DDPM: Denoising Diffusion Probabilistic Models

jojonki

May 2022

1 Introduction

DDPM で使われている式を導出するよ．最初に拡散過程で次に逆拡散過程．変数が太字ベクトルになっていないのは許してほしい．

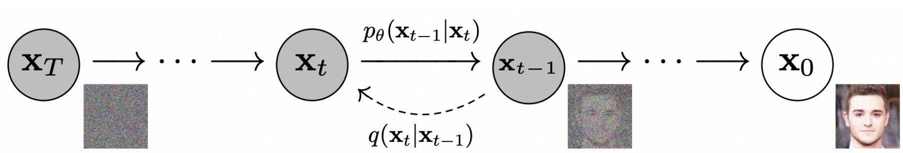

図1　DDPM

DDPM は最終的に下記のようなアルゴリズムとなる．なぜこのような定式化ができるかをこの文書で追っていく．

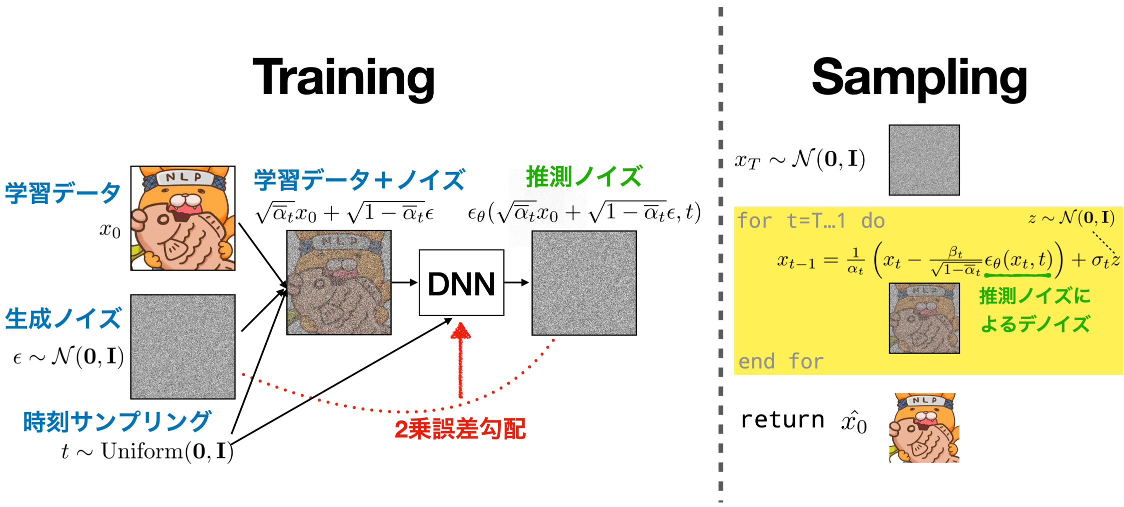

図2　DDPM の Training と Sampling

2　拡散過程，Diffusion process, Forward process

ガウスノイズを付加していく過程．図 1の左方向．純粋にガウスを足していくだけなので，学習するものではなく解析的に導出できる．

まず 1 ステップ分を見る．スケジュールパラメタ β_t を用意．学習してもよいが DDPM ではハイパラとして用意している．

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

x_t を求めるためには，x_0 からこの式を t 回展開していけばよいが，実は x_0 から reparametrization trick 等を使って 1 発で求めることが可能．

$$\begin{aligned}
x_t &= \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_{t-1}; \mathrm{where} \epsilon \sim \mathcal{N}(0, \mathbf{I}) \\
&\quad 1-\beta_t = \alpha_t \text{ とおく} \\
&= \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1} \\
&\quad x_{t-1} \text{ を次に展開} \\
&= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_{t-1}}\epsilon_{t-2}) + \sqrt{1-\alpha_t}\epsilon_{t-1} \\
&\quad 2 \text{ つのガウス分布 } \mathcal{N}(0, \sigma_a^2)\text{，}\mathcal{N}(0, \sigma_b^2) \text{ をマージすると分散は } \sigma_a^2 + \sigma_b^2 \text{ になることを利用し} \\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{(\alpha_t - \alpha_t\alpha_{t-1}) + (1-\alpha_t)}\bar{\epsilon}_{t-2} \\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\epsilon}_{t-2} \\
&\quad x_{t-2} \text{ を次に展開} \\
&= \sqrt{\alpha_t\alpha_{t-1}}(\sqrt{\alpha_{t-2}}x_{t-3} + \sqrt{1-\alpha_{t-2}}\epsilon_{t-3}) + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\epsilon}_{t-2} \\
&= \sqrt{\alpha_t\alpha_{t-1}\alpha_{t-2}}x_{t-3} + \sqrt{1-\alpha_t\alpha_{t-1}\alpha_{t-2}}\bar{\epsilon}_{t-3} \\
&\quad \text{続けていくと} \\
&= \sqrt{\alpha_t\alpha_{t-1}...\alpha_1}x_0 + \sqrt{1-\alpha_t\alpha_{t-1}...\alpha_1}\epsilon \\
&\quad \overline{\alpha}_t = \prod_{i=1}^{t}\alpha_i \text{ とおく} \\
&= \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon
\end{aligned} \tag{2}$$

以上から，1 つのガウス分布で x_0 から直接 x_t を表すことができることがわかる．

$$q(x_t|x_0) := \mathcal{N}(x_t; \sqrt{\overline{\alpha}_t}x_0, (1-\overline{\alpha}_t)\mathbf{I}) \tag{3}$$

3　逆拡散過程，Reverse process

逆拡散過程はノイズを除去する方向．図 1の右方向．$x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ からスタートし，$q(x_{t-1}|x_t)$ を経由して x_0 にたどり着く．β_t が十分小さい場合，$q(x_{t-1}|x_t)$ はガウス分布に従う．ただし $q(x_{t-1}|x_t)$ は容易に求められないので学習パラメタ θ を導入し，このガウス分布の平均・分散をモデルで推定する話になる．

この θ で置き換えて確率モデル p_θ を導入し，下記のように定義する．この 1 ステップはガウス分布に従うので，そのガウス分布の平均と分散が θ でパラメタライズされる．

$$p_\theta(x_{0:T}) := p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{4}$$

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{5}$$

ちなみに $q(x_{t-1}|x_t)$ は容易に求められないが，x_0 で条件づけした $q(x_{t-1}|x_t, x_0)$ は計算可能．ベイズの定例に従って計算していく．

$$\begin{aligned}
q(x_{t-1}|x_t, x_0) &= \mathcal{N}(x_{t-1} : \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t\mathbf{I}) \\
&= \frac{q(x_t, x_{t-1}, x_0)}{q(x_t, x_0)} \\
&= \frac{q(x_t|x_{t-1}, x_0)q(x_{t-1}, x_0)}{q(x_t|x_0)q(x_0)} \\
&= \frac{q(x_t|x_{t-1}, x_0)q(x_{t-1}|x_0)q(x_0)}{q(x_t|x_0)q(x_0)} \\
&= \frac{q(x_t|x_{t-1}, x_0)q(x_{t-1}|x_0)}{q(x_t|x_0)} \\
&\quad \text{分子第 1 項について } x_t \text{ は直前のステップ } x_{t-1} \text{ で条件づけしており，マルコフ性より } x_0 \text{ は不要} \\
&= \frac{q(x_t|x_{t-1})q(x_{t-1}|x_0)}{q(x_t|x_0)} \\
&\quad \text{これは式 (1), (3) で表せており，すべてガウス分布になっている} \\
&\propto \frac{\exp\left(-\frac{1}{2}\frac{(x_t-\sqrt{\alpha_t}x_{t-1})^2}{\beta_t}\right)\exp\left(-\frac{1}{2}\frac{(x_{t-1}-\sqrt{\overline{\alpha}_{t-1}}x_0)^2}{1-\overline{\alpha}_{t-1}}\right)}{\exp\left(-\frac{1}{2}\frac{(x_t-\sqrt{\overline{\alpha}_t}x_0)^2}{1-\overline{\alpha}_t}\right)} \\
&= \exp\left(-\frac{1}{2}\left(\frac{(x_t-\sqrt{\alpha_t}x_{t-1})^2}{\beta_t} + \frac{(x_{t-1}-\sqrt{\overline{\alpha}_{t-1}}x_0)^2}{1-\overline{\alpha}_{t-1}} - \frac{(x_t-\sqrt{\overline{\alpha}_t}x_0)^2}{1-\overline{\alpha}_t}\right)\right) \\
&\quad \text{最後の項は } x_{t-1} \text{ と関係ないので } C \text{ として出す} \\
&= \exp\left(-\frac{1}{2}\left(\frac{x_t^2-2\sqrt{\alpha_t}x_t x_{t-1}+\alpha_t x_{t-1}^2}{\beta_t} + \frac{x_{t-1}^2-2\sqrt{\overline{\alpha}_{t-1}}x_{t-1}x_0+\alpha_{t-1}x_0^2}{1-\overline{\alpha}_{t-1}} + C(x_t, x_0)\right)\right) \\
&\quad x_{t-1} \text{ の項でまとめる} \\
&= \exp\left(-\frac{1}{2}\left(\left(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\overline{\alpha}_{t-1}}\right)x_{t-1}^2 - \left(\frac{2\sqrt{\alpha_t}}{\beta_t}x_t + \frac{2\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_{t-1}}x_0\right)x_{t-1} + C(x_t, x_0)\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\overline{\alpha}_{t-1}}\right)x_{t-1}^2 + \left(\frac{\sqrt{\alpha_t}}{\beta_t}x_t + \frac{\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_{t-1}}x_0\right)x_{t-1} + C(x_t, x_0)\right)
\end{aligned} \tag{6}$$

ガウス分布 $\mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t\mathbf{I}) \propto \exp\left(-\frac{1}{2\tilde{\beta}_t}||x_{t-1}-\tilde{\mu}_t(x_t, x_0)||^2\right) = \exp\left(-\frac{1}{2}\frac{1}{\tilde{\beta}_t}x_{t-1}^2 + \frac{1}{\tilde{\beta}_t}\tilde{\mu}_t x_{t-1} + const\right)$

と係数を比較して $\tilde{\beta}_t$ と $\tilde{\mu}_t(x_t, x_0)$ を求める．

$$
\begin{aligned}
\tilde{\beta}_t &= \left(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\overline{\alpha}_{t-1}}\right)^{-1} \\
&= \left(\frac{\alpha_t(1-\overline{\alpha}_{t-1}) + \beta_t}{\beta_t(1-\overline{\alpha}_{t-1})}\right)^{-1} \\
&= \left(\frac{\alpha_t - \overline{\alpha}_t + 1 - \alpha_t}{\beta_t(1-\overline{\alpha}_{t-1})}\right)^{-1} \\
&= \left(\frac{1-\overline{\alpha}_t}{\beta_t(1-\overline{\alpha}_{t-1})}\right)^{-1} \\
&= \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t
\end{aligned}
\tag{7}
$$

$$
\begin{aligned}
\tilde{\mu}_t(x_t, x_0) &= \tilde{\beta}_t \left(\frac{\sqrt{\alpha_t}}{\beta_t}x_t + \frac{\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_{t-1}}x_0\right) \\
&= \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t \left(\frac{\sqrt{\alpha_t}}{\beta_t}x_t + \frac{\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_{t-1}}x_0\right) \\
&= \sqrt{\alpha_t}\frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}x_t + \frac{\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_t}\beta_t x_0 \\
&\quad \text{式 (2) より } x_t = \sqrt{\tilde{\alpha}_t}x_0 + \sqrt{1-\tilde{\alpha}_t}\epsilon_t \text{ なので } x_0 = \text{ の形にして代入} \\
&= \sqrt{\alpha_t}\frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}x_t + \frac{\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_t}\beta_t \frac{x_t - \sqrt{1-\overline{\alpha}_t}\epsilon_t}{\sqrt{\overline{\alpha}_t}} \\
&= \sqrt{\alpha_t}\frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}x_t + \frac{\sqrt{\alpha_1\alpha_2 \ldots \alpha_{t-1}}}{1-\overline{\alpha}_t}\beta_t \frac{x_t - \sqrt{1-\overline{\alpha}_t}\epsilon_t}{\sqrt{\alpha_1\alpha_2 \ldots \alpha_{t-1}\alpha_t}} \\
&= \sqrt{\alpha_t}\frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}x_t + \frac{1}{1-\overline{\alpha}_t}\beta_t \frac{x_t - \sqrt{1-\overline{\alpha}_t}\epsilon_t}{\sqrt{\alpha_t}} \\
&= \frac{1}{1-\overline{\alpha}_t}\left(\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})x_t + \beta_t \frac{x_t - \sqrt{1-\overline{\alpha}_t}\epsilon_t}{\sqrt{\alpha_t}}\right) \\
&= \frac{1}{1-\overline{\alpha}_t}\left(\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})x_t + \frac{\beta_t}{\sqrt{\alpha_t}}x_t - \beta_t \frac{\sqrt{1-\overline{\alpha}_t}\epsilon_t}{\sqrt{\alpha_t}}\right) \\
&= \frac{1}{1-\overline{\alpha}_t}\left(\frac{\alpha_t(1-\overline{\alpha}_{t-1}) + \beta_t}{\sqrt{\alpha_t}}x_t - \frac{\beta_t\sqrt{1-\overline{\alpha}_t}\epsilon}{\sqrt{\alpha_t}}\right) \\
&= \frac{1}{1-\overline{\alpha}_t}\left(\frac{1-\overline{\alpha}_t}{\sqrt{\alpha_t}}x_t - \frac{\beta_t\sqrt{1-\overline{\alpha}_t}\epsilon_t}{\sqrt{\alpha_t}}\right) \\
&= \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_t\right)
\end{aligned}
\tag{8}
$$

3.1 学習

DDPM では負の対数尤度を取って最小化する．VAD の導出と基本的に同様．

2 通りの方法で示す．まず 1 つ目は KLD が正であることを利用．

$$\begin{aligned}
-\log p_\theta(x_0) &\leq -\log p_\theta(x_0) + D_{KL}(q(x_{1:T}|x_0)|p_\theta(x_{1:T}|x_0)) \\
&= -\log p_\theta(x_0) + \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T}|x_0)}\right] \\
&= -\log p_\theta(x_0) + \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{q(x_{1:T}|x_0)}{\frac{p_\theta(x_{1:T},x_0)}{p_\theta(x_0)}}\right] \\
&= -\log p_\theta(x_0) + \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{q(x_{1:T}|x_0)}{\frac{p_\theta(x_{0:T})}{p_\theta(x_0)}}\right] \\
&= -\log p_\theta(x_0) + \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})} + \log p_\theta(x_0)\right] \\
&= -\log p_\theta(x_0) + \log p_\theta(x_0) + \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\right] \\
&= \mathcal{L}_{VLB}
\end{aligned} \tag{9}$$

2 つ目は Cross-Entropy loss を考えるところから始めても同じ変分下限にたどり着く．

$$\begin{aligned}
\mathcal{L}_{CE} &= -\mathbb{E}_{q(x_0)} \log p_\theta(x_0) \\
&= -\mathbb{E}_{q(x_0)} \log \left(\int p_\theta(x_{0:T}) dx_{1:T} \right) \\
&\quad \text{天下り的に } q(x_{1:T}|x_0) \text{ を導入} \\
&= -\mathbb{E}_{q(x_0)} \log \left(\int q(x_{1:T}|x_0) \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} dx_{1:T} \right) \\
&= -\mathbb{E}_{q(x_0)} \log \left(\mathbb{E}_{q(x_{1:T}|x_0)} \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right) \\
&\quad \text{イェンセンの不等式より対数と期待値を入れ替え} \\
&\leq -\mathbb{E}_{q(x_0)} \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right] \\
&= -\mathbb{E}_{q(x_{0:T})} \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right] \\
&= \mathbb{E}_{q(x_{0:T})} \left[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\right] \\
&= \mathcal{L}_{VLB}
\end{aligned} \tag{10}$$

ではこの変分下限 $\mathcal{L}_{VLB}$ を展開していく．

$$
\begin{aligned}
\mathcal{L}_{VLB} &= \mathbb{E}_{q(x_{0:T})}\left[\log\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\right] \\
&\quad \text{式 (4) を代入して分母を展開} \\
&= \mathbb{E}_{q(x_{0:T})}\left[\log\frac{\prod_{t=1}^{T} q(x_t|x_{t-1})}{p_\theta(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}\right] \\
&= \mathbb{E}_{q(x_{0:T})}\left[-\log p_\theta(x_T) + \sum_{t=1}^{T}\log\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)}\right] \\
&\quad t=1 \text{ を外に出す} \\
&= \mathbb{E}_{q(x_{0:T})}\left[-\log p_\theta(x_T) + \sum_{t=2}^{T}\log\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)} + \log\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right] \\
&\quad \text{ここで式 (6) の途中式を使い，} q(x_t|x_{t-1}) \text{ 部分を } x_0 \text{ 条件付式に置き換える} \\
&= \mathbb{E}_{q(x_{0:T})}\left[-\log p_\theta(x_T) + \sum_{t=2}^{T}\log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\frac{q(x_t|x_0)}{q(x_{t-1}|x_0)}\right) + \log\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right] \\
&= \mathbb{E}_{q(x_{0:T})}\left[-\log p_\theta(x_T) + \sum_{t=2}^{T}\log\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)} + \sum_{t=2}^{T}\log\frac{q(x_t|x_0)}{q(x_{t-1}|x_0)} + \log\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right] \\
&\quad \text{第 3 項はシグマ展開すると対数の分母分子が打ち消し合い } t=T,2 \text{ のときだけ残る} \\
&= \mathbb{E}_{q(x_{0:T})}\left[-\log p_\theta(x_T) + \sum_{t=2}^{T}\log\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)} + \log\frac{q(x_T|x_0)}{q(x_1|x_0)} + \log\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right] \\
&\quad \text{第 1,3,4 項を整理} \\
&= \mathbb{E}_{q(x_{0:T})}\left[\log\frac{q(x_T|x_0)}{p_\theta(x_T)} + \sum_{t=2}^{T}\log\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)} - \log p_\theta(x_0|x_1)\right] \\
&\quad \text{KL ダイバージェンス形式でまとめる} \\
&= \mathbb{E}_{q(x_{0:T})}\left[\underbrace{D_{KL}(q(x_T|x_0)||p_\theta(x_T))}_{L_T} + \sum_{t=2}^{T}\underbrace{D_{KL}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t))}_{L_{t-1}}\underbrace{-\log p_\theta(x_0|x_1)}_{L_0}\right]
\end{aligned}
\tag{11}
$$

L_T は学習パラメタなしなので無視，L_{t-1} を説明する[*1]．

まず L_{t-1} から式を展開する．式 (5) の $p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1};\mu_\theta(x_t,t),\Sigma_\theta(x_t,t))$ をもう一度思い出す．DDPM では，$\Sigma_\theta(x_t,t)=\sigma_t^2$ と時間依存の定数にして学習パラメタ θ 非依存にする．そうすると，ガウス分布間の KL ダイバージェンスの式を使うと[*2]，μ_θ を含む項だけ考えれば良くなる．

[*1] DDPM では L_0 を独立したガウス分布を仮定．最終的な損失関数にでてこないのでここでは省略．

[*2] `https://github.com/jojonki/AutoEncoders/blob/master/kl_divergence_between_two_gaussians.pdf` 等を見てね

$$
\begin{aligned}
L_{t-1} &= D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) \\
&= \mathbb{E}_{x_0,\epsilon}\left[\frac{1}{2\sigma_t^2}||\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)||^2\right] + C \\
&\quad \text{式 (8) を使って} \\
&= \mathbb{E}_{x_0,\epsilon}\left[\frac{1}{2\sigma_t^2}||\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_t\right) - \mu_\theta(x_t, t)||^2\right] + C \\
&\quad \mu_\theta \text{ が } \tilde{\mu}_t \text{ を推測するようにモデル化できるので，} \mu_\theta \text{ を以下のように設定} \\
&= \mathbb{E}_{x_0,\epsilon}\left[\frac{1}{2\sigma_t^2}||\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_t\right) - \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)\right)||^2\right] + C \\
&\quad \text{項を整理} \\
&= \mathbb{E}_{x_0,\epsilon}\left[\frac{1}{2\sigma_t^2}||\frac{\beta_t}{\sqrt{\alpha_t(1-\overline{\alpha}_t)}}\left(-\epsilon_t + \epsilon_\theta(x_t, t)\right)||^2\right] + C \\
&= \mathbb{E}_{x_0,\epsilon}\left[\frac{\beta_t^2}{2\alpha_t(1-\overline{\alpha}_t)\sigma_t^2}||\epsilon_t - \epsilon_\theta(x_t, t)||^2\right] + C \\
&\quad \text{式 (2) より } x_t \text{ は } x_0 \text{ から一気に表せる} \\
&= \mathbb{E}_{x_0,\epsilon}\left[\frac{\beta_t^2}{2\alpha_t(1-\overline{\alpha}_t)\sigma_t^2}||\epsilon_t - \epsilon_\theta(\sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, t)||^2\right] + C
\end{aligned}
\tag{12}
$$

これは付加したノイズの推定誤差となっている．そのため，対数尤度を最大化するためには，ノイズ付き画像からノイズを推定し，それを最小化すれば良い．

3.2 推論

推論時（サンプリング）は，$x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ から初めて，T 回だけ $p_\theta(x_{t-1}|x_t)$ を繰り返す．式 (5) の $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2)$ をもう一度思い出して reparametrization trick.

$$
\begin{aligned}
x_{t-1} &= \mu_\theta(x_t, t) + \sigma_t z \\
&= \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z
\end{aligned}
\tag{13}
$$

$z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ である．この式は x_t からノイズ $\epsilon_\theta(x_t, t)$ を推測し，取り除く作業になる．

4 まとめ

式がごちゃごちゃするのだが，改めて最初の図 2を見ると良いかもしれない．キーとなるアイデアは，ちょっとずつガウスノイズを足していき（実際には 1 ステップで計算可能），逆拡散過程はその付加したノイズを予測して除去していく，というものであり，処理フローはイメージはしやすいと思う．

5 References

- Denoising Diffusion Probabilistic Models, Ho et al. `https://arxiv.org/abs/2006.11239`

- What are diffusion models?, Lil'Log by Lilian Weng. https://lilianweng.github.io/posts/2021-07-11-diffusion-models/
- 【Deep Learning 研修（発展)】データ生成・変換のための機械学習, nnabla ディープラーニングチャンネル https://www.youtube.com/watch?v=10ki2IS55Q4
- KL Divergence between two gaussians, jojonki https://github.com/jojonki/AutoEncoders/blob/master/kl_divergence_between_two_gaussians.pdf