2: Reconstruction error with DDIM on CIFAR-10 test set, rounded to 10^{-4}.

S	10	20	50	100	200	500	1000
Error	0.014	0.0065	0.0023	0.0009	0.0004	0.0001	0.0001

Appendix. A. Non-markovian Forward Processes for a Discrete Case.

Discrete data와 그에 해당하는 variational objectives에 대한 non-markovian을 설명

Appendix. B.

$$q_\sigma(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) := q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0)\prod_{t=2}^{T} q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) \tag{6}$$

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\frac{\boldsymbol{x}_t-\sqrt{\alpha_t}\boldsymbol{x}_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\boldsymbol{I}\right). \tag{7}$$

여기서 평균은 $q_\sigma(x_t|x_0) = N(\sqrt{\alpha_t}x_0, (1-\alpha_t)I)$를 만족하기 위해 결정됨.

Lemma 1. *For $q_\sigma(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$ defined in Eq. (6) and $q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$ defined in Eq. (7), we have:*

$$q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{x}_0, \sqrt{1-\alpha_t}\boldsymbol{I}) \tag{22}$$

이거면

이거

Proofs.

$t<T$ 일 때, $q_\sigma(x_t|x_0) = N(\sqrt{\alpha_t}x_0, \sqrt{1-\alpha_t}I)$라 가정.

만약, $q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0, \sqrt{1-\alpha_{t-1}}\boldsymbol{I})$ 가 성립하면,

t일 때가 이미 성립하기 때문에, T→1로의 t의 귀납으로 증명 가능

먼저,

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0) := \int_{\boldsymbol{x}_t} q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_0)q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}_t$$

이고,

$$q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{x}_0, (1-\alpha_t)\boldsymbol{I}) \tag{24}$$

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\frac{\boldsymbol{x}_t-\sqrt{\alpha_t}\boldsymbol{x}_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\boldsymbol{I}\right). \tag{25}$$

다

PRML 2.115 에서 보면,

Marginal and Conditional Gaussians

Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.113}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1}) \tag{2.114}$$

the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \tag{2.115}$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \tag{2.116}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}. \tag{2.117}$$

$x_t = X,\ x_{t-1}|x_0 = y$ 일 때,

$\mu = \sqrt{\alpha_t}\,x_0,\ \Lambda^{-1} = (1-\alpha_t)I.$

$$Ax_t + b = \sqrt{\alpha_{t-1}}\,x_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\frac{x_t - \sqrt{\alpha_t}\,x_0}{\sqrt{1-\alpha_t}}$$

$$A = \frac{\sqrt{1-\alpha_{t-1}-\sigma_t^2}}{\sqrt{1-\alpha_t}},\quad b = \left(\sqrt{\alpha_{t-1}} - \frac{\sqrt{1-\alpha_{t-1}-\sigma_t^2}}{\sqrt{1-\alpha_t}}\cdot\sqrt{\alpha_t}\right)\cdot x_0$$

$$L^{-1} = \sigma_t^2 I$$

$$p(y) = p(x_{t-1}|x_0) = N(A\mu + b,\ L^{-1} + A\Lambda^{-1}A^T)$$

$$A\mu + b = \cancel{\frac{\sqrt{1-\alpha_{t-1}-\sigma_t^2}}{\sqrt{1-\alpha_t}}\cdot\sqrt{\alpha_t}\,x_0} + \left(\sqrt{\alpha_{t-1}} - \cancel{\frac{\sqrt{1-\alpha_{t-1}-\sigma_t^2}}{\sqrt{1-\alpha_t}}\cdot\sqrt{\alpha_t}}\right)\cdot x_0$$

$$= \sqrt{\alpha_{t-1}}\,x_0$$

$$L^{-1} + A\Lambda^{-1}A^T = \sigma_t^2 I + \frac{\sqrt{1-\alpha_{t-1}-\sigma_t^2}}{\sqrt{1-\alpha_t}}\cdot(1-\alpha_t)I\cdot\frac{\sqrt{1-\alpha_{t-1}-\sigma_t^2}}{\sqrt{1-\alpha_t}}$$

$$= \sigma_t^2 I + \frac{1-\alpha_{t-1}-\sigma_t^2}{\cancel{1-\alpha_t}}\cdot\cancel{(1-\alpha_t)}I$$

$$= (1-\alpha_{t-1})I$$

$$\therefore q_\sigma(x_{t-1}|x_0) = N(\sqrt{\alpha_{t-1}}\,x_0,\ \sqrt{1-\alpha_{t-1}}\,I)$$

$$L_\gamma(\epsilon_\theta) := \sum_{t=1}^{T} \gamma_t \mathbb{E}_{\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), \epsilon_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})} \left[\left\| \epsilon_\theta^{(t)}(\sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\epsilon_t) - \epsilon_t \right\|_2^2 \right] \tag{5}$$

γ는 DDPM에서는 1로 취급되는 Σ의 weight.

$$J_\sigma(\epsilon_\theta) := \mathbb{E}_{\boldsymbol{x}_{0:T} \sim q_\sigma(\boldsymbol{x}_{0:T})}[\log q_\sigma(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) - \log p_\theta(\boldsymbol{x}_{0:T})] \tag{11}$$

$$= \mathbb{E}_{\boldsymbol{x}_{0:T} \sim q_\sigma(\boldsymbol{x}_{0:T})} \left[q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0) + \sum_{t=2}^{T} \log q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) - \sum_{t=1}^{T} \log p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) - \log p_\theta(\boldsymbol{x}_T) \right]$$

Theorem 1.

Theorem 1. *For all $\sigma > \mathbf{0}$, there exists $\gamma \in \mathbb{R}_{>0}^T$ and $C \in \mathbb{R}$, such that $J_\sigma = L_\gamma + C$.*

Proofs.

$$J_\sigma(\epsilon_\theta) := \mathbb{E}_{\boldsymbol{x}_{0:T} \sim q(\boldsymbol{x}_{0:T})} \left[\log q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0) + \sum_{t=2}^{T} \log q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) - \sum_{t=1}^{T} \log p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \right] \tag{29}$$

$$\equiv \mathbb{E}_{\boldsymbol{x}_{0:T} \sim q(\boldsymbol{x}_{0:T})} \left[\sum_{t=2}^{T} D_{\mathrm{KL}}(q_{\boldsymbol{\sigma}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)) \| p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) - \log p_\theta^{(1)}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right]$$

t>1 일 때,

$$\mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{x}_t \sim q(\boldsymbol{x}_0, \boldsymbol{x}_t)}[D_{\mathrm{KL}}(q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)) \| p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

$$= \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{x}_t \sim q(\boldsymbol{x}_0, \boldsymbol{x}_t)}[D_{\mathrm{KL}}(q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)) \| q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, f_\theta^{(t)}(\boldsymbol{x}_t)))]$$

ㄴ ϵ_θ^t로부터 온

$$= \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{x}_t \sim q(\boldsymbol{x}_0, \boldsymbol{x}_t)} \left[\frac{\| \boldsymbol{x}_0 - f_\theta^{(t)}(\boldsymbol{x}_t) \|_2^2}{2\sigma_t^2} \right] \tag{30}$$

→ Gaussian 식의 지수부

$$= \mathbb{E}_{\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\epsilon} \left[\frac{\| (\boldsymbol{x}_t - \epsilon)/\sqrt{\alpha_t} - (\boldsymbol{x}_t - \epsilon_\theta^{(t)}(\boldsymbol{x}_t))/\sqrt{\alpha_t} \|_2^2}{2\sigma_t^2} \right] \tag{31}$$

$$= \mathbb{E}_{\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\epsilon} \left[\frac{\| \epsilon - \epsilon_\theta^{(t)}(\boldsymbol{x}_t) \|_2^2}{2d\sigma_t^2\alpha_t} \right] \tag{32}$$

(30) → (32)

제는 제곱되어서 빠지기 때문에

$$f_\theta^{(t)}(\boldsymbol{x}_t) := (\boldsymbol{x}_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta^{(t)}(\boldsymbol{x}_t))/\sqrt{\alpha_t}.$$

$$\frac{\cancel{x_t} - \sqrt{1-\alpha_t}\,\epsilon}{\sqrt{\alpha_t}} - \frac{\cancel{x_t} - \sqrt{1-\alpha_t} \cdot \epsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}} \Rightarrow \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}}(\epsilon - \epsilon_\theta^{(t)}(x_t))$$

그리고,

$$\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{x}_1\sim q(\boldsymbol{x}_0,\boldsymbol{x}_1)}\left[-\log p_\theta^{(1)}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] \equiv \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{x}_1\sim q(\boldsymbol{x}_0,\boldsymbol{x}_1)}\left[\frac{\|\boldsymbol{x}_0 - f_\theta^{(t)}(\boldsymbol{x}_1)\|_2^2}{2\sigma_1^2}\right] \tag{33}$$

$$= \mathbb{E}_{\boldsymbol{x}_0\sim q(\boldsymbol{x}_0),\epsilon\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}),\boldsymbol{x}_1=\sqrt{\alpha_1}\boldsymbol{x}_0+\sqrt{1-\alpha_t}\epsilon}\left[\frac{\|\epsilon - \epsilon_\theta^{(1)}(\boldsymbol{x}_1)\|_2^2}{2d\sigma_1^2\alpha_1}\right] \tag{34}$$

따라서,

$$J_\sigma(\epsilon_\theta) \equiv \sum_{t=1}^{T}\frac{1}{2d\sigma_t^2\alpha_t}\mathbb{E}\left[\|\epsilon_\theta^{(t)}(\boldsymbol{x}_t) - \epsilon_t\|_2^2\right] = L_\gamma(\epsilon_\theta)$$

이고, $J_\sigma = L_\gamma + C$ 이기 때문에

DDIM 과 DDPM은 같은 구조의 손실함수로 optimal로 가질 수 있다

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}}\underbrace{\left(\frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(\boldsymbol{x}_t)}{\sqrt{\alpha_t}}\right)}_{\text{" predicted } \boldsymbol{x}_0\text{"}} + \underbrace{\sqrt{1-\alpha_{t-1}-\cancel{\sigma_t^2}^{\,0}}\cdot\epsilon_\theta^{(t)}(\boldsymbol{x}_t)}_{\text{"direction pointing to } \boldsymbol{x}_t\text{"}} + \underbrace{\cancel{\sigma_t\epsilon_t}^{\,0}}_{\text{random noise}} \tag{12}$$

$1 \to \Delta t$

rewrite

$$\frac{\boldsymbol{x}_{t-\Delta t}}{\sqrt{\alpha_{t-\Delta t}}} = \frac{\boldsymbol{x}_t}{\sqrt{\alpha_t}} + \left(\sqrt{\frac{1-\alpha_{t-\Delta t}}{\alpha_{t-\Delta t}}} - \sqrt{\frac{1-\alpha_t}{\alpha_t}}\right)\epsilon_\theta^{(t)}(\boldsymbol{x}_t) \tag{13}$$

$d\sigma(t)$

$\sigma = (\sqrt{1-\alpha}/\sqrt{\alpha})$, $\bar{x} = x/\sqrt{\alpha}$

$$\mathrm{d}\bar{\boldsymbol{x}}(t) = \epsilon_\theta^{(t)}\left(\frac{\bar{\boldsymbol{x}}(t)}{\sqrt{\sigma^2+1}}\right)\mathrm{d}\sigma(t), \tag{14}$$

Proposition.1.

Proposition 1. *The ODE in Eq. (14) with the optimal model $\epsilon_\theta^{(t)}$ has an equivalent probability flow ODE corresponding to the "Variance-Exploding" SDE in Song et al. (2020).*

Appendix C. Additional Derivations.

C.1. Accelerated Sampling Processes

Inference는

$$q_{\sigma,\tau}(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = q_{\sigma,\tau}(\boldsymbol{x}_{\tau_S}|\boldsymbol{x}_0)\prod_{i=1}^{S} q_{\sigma,\tau}(\boldsymbol{x}_{\tau_{i-1}}|\boldsymbol{x}_{\tau_i},\boldsymbol{x}_0)\prod_{t\in\bar{\tau}} q_{\sigma,\tau}(\boldsymbol{x}_t|\boldsymbol{x}_0) \tag{52}$$