Denoising Diffusion Probabilistic Models

- Abstract
  - 非平衡熱力学の考えに基づいた潜在変数モデルの一種である拡散モデルを使用して，高品質な画像生成を可能にした
  - 拡散モデルとランジュバン動力学を用いたデノイジングスコアマッチングとの新しい関連性に基づいて設計された重み付き変分境界で学習した
- Introduction
  - 拡散モデルは．変分推論を用いて学習されたマルコフ連鎖に基づく生成モデルであり，有限のステップを経て与えられたデータ分布に一致するサンプルを生成することを目的とする．これは，データに段階的にノイズを加えてその構造を破壊していく過程を逆転させるように学習される．具体的には，拡散過程においては元のデータに対して徐々にノイズが追加され，最終的にはデータの構造が完全に失われるまで進行する．この過程を逆転させるための遷移確率をニューラルネットワークによってパラメータ化し，元のデータを再構成する能力を獲得する．これにより，拡散モデルは複雑なデータ分布を捉えたサンプル生成を可能とする
  - 拡散モデルの特定のパラメータ設定により，トレーニング時に複数のノイズレベルをのせるデノイジングスコアマッチングと，サンプリング時に行うannealed Langevin Dynamicsと同等であることを示した
  - サンプルの品質にもかかわらず，拡散モデルは他の確率ベースのモデルと比較して対数尤度（log likelihood）が劣る
- Background
  - 拡散モデルは，以下のような潜在変数モデルとして形式化される

    $p(\mathbf{x}_0) := \int p_\theta(\mathbf{x}_{0:T})\, d\mathbf{x}_{1:T} \quad \cdots (1)$
  - ここで，$\mathbf{x}_0 \sim q(\mathbf{x}_0)$ は観測データを表し，$\mathbf{x}_1, \ldots, \mathbf{x}_T$は元のデータ $\mathbf{x}_0$と同次元の潜在変数である．これらの潜在変数を介してデータ生成過程をモデル化することにより，拡散モデルは高品質なサンプル生成を実現する．同時分布$p_\theta(\mathbf{x}_{0:T})$は逆拡散過程と呼ばれ，$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$から始まる学習された遷移確率を持つ以下のようなマルコフ連鎖で定義される

    $p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t),$
    $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \quad \cdots (2)$
  - 拡散モデルは，他の種類の潜在変数モデルと異なる点として，事後分布の近似，$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)$（拡散過程）が，分散スケジュール$\beta_1, \ldots, \beta_T$に従ってデータに徐々にガウスノイズを追加する以下のようなマルコフ連鎖であることが挙げられる．

    $q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \quad \cdots (3)$
  - ここで，各$q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$はガウス分布に従い、以下のように定義される．

    $q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}) \quad \cdots (4)$
  - 学習は，負の変分下界の最小化によって実現される．

    $\mathbb{E}[-\log p_\theta(\mathbf{x}_0)] \leq \mathbb{E}q\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] = \mathbb{E}q\left[-\log p(\mathbf{x}_T) - \sum_{t\geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right] := L \quad \cdots (5)$
  - 拡散過程の分散β_tは，再パラメータ化（reparameterization）によって学習することも，ハイパーパラメータとして一意に保つこともできる．拡散過程の注目すべき特性は，任意のタイムステップtでx_tを閉形式でサンプリングできることである．$\alpha_t := 1 - \beta_t$および$\bar{\alpha}_t := \prod_{s=1}^{t}(1-\beta_s)$を用いると，以下が成り立つ．

    $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \quad \cdots (6)$
  - したがって，確率的勾配降下法を用いて L のランダムな項を最適化することで、効率的なトレーニングが可能になる．Lを変形すると，

    - -

$\mathcal{L} = \mathbb{E}q\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right]$
$= \mathbb{E}q\left[-\log p(\mathbf{x}_T) - \sum_{t\geq 1}\log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right]$
$= \mathbb{E}q\left[-\log p(\mathbf{x}_T) - \sum_{t>1}\log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} - \log \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)}\right] \quad \cdots(7)$

マルコフ過程とベイズの定理により，(8)が成り立つので，

$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = q(\mathbf{x}_{t-1}|\mathbf{x}_t) = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_t)}{q(\mathbf{x}_{t-1})} = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{x}_0)q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)} = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)} \quad \cdots(8)$

$\mathcal{L} = \mathbb{E}q\left[-\log p(\mathbf{x}_T) - \sum_{t>1}\log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}\cdot\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} - \log \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)}\right]$
$= \mathbb{E}q\left[-\log p(\mathbf{x}_T) - \sum_{t>1}\log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)} - \log q(\mathbf{x}_1|\mathbf{x}_0) + \log q(\mathbf{x}_T|\mathbf{x}_0) - \log \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)}\right]$
$= \mathbb{E}q\left[-\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)} - \sum_{t>1}\log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)} - \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)\right] \quad \cdots(9)$

$D_{KL}(P \parallel Q) = \mathbb{E}_P\left[\log \frac{P(x)}{Q(x)}\right]$より，

$\mathbb{E}q\left[-\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)}\right] = D_{KL}(q(\mathbf{x}_T|\mathbf{x}_0)\|p(\mathbf{x}_T)),$

$\mathbb{E}q\left[-\sum_{t>1}\log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}\right] = \sum_{t>1}\mathbb{E}q\left[-\log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}\right] =$
$\sum_{t>1} D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\|p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) \quad \cdots(10)$

が成り立つので，

$L =$
$\mathbb{E}q\left[\underbrace{D_{KL}(q(\mathbf{x}_T|\mathbf{x}_0)\|p(\mathbf{x}_T))}_{L_T} + \sum_{t>1}\underbrace{D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\|p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \underbrace{-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}\right] \quad \cdots(11)$

- この式は，KLダイバージェンスを使用して，拡散過程の事後分布と$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$を直接比較する．これらは，以下のように$\mathbf{x}_0$を条件とした場合に扱いやすくなる．

  $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I}) \quad \cdots(12)$

- ここで，$\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t$，$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$である．

- これにより, 式(11)に現れる全てのKLダイバージェンスはガウス分布間の比較に帰着し, 高分散を伴うモンテカルロ推定を用いることなくRao-Blackwell法によって計算することが可能となる. 実際, ガウス分布同士のKLダイバージェンスは以下のように閉形式で計算できるため, モンテカルロ法に起因する高い分散（不確実性）を回避し, KLダイバージェンスをより安定かつ効率的に計算するためにRao-Blackwell法を適用できる.

  $D_{KL}(\mathcal{N}(\mu_0, \Sigma_0) \parallel \mathcal{N}(\mu_1, \Sigma_1)) =$
  $\frac{1}{2}\left(\mathrm{tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_1-\mu_0)^\top\Sigma_1^{-1}(\mu_1-\mu_0) - k + \log\frac{\det\Sigma_1}{\det\Sigma_0}\right) \quad \cdots(13)$

  - 式（13）の証明

    KLダイバージェンスは，2つの確率分布PとQ間で以下のように定義される．

    $D_{KL}(P \parallel Q) = \int p(x)\log\frac{p(x)}{q(x)}\,dx$

    ここで，p(x)とq(x)はそれぞれPとQの確率密度関数である．
    多変量正規分布
    $\mathcal{N}(\mu, \Sigma)$の確率密度関数は以下のように表される．

    $p(x) = \frac{1}{(2\pi)^{k/2}\det(\Sigma)^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)$

    ここで，$P = \mathcal{N}(\mu_0, \Sigma_0)$と$Q = \mathcal{N}(\mu_1, \Sigma_1)$を考える．

    $p(x) = \frac{1}{(2\pi)^{k/2}\det(\Sigma_0)^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_0)^\top\Sigma_0^{-1}(x-\mu_0)\right)$
    $q(x) = \frac{1}{(2\pi)^{k/2}\det(\Sigma_1)^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1)\right)$

    これらをKLダイバージェンスの定義に代入する．

    $D_{KL}(P \parallel Q) = \int p(x)\left[\log p(x) - \log q(x)\right]dx$

まず，それぞれの対数を計算すると，

$\log p(x) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log\det(\Sigma_0) - \frac{1}{2}(x-\mu_0)^\top\Sigma_0^{-1}(x-\mu_0)$
$\log q(x) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log\det(\Sigma_1) - \frac{1}{2}(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1)$

したがって，

$\log p(x) - \log q(x) = \frac{1}{2}\left(\log\frac{\det\Sigma_1}{\det\Sigma_0}\right) + \frac{1}{2}\left[(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^\top\Sigma_0^{-1}(x-\mu_0)\right]$

これをKLダイバージェンスの式に代入すると，

$D_{KL}(P\parallel Q) = \frac{1}{2}\left(\log\frac{\det\Sigma_1}{\det\Sigma_0}\right) +$
$\frac{1}{2}\int p(x)\left[(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^\top\Sigma_0^{-1}(x-\mu_0)\right]dx$

次に積分部分を計算する．

$\int p(x)(x-\mu_0)^\top\Sigma_0^{-1}(x-\mu_0)\,dx = \mathbb{E}_P\left[(x-\mu_0)^\top\Sigma_0^{-1}(x-\mu_0)\right]$
$\int p(x)(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1)\,dx = \mathbb{E}_P\left[(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1)\right]$

まず，$\mathbb{E}_P\left[(x-\mu_0)^\top\Sigma_0^{-1}(x-\mu_0)\right]$を計算する．展開すると，

$(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1) = (x-\mu_0+\mu_0-\mu_1)^\top\Sigma_1^{-1}(x-\mu_0+\mu_0-\mu_1)$
$= (x-\mu_0)^\top\Sigma_1^{-1}(x-\mu_0) + 2(x-\mu_0)^\top\Sigma_1^{-1}(\mu_0-\mu_1) + (\mu_0-\mu_1)^\top\Sigma_1^{-1}(\mu_0-\mu_1)\cdots(*)$

まず(*)の第一項について考える．任意のベクトルvと対称行列Aに対して，以下が成り立つことを考慮すると，

$v^\top Av = \mathrm{tr}(Avv^\top)$

$(x-\mu_0)^\top\Sigma_1^{-1}(x-\mu_0) = \mathrm{tr}\left(\Sigma_1^{-1}(x-\mu_0)(x-\mu_0)^\top\right)$

が成り立つ．また，行列のトレース演算子と期待値演算子は交換可能なので，

$\mathbb{E}_P\left[\mathrm{tr}\left(\Sigma_1^{-1}(x-\mu_0)(x-\mu_0)^\top\right)\right] = \mathrm{tr}\left(\Sigma_1^{-1}\mathbb{E}_P\left[(x-\mu_0)(x-\mu_0)^\top\right]\right)$

ここで，$x\sim\mathcal{N}(\mu_0,\Sigma_0)$であるため、

$\mathbb{E}_P\left[(x-\mu_0)(x-\mu_0)^\top\right] = \Sigma_0$が成り立ち，代入すると，

$\mathrm{tr}\left(\Sigma_1^{-1}\mathbb{E}_P\left[(x-\mu_0)(x-\mu_0)^\top\right]\right) = \mathrm{tr}\left(\Sigma_1^{-1}\Sigma_0\right).$

次に(*)の第二項について考える．

$2\mathbb{E}_P\left[(\mu_0-\mu_1)^\top\Sigma_1^{-1}(x-\mu_0)\right]$

これは，期待値の線形性を利用して，

$2(\mu_0-\mu_1)^\top\Sigma_1^{-1}\mathbb{E}_P\left[x-\mu_0\right]$と書き換えられる．ここで，$\mathbb{E}_P\left[x-\mu_0\right]=0$なので，この項は0になる．

最後に(*)の第三項について考える．

$\mathbb{E}_P\left[(\mu_0-\mu_1)^\top\Sigma_1^{-1}(\mu_0-\mu_1)\right]$

これは定数項であり，期待値を取ってもそのまま値になる．

$= (\mu_0-\mu_1)^\top\Sigma_1^{-1}(\mu_0-\mu_1)$

よって，以下が成り立つ．

$\mathbb{E}_P\left[(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1)\right] = \mathrm{tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_0-\mu_1)^\top\Sigma_1^{-1}(\mu_0-\mu_1)$

次に，$\mathbb{E}_P\left[(x-\mu_0)^\top\Sigma_0^{-1}(x-\mu_0)\right]$を計算する．これは，マハラノビス距離の期待値であり，

$\mathbb{E}_P\left[(x-\mu_0)^\top\Sigma_0^{-1}(x-\mu_0)\right] = \mathrm{tr}(\Sigma_0^{-1}\Sigma_0) = \mathrm{tr}(I_k) = k.$

以上をまとめると，

$D_{KL}(P\parallel Q) = \frac{1}{2}\left(\log\frac{\det\Sigma_1}{\det\Sigma_0}\right) + \frac{1}{2}\left(\mathrm{tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_1-\mu_0)^\top\Sigma_1^{-1}(\mu_1-\mu_0) - k\right).$

これを整理すると，

$$D_{KL}(\mathcal{N}(\mu_0, \Sigma_0) \parallel \mathcal{N}(\mu_1, \Sigma_1)) = \frac{1}{2}\left(\mathrm{tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^\top \Sigma_1^{-1}(\mu_1 - \mu_0) - k + \log \frac{\det \Sigma_1}{\det \Sigma_0}\right)$$

となり，式(13)に一致する．

- 拡散モデルは実装上，多くの自由度を設定できる．拡散過程の分散β_tの選択や，逆拡散過程のモデルアーキテクチャおよびガウス分布のパラメータ化を決定する必要がある．この自由度の選択に関して以下に説明する．
- まず，DDPMでは，β_tを定数に固定している．従って，事後分布qに学習可能なパラメータが存在しないため，L_Tは訓練中に一定の値に保たれ，無視することができる．
- 次に，$1 < t \leq T$における$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t; t), \Sigma_\theta(\mathbf{x}_t; t))$の選択について議論する．まず，$\Sigma_\theta(\mathbf{x}_t, t) = \sigma_t^2 \mathbf{I}$を未学習の時間に依存する定数として設定する．次に，$\mu_\theta(\mathbf{x}_t, t)$を考える．ここで以下のような再パラメータ化を考える

  $D_{KL}(\mathcal{N}(\mu_0, \Sigma_0) \parallel \mathcal{N}(\mu_1, \Sigma_1)) =$
  $\frac{1}{2}\left(\mathrm{tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^\top \Sigma_1^{-1}(\mu_1 - \mu_0) - k + \log \frac{\det \Sigma_1}{\det \Sigma_0}\right) \quad \cdots (14)$

  と，

  $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}),$
  $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t; t), \sigma_t^2 \mathbf{I}) \quad \cdots (15)$

  より，

  $D_{KL}\left(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\right) = \frac{1}{2}\left(\log \frac{\sigma_t^2}{\tilde{\beta}_t} - k + \frac{\tilde{\beta}_t}{\sigma_t^2} \cdot k\right) +$
  $\frac{1}{2\sigma_t^2} \left\| \boldsymbol{\mu}_\theta(\mathbf{x}_t, t) - \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) \right\|^2 \quad \cdots (16)$

  なので，以下が成り立つ（Cは定数）

  $\mathcal{L}_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} ||\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, t)||^2 \right] + C \quad \cdots (17)$

  $\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon},$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$と変換すると，

  $\mathcal{L}_{t-1} - C = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}} \left[\frac{1}{2\sigma_t^2} \| \tilde{\mu}_t((\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}), \frac{1}{\sqrt{\bar{\alpha}}}(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}) - \mu_\theta(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}), t) \|^2 \right] \quad \cdots (18)$こ
  こで，$\tilde{\mu}_t((\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}), \frac{1}{\sqrt{\bar{\alpha}}}(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon})$について考える．$\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t}\mathbf{x}_0 +$
  $\frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}\mathbf{x}_t$なので，代入して，

  $\tilde{\mu}_t((\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}), \frac{1}{\sqrt{\bar{\alpha}}}(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{\sqrt{\bar{\alpha}_t}(1 - \bar{\alpha}_t)}(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}) +$
  $\frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) \quad \cdots (19)$

  $\bar{\alpha}_{t-1} = \frac{\bar{\alpha}_t}{\alpha_t}, \beta_t = 1 - \alpha_t$なので，

  $= \frac{\sqrt{\frac{\bar{\alpha}_t}{\alpha_t}}(1 - \alpha_t) + \sqrt{\alpha_t \bar{\alpha}_t}(1 - \frac{\bar{\alpha}_t}{\alpha_t})}{(1 - \bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \frac{\sqrt{\frac{\bar{\alpha}_t}{\alpha_t}}\beta_t\boldsymbol{\epsilon}}{\sqrt{\bar{\alpha}_t(1 - \bar{\alpha}_t)}} \quad \cdots (20)$
  $= \frac{\sqrt{\bar{\alpha}_t}(1 - \alpha_t) + \sqrt{\bar{\alpha}_t}(\alpha_t - \bar{\alpha}_t)}{(1 - \alpha_t)\sqrt{\alpha_t \bar{\alpha}_t}}\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \frac{\beta_t\boldsymbol{\epsilon}}{\sqrt{\alpha_t(1 - \bar{\alpha}_t)}} \quad \cdots (21)$
  $= \frac{1}{\sqrt{\alpha_t}}\left[(\frac{1 - \alpha_t + \alpha_t - \bar{\alpha}_t}{1 - \bar{\alpha}_t})\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\boldsymbol{\epsilon} \right] \quad \cdots (22)$
  $= \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\boldsymbol{\epsilon}) \quad \cdots (23)$

  が成り立つ．よって，

  $\mathcal{L}_{t-1} - C = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}} \left[\frac{1}{2\sigma_t^2} \| \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\boldsymbol{\epsilon}) - \mu_\theta(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}), t) \|^2 \right]. \quad \cdots (24)$

  ここで以下のように再パラメータ化を行う．

  $\mu_\theta(\mathbf{x}_t, t) = \tilde{\mu}_t(\mathbf{x}_t, \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t))) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)) \quad \cdots (25)$

  $\boldsymbol{\epsilon}_\theta$は，$\mathbf{x}_t$から$\boldsymbol{\epsilon}$を予測する関数である．このパラメータ化によって，式(18)は以下のように変形できる

  $= \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}}[\frac{1}{2\sigma_t^2} \parallel \frac{1}{\sqrt{\alpha_t}}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon} - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\boldsymbol{\epsilon})$

$$-\frac{1}{\sqrt{\alpha_t}}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}-\frac{\beta_t}{\sqrt{1-\bar{\alpha}t}}\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t))\parallel^2] \quad \cdots(26)$$

$$=\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}t)}\parallel\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t)\parallel^2\right] \quad \cdots(27)$$

- これは複数のノイズスケールに渡るノイズ除去のスコアマッチングに類似している，(27)は，ランジュバン仕様の逆拡散過程の変分下界の一つの項と等しいため，ノイズ除去スコアマッチングに類似した目的関数を最適化することは，ランジュバンダイナミクスに類似した有限時間のマージンに適応させるために変分推論を使用することと同等である．まとめると，逆拡散過程でμ_θによって$\tilde{\mu}_t$を予測するように訓練することもできる上，そののパラメータ化を変更することで$\boldsymbol{\epsilon}$を予測するように訓練することも可能である．$\boldsymbol{\epsilon}$のパラメータ化は，ランジュバンダイナミクスに類似しているだけでなく，拡散モデルの変分境界をノイズ除去スコアマッチングに類似した目的関数に単純化できることもできる．
- 画像データが$\{0,1,\ldots,255\}$の整数から成り，それが[-1, 1]に線形にスケールされていると仮定する．これは，ニューラルネットワークの逆拡散過程が標準正規分布の事前分布$p(x_T)$から始まる一貫してスケールされた入力で動作することを保証する．離散対数尤度を得るために、逆拡散過程の最後の項をガウス分布$\mathcal{N}(\mathbf{x}_0;\mu_\theta(\mathbf{x}_1,1),\sigma_1^2\mathbf{I})$から導かれる独立な離散デコーダに設定する

  $$p_\theta(\mathbf{x}_0\mid\mathbf{x}_1)=\prod_{i=1}^D\int_{\delta_-(x_0^i)}^{\delta_+(x_0^i)}\mathcal{N}(x;\mu_\theta^i(\mathbf{x}_1,1),\sigma_1^2)\,dx \quad \cdots(28)$$

  ここで

  $$\delta_+(x)=\begin{cases}\infty & \text{if} \quad x=1\\ x+\frac{1}{255} & \text{if} \quad x<1\end{cases} \quad \delta_-(x)=\begin{cases}-\infty & \text{if} \quad x=-1\\ x-\frac{1}{255} & \text{if} \quad x>-1.\end{cases}$$

  Dはデータの次元数を表し，iの上付き文字は1つの座標の抽出を示す．VAEデコーダや自己回帰モデルで用いられる離散化された連続分布と類似しており，ここでの選択は変分境界が離散データの可逆符号長となることを保証する．データにノイズを加えたり，スケーリング操作のヤコビアンを対数尤度に組み込む必要がない．サンプリングの最後に，$\mu_\theta(\mathbf{x}_1,1)$をノイズなしで表示できる．
- 逆拡散過程とデコーダを前述のように定義すると，式（27）および式（28）から導出される項からなる変分境界は微分可能である．サンプルの品質を向上させるため，以下の変分境界を用いて訓練することが有益であることがわかった．

  $$\mathcal{L}_{\text{simple}}(\theta):=\mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\parallel\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t)\parallel^2\right] \quad \cdots(29)$$

  ここで，tは1からTまでの一様な値である．t = 1の場合はL_0に対応し，離散デコーダの定義（式（28））における積分は，正規分布にビン幅を掛けたもので近似される．t > 1の場合は，式（27）の重み付けされていないバージョンに対応し，NCSN（デノイジングスコアマッチングモデル）で使用される損失の重み付けに類似している．