Understanding Diffusion Models: A Unified Perspective

NTTコミュニケーションズ 藤原大悟

© NTT Communications Corporation All Rights Reserved.

本日のモチベーション

- 個人的に後回しにしていた **Diffusion Model** について
  真面目に理解する
  - ざっくりノイズを足すとか拡散過程を使ってることは知っているけど正確な定式化とか知らない
    - 数式が多いですがお気持ちを説明するようにします
  - なんで上手くいくのか直感的に理解できてない
  - 色んな生成モデルを復習して関係性を理解したい
- ちょっと業務逼迫で急ピッチで読んだ&作ったので粗いところがあったらごめんなさい

© NTT Communications Corporation All Rights Reserved.

生成モデルとは

- 生成モデルの大まかな目的
  - データ $D = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)^{\mathrm{T}}$ が与えられた時に、データ $\boldsymbol{x}_i$ を生み出す分布関数 $p(\boldsymbol{x})$ をモデリングし、そこからのサンプリングを行うこと

- 生成モデルの系譜 (知ってる人向けの説明)
  - Generative Adversarial Networks (GANs)
    - みなさんご存知Diffusion登場以前の最強モデル。Discriminator(識別器)とGenerator(生成器)を戦わせて精度アップ。
    - 背後には密度比推定とかの理論が
  - ”Likelihood-Based” (原文ママ)
    - パラメトリックな確率モデル p_θ を考えて、データに対する尤度を最大化してフィッティング
    - Autoregressive Models
      - (入出力を繰り返し適用すればデータ生成できるからそういう意味で生成モデルとして挙げられている？)
    - Normalizing Flows
      - 正規ノイズにNNなどによる非線形変換を何回もかましてデータ分布に近づけようみたいなやつ
      - カップリングとかで逆変換可能な非線形変換を作って、尤度Lossを計算可能にする工夫。変分推論を利用。
    - Variational Auto Encoders (VAEs)
      - Auto Encoder構造を基礎に、中間層(低次元)は正規ノイズから生成されるものとして尤度ベースのLossで学習。
      - Encoder $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ とDecoder $p_\phi(\boldsymbol{x}|\boldsymbol{z})$ を同時に学習。変分推論の利用。

© NTT Communications Corporation All Rights Reserved.

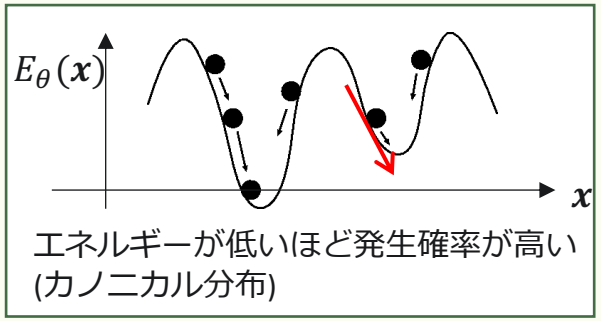
生成モデル(エネルギーモデル,スコアベースモデル)

- 生成モデルの系譜 (知ってる人向けの説明)
  - ”Likelihood-Based” (原文ママ)
    - パラメトリックな確率モデルを考えて、データに対する尤度を最大化してフィッティング
      - Energy Based Models
        - よく統計力学とかでありそうな分布形、$p(\boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{\theta}}}\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))$を仮定して尤度Lossで$E_{\boldsymbol{\theta}}(\boldsymbol{x})$を学習
        - 統計力学の知見を流用しつつ、尤度に登場する期待値(規格化定数 $Z_{\boldsymbol{\theta}}$)計算を近似とサンプリングで上手く回避
          - $Z_{\boldsymbol{\theta}}$を求めず分布 $p(\boldsymbol{x})$ からサンプリングする方法 (MCMC+ランジュバンダイナミクス)
            - $\boldsymbol{x}^{t+1} = \boldsymbol{x}^{t} + \eta\, \nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{\omega}$　($\boldsymbol{\omega}$: ガウスノイズ, η: ハイパーパラメタ)
          - ※ スコア $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log(\frac{1}{Z_{\boldsymbol{\theta}}}\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))) = \nabla_{\boldsymbol{x}} \{\log(\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x})) - \log(\mathrm{Z}_{\boldsymbol{\theta}})\} = -\nabla_{\boldsymbol{x}} E_{\boldsymbol{\theta}}(\boldsymbol{x})$
      - Score Based Models
        - 発想：$E_{\boldsymbol{\theta}}(\boldsymbol{x})$ を介さずに直接 $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})$ だけ学習したらいいのでは？？
          - スコアマッチング：
            - データからスコアをフィッティングするLossを計算するための工夫(変形)

Fisher Divergence の最小化問題 →

$$\mathrm{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\frac{1}{2}\left|\left|\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x}) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\right|\right|_2^2\right] = \mathrm{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\frac{1}{2}\left|\left|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\right|\right|_2^2 + \mathrm{tr}\left(\nabla_{\boldsymbol{x}}\, \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right]$$

ここがわからなくてもOK



エネルギーが低いほど発生確率が高い
(カノニカル分布)

筆者曰く、Diffusion Modelは尤度ベース、スコアベースどちらとしても解釈可能とのこと (楽しみ)

© NTT Communications Corporation All Rights Reserved.

変分推論ってなんだっけ？

前提
データ$\boldsymbol{x}$に対して、その根本構造を有する(低次元)潜在変数 $\boldsymbol{z}$ の存在を考えるとうまく行くことが多い。
(MNIST画像 $\boldsymbol{x}$ に対して数字や手癖 $\boldsymbol{z}$ の存在を考えるのは自然)

ということでその関係性 $p(\boldsymbol{x}, \boldsymbol{z})$ に興味があり、モデル化したい。
(※実際は多くの場合 $p(\boldsymbol{z}|\boldsymbol{x})$ をモデル化する)

そこで、我々は真の分布 $p(\boldsymbol{z}|\boldsymbol{x})$ に対して 適当な提案分布のクラス $q(\boldsymbol{z}|\boldsymbol{x})(= q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}))$ で近似を図る

で、天下り的だが右が成り立つ

KL divergence 最小化をELBOを介して行うことができる

$$
\begin{aligned}
\log p(\boldsymbol{x}) &= \log p(\boldsymbol{x}) \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} && \text{(Multiply by } 1 = \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z}\text{)} && (9)\\
&= \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) (\log p(\boldsymbol{x})) d\boldsymbol{z} && \text{(Bring evidence into integral)} && (10)\\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} [\log p(\boldsymbol{x})] && \text{(Definition of Expectation)} && (11)\\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})} \right] && \text{(Apply Equation 2)} && (12)\\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}) q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z}|\boldsymbol{x}) q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \right] && \text{(Multiply by } 1 = \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\text{)} && (13)\\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \right] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z}|\boldsymbol{x})} \right] && \text{(Split the Expectation)} && (14)\\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \right] + D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z}|\boldsymbol{x})) && \text{(Definition of KL Divergence)} && (15)\\
&\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \right] && \text{(KL Divergence always} \geq 0\text{)} && (16)
\end{aligned}
$$

定数
(for ϕ)

ここをでかくすればOK
=ELBO

ゼロにしたら近似成功

© NTT Communications Corporation All Rights Reserved.

変分推論ってなんだっけ？

ちなみに

VAEなんかでよくやる手は、 $p(\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{x})$ (や、 $p(\boldsymbol{x}|\boldsymbol{z}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$) とおいて、$\boldsymbol{\theta}$に対する尤度の最大化もいっぺんにやってしまう (※EMアルゴリズムっぽい)

とにかくELBOをでかくすればKLも小さくなって尤度($\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$)もでかくなるだろうと

※EMアルゴリズムはKL divergence最小化と尤度最大化を交互に行う

$\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$

定数
(for $\boldsymbol{\phi}$)

$$
\begin{aligned}
\log p(\boldsymbol{x}) &= \log p(\boldsymbol{x}) \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} && \left(\text{Multiply by } 1 = \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z}\right) && (9)\\
&= \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) (\log p(\boldsymbol{x})) d\boldsymbol{z} && \text{(Bring evidence into integral)} && (10)\\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} [\log p(\boldsymbol{x})] && \text{(Definition of Expectation)} && (11)\\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})}\right] && \text{(Apply Equation 2)} && (12)\\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}) q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z}|\boldsymbol{x}) q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] && \left(\text{Multiply by } 1 = \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right) && (13)\\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z}|\boldsymbol{x})}\right] && \text{(Split the Expectation)} && (14)\\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] + D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z}|\boldsymbol{x})) && \text{(Definition of KL Divergence)} && (15)\\
&\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] && \text{(KL Divergence always } \geq 0) && (16)
\end{aligned}
$$

ここをでかくすればOK
=ELBO

ゼロにしたら近似成功

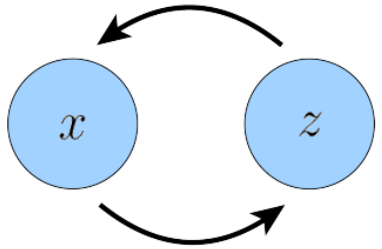
© NTT Communications Corporation All Rights Reserved.

VAE → Hierarchical VAE

例えば白黒画像ならベルヌーイ分布

$p(x|z) = \mathrm{Bern}(\boldsymbol{x} \mid \mu_{\boldsymbol{\theta}}(\boldsymbol{z}))$とか

VAE



$\sim \mathcal{N}(\boldsymbol{z} \mid 0, \mathbf{I})$

$q(z|x) = \mathcal{N}(\boldsymbol{z} \mid \mu_{\boldsymbol{\phi}}(\boldsymbol{x}), \sigma^2_{\boldsymbol{\phi}}(\boldsymbol{x})\mathbf{I})$

Figure 1: A Variational Autoencoder graphically represented. Here, encoder $q(\boldsymbol{z}|\boldsymbol{x})$ defines a distribution over latent variables $\boldsymbol{z}$ for observations $\boldsymbol{x}$, and $p(\boldsymbol{x}|\boldsymbol{z})$ decodes latent variables into observations.

Loss (ELBO)

$$\begin{aligned}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] \\ &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] \\ &= \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \,\|\, p(\boldsymbol{z}))}_{\text{prior matching term}}\end{aligned}$$

階層化しただけ

HVAE

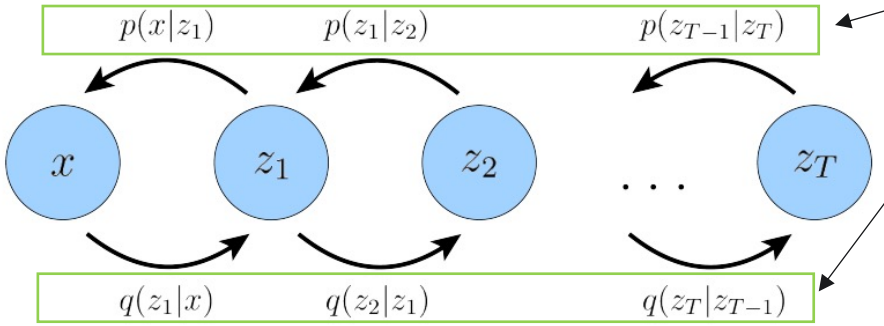


Figure 2: A Markovian Hierarchical Variational Autoencoder with T hierarchical latents. The generative process is modeled as a Markov chain, where each latent $\boldsymbol{z}_t$ is generated only from the previous latent $\boldsymbol{z}_{t+1}$.

$$p(\boldsymbol{x}, \boldsymbol{z}_{1:T}) = p(\boldsymbol{z}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)$$

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x}) = q_{\boldsymbol{\phi}}(\boldsymbol{z}_1|\boldsymbol{x})\prod_{t=2}^{T} q_{\boldsymbol{\phi}}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$$

Loss (ELBO)

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z}_{1:T})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x})}\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{z}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)}{q_{\boldsymbol{\phi}}(\boldsymbol{z}_1|\boldsymbol{x})\prod_{t=2}^{T} q_{\boldsymbol{\phi}}(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})}\right]$$

© NTT Communications Corporation All Rights Reserved.

Hierarchical VAE → Diffusion Models

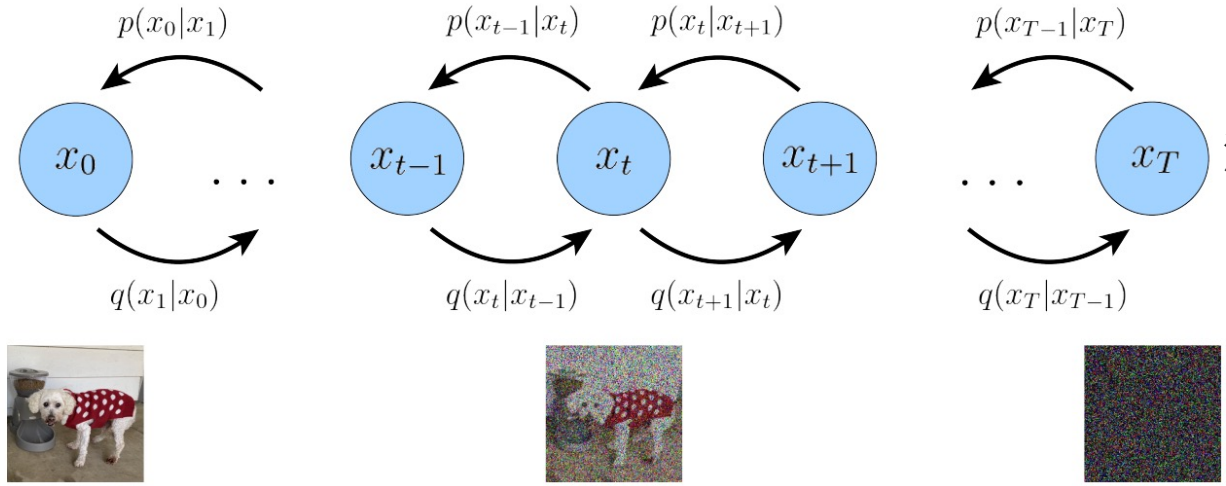
Variational Diffusion Models
この文脈でのDiffusion Modelsは以下の制約を持つHVAEと解釈できる

- 潜在次元はデータ次元と等しい。
- 各タイムステップにおけるエンコーダー$q(\boldsymbol{z}|\boldsymbol{x})$ は学習されない。
  前のタイムステップの出力を中心とする**線形ガウス分布モデルとして(ハイパラが与えられ)定義されている。**
- 潜在変数は、**最終タイムステップTにおける分布が標準ガウス分布となるように、各ステップで時間発展する**
  - (Tが十分大きければ最終ステップの分布は勝手に標準ガウス分布になる)

- 前後で分散が同程度になるよう調整されている
- ハイパラα_tはtによって異なる(学習してもいい)

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})$$

データ空間サイド



潜在空間サイド

Figure 3: A visual representation of a Variational Diffusion Model; $\boldsymbol{x}_0$ represents true data observations such as natural images, $\boldsymbol{x}_T$ represents pure Gaussian noise, and $\boldsymbol{x}_t$ is an intermediate noisy version of $\boldsymbol{x}_0$. Each $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ is modeled as a Gaussian distribution that uses the output of the previous state as its mean.

© NTT Communications Corporation All Ri

Diffusion Models のナイーブなLoss

Variational Diffusion Models

(H)VAEと同様にELBOの最大化に落とせる

ここをでかくすればOK
=ELBO

まあ変形は置いといて最後の方を見ると以下の3つに分解できる。

- 再構成誤差
  - データとの直接のDecoder Loss
- 事前分布のKL divergence
  - 学習可能なパラメータがないため実質定数
  - Tが十分大きければ0に
- **EncoderとDecoder過程の一致を要求する項**
  - Decoderはθが学習可能なので、Encoder (=与えられたガウス分布拡散)に合わせなければいけない

$$
\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} \\
&= \log \int \frac{p(\boldsymbol{x}_{0:T}) q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} d\boldsymbol{x}_{1:T} \\
&= \log \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] \\
&\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \prod_{t=1}^{T-1} p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})} \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \prod_{t=1}^{T-1} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})} \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\sum_{t=1}^{T-1} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_{T-1},\boldsymbol{x}_T|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_t,\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \\
&= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_{\theta}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right]}_{\text{reconstruction term}} + \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \parallel p(\boldsymbol{x}_T)) \right]}_{\text{prior matching term}} \\
&\quad + \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \parallel p_{\theta}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})) \right]}_{\text{consistency term}}
\end{aligned}
$$

参考：VAEのELBO

$$
\begin{aligned}
\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \right] &= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \right] \\
&= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) \right] + \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \right] \\
&= \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) \right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z}))}_{\text{prior matching term}}
\end{aligned}
$$

© NTT Communications Corporation All Rights Reserved.

Diffusion Models の ナイーブなLoss

Variational Diffusion Models

$$-\sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \parallel p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}))\right]}_{\text{consistency term}}$$

補足

- **EncoderとDecoder過程の一致を要求する項**
  - Decoderは $\boldsymbol{\theta}$ が学習可能なので、Encoder (=与えられたガウス分布拡散)に合わせなければいけない
  - x_tを中心に両側から挟む感じ(ピンクと緑の矢印)
- **悪い点**
  - 実際のロスの計算は期待値のモンテカルロ推定が必要だが、第3項については2変数$(\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t+1})$で期待値を取るので、**推定値の誤差(分散)が大きくなってしまう**

→修正

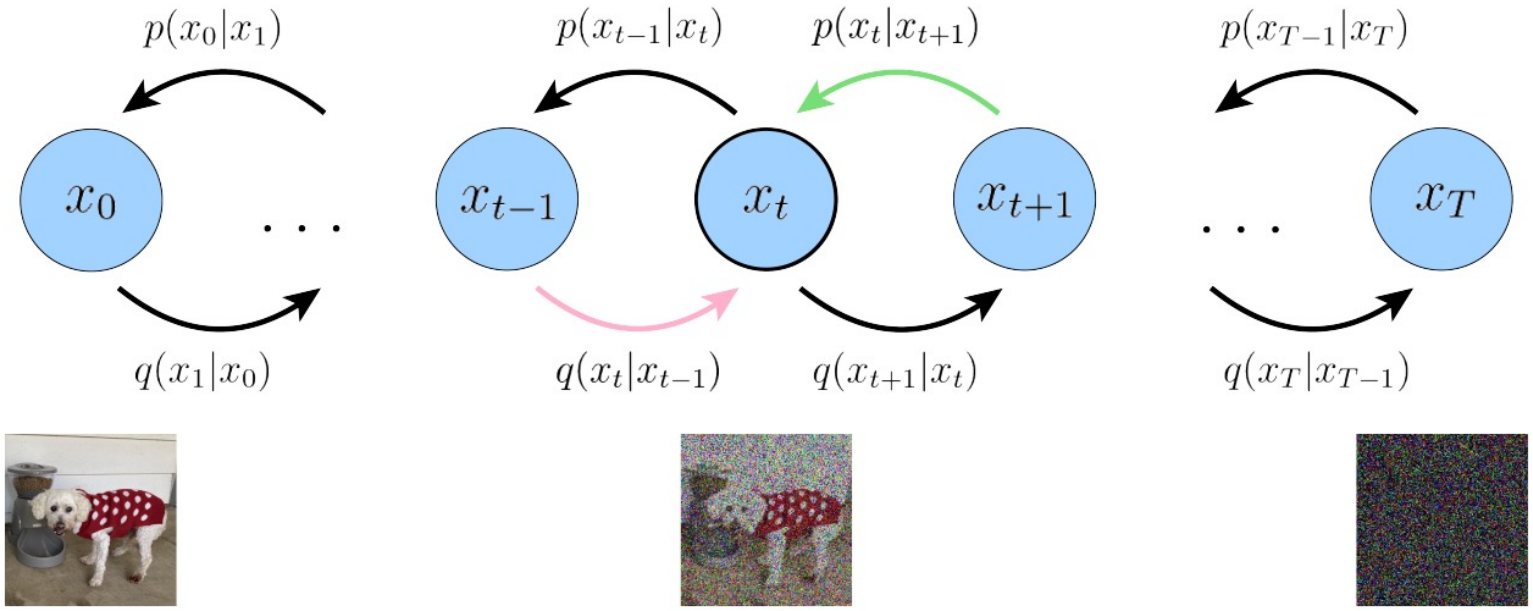


Figure 4: Under our first derivation, a VDM can be optimized by ensuring that for every intermediate $\boldsymbol{x}_t$, the posterior from the latent above it $p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})$ matches the Gaussian corruption of the latent before it $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$. In this figure, for each intermediate $\boldsymbol{x}_t$, we minimize the difference between the distributions represented by the pink and green arrows.

© NTT Communications Corporation All Rights Reserved.

Diffusion Models の 修正したLoss

計算が
めちゃめちゃ長いのでざっくり説明 (次ページ)

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

同じELBOからスタート

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)\prod_{t=2}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)\prod_{t=2}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0)}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log\prod_{t=2}^{T}\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0)}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log\prod_{t=2}^{T}\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{\cancel{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}} + \log \frac{\cancel{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} + \log\prod_{t=2}^{T}\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} + \sum_{t=2}^{T}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log\frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\right] + \sum_{t=2}^{T}\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] + \mathbb{E}_{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\left[\log\frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\right] + \sum_{t=2}^{T}\mathbb{E}_{q(\boldsymbol{x}_t,\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}\left[\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}\right]$$

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\parallel p(\boldsymbol{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^{T}\underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

同じ　　ほぼ同じ　　新しいLoss

© NTT Communications Corporation All Rights Reserved.

Diffusion Models の 修正したLoss

Variational Diffusion Models

$$-\sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

新しいLoss (時間が揃ってて嬉しい)

$$-\sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \parallel p_{\theta}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}))\right]}_{\text{consistency term}}$$

元のLoss (時間がずれててうざい)

- **EncoderとDecoder過程の一致を要求する項を修正**
- **基本発想：**
  - **ベイズの定理使えば $q(x_t|x_{t-1})$ を逆転して$q(x_{t-1}|x_t)$みたいな形に持っていけるんじゃね？**
    ↓
  - ピンクと緑の矢印の場所が揃った！！
- **良い点**
  - **期待値推定がx_tについてのみになり精度向上**

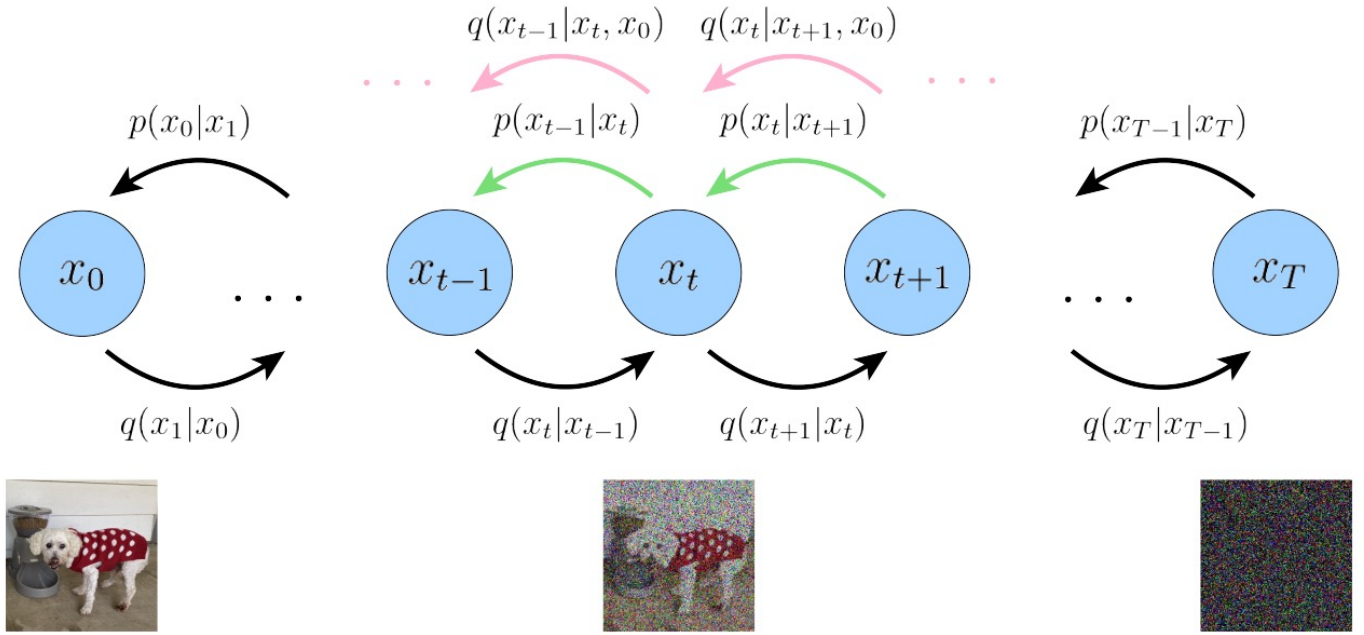


Figure 5: Depicted is an alternate, lower-variance method to optimize a VDM; we compute the form of ground-truth denoising step $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ using Bayes rule, and minimize its KL Divergence with our approximate denoising step $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$. This is once again denoted visually by matching the distributions represented by the green arrows with those of the pink arrows. Artistic liberty is at play here; in the full picture, each pink arrow must also stem from $\boldsymbol{x}_0$, as it is also a conditioning term.

© NTT Communications Corporation All Rights Reserved.

Diffusion Models の 修正したLoss

Variational Diffusion Models

$$-\sum_{t=2}^{T}\underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

新しいLoss (時間が揃ってて嬉しい)

$$-\sum_{t=1}^{T-1}\underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\parallel p_{\theta}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}))\right]}_{\text{consistency term}}$$

元のLoss (時間がずれててうざい)

- EncoderとDecoder過程の一致を要求する項を修正
- 基本発想：
  - ベイズの定理使えば $q(x_t|x_{t-1})$ を逆転して$q(x_{t-1}|x_t)$みたいな形に持っていけるんじゃね？
    ↓
  - ピンクと緑の矢印の場所が揃った！！
- 良い点
  - 期待値推定がx_tについてのみになり精度向上

ポイント！！！

- $q(x_{t-1}|x_t,x_0)$が解析的に計算できるのは今回だけ！
  - なぜなら単純な線形ガウス変換(拡散過程)を考えていたから
  - 一般のHVAEでは計算困難
- $q(x_{t-1}|x_t,x_0)=\dfrac{q(x_t\,|\,x_{t-1},x_0)\,q(x_{t-1}|\,x_0)}{q(x_t\,|\,x_0)}$

$q(x_t\,|\,x_{t-1},)$ 拡散過程(既知)

拡散過程

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})=\mathcal{N}(\boldsymbol{x}_t;\sqrt{\alpha_t}\boldsymbol{x}_{t-1},(1-\alpha_t)\mathbf{I})$$

↓等式で書き下すとこう

$$\boldsymbol{x}_t=\sqrt{\alpha_t}\boldsymbol{x}_{t-1}+\sqrt{1-\alpha_t}\boldsymbol{\epsilon}\quad\text{with }\boldsymbol{\epsilon}\sim\mathcal{N}(\boldsymbol{\epsilon};\mathbf{0},\mathbf{I})$$

- $q(x_t\,|\,x_0)$について、
  - x_t に再起的に繰り返し代入すると、

$$\boldsymbol{x}_t=\sqrt{\prod_{i=1}^{t}\alpha_i}\,\boldsymbol{x}_0+\sqrt{1-\prod_{i=1}^{t}\alpha_i}\,\boldsymbol{\epsilon}_0=\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0$$

$$\sim q(\boldsymbol{x}_t\,|\,\boldsymbol{x}_0)=\mathcal{N}(\boldsymbol{x}_t;\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0,(1-\bar{\alpha}_t)\mathbf{I})$$

ということで計算できる(代入するだけ)

© NTT Communications Corporation All Rights Reserved.

Diffusion Models による画像生成(元画像推定)

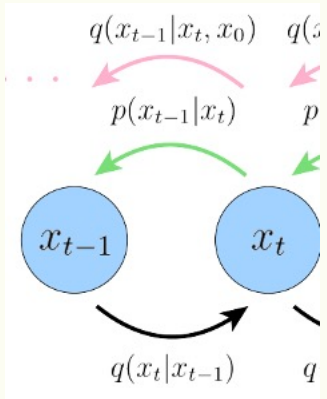
Variational Diffusion Models

先の結果を$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$に代入して具体的に計算すると、

$$\propto \mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)})$$

$$-\sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

新しいLoss (時間が揃ってて嬉しい)



のようにガウス分布になる。(ガウス分布×ガウス分布÷ガウス分布なので自明)

==========

さて我々は$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$にどんな分布を持ってきても良いわけだが、形が分かってるので、**ガウス分布を仮定してみよう。しかも分散は既知(α_tから計算可)なので、正解の分散を教えてあげることにしよう。**

真の平均がどんなだったかというと、**こんな**感じ。
では、ここも形が分かってるので真似して、**こういう形**を仮定しよう。

Loss, KL最小化は二乗誤差 $\left|\left|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{x}_0\right|\right|^2$の最小化に帰着する。

$\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$は元画像$\boldsymbol{x}_0$の推定器になる！！！

$$\boldsymbol{\mu}_q(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}$$

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)}{1-\bar{\alpha}_t}$$

NNで与えてあげる

© NTT Communications Corporation All Rights Reserved.

Diffusion Models による画像生成(元画像推定)

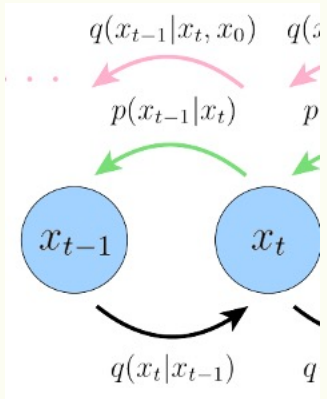
Variational Diffusion Models

先の結果を$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$に代入して具体的に計算すると、

$$\propto \mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)})$$

$$-\sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

新しいLoss (時間が揃ってて嬉しい)



のようにガウス分布になる。(ガウス分布×ガウス分布÷ガウス分布なので自明)

==========

さて我々は$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$にどんな分布を持ってきても良いわけだが、形が分かってるので、**ガウス分布を仮定してみよう。しかも分散は既知(α_tから計算可)なので、正解の分散を教えてあげることにしよう。**

真の平均がどんなだったかというと、**こんな**感じ。

では、ここも形が分かってるので真似して、**こういう形**を仮定しよう。

Loss, KL最小化は二乗誤差$\left|\left|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{x}_0\right|\right|^2$の最小化に帰着する。

$\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$は元画像$\boldsymbol{x}_0$の推定器になる！！！

$$\boldsymbol{\mu}_q(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}$$

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)}{1-\bar{\alpha}_t}$$

NNで与えてあげる

素朴な疑問：この結果がどこまで見えていて階層的な拡散過程考えたんだろう、、、

© NTT Communications Corporation A

雑多な感想

- 解析的に逆拡散過程 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ が計算できるため、好きなパラメトライズを選択できる $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ に分かってる情報を全て組み込むと、必然的に二つの分布の間で足りない情報(=$\boldsymbol{x}_0$)を推定する形になるのかもしれないと思った

- (ノイズ込みで)分布をモデリングして、そこからサンプリングするといった生成方法ではなく、あくまで「元画像の推定」のテイなので、めちゃくちゃ解像度の高い画像を出力できるのかなと思った

- ただ原理的に「元画像を推定する」点で、Training Dataを再現する傾向が強うそうで汎化性能に疑問が残るので、AI絵師、著作権関連で言われる「既存画像のパクリを出力してるだけ」説は結構正しそうな気がしてしまった

© NTT Communications Corporation All Rights Reserved.

ハイパラ(α_t)の学習

ここでは割愛

- α_tではなく$\bar{\alpha}_t$をNNでモデリングすること
  - 結局計算には$\bar{\alpha}_t$だけ使えば十分だし、全ての時間tで掛け合わせないといけないので非効率

- 時間と共に単調減少する関数としてモデリングする必要があること
  - 最終的にノイズに終着してもらう必要があるため。

がポイント

© NTT Communications Corporation All Rights Reserved.

付加ノイズ推定としての解釈

ここでは割愛。ざっくり言うと、

- 元画像そのものを推定する問題として定式化したが、等価な方法として、tステップ時点で元画像に加わった合計ノイズ、を推定する方法もある。(簡単な式変形で求まる)

- なぜかノイズを学習した方が経験的に精度が良いと言う話もあるらしい
  - 等価なのになぜ。。。？

© NTT Communications Corporation All Rights Reserved.

スコア推定としての解釈

- **Tweedieの公式**というものを使う

  指数関数型分布の真の平均は、その分布から得られた標本が与えられたとき、標本による最尤推定値（別名：経験平均）に推定値のスコアを含む何らかの補正項を加えたもので推定できることを述べている。

ガウス分布の標本が一つある場合：

Mathematically, for a Gaussian variable $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$, Tweedie's Formula states that:

$$\mathbb{E}\left[\boldsymbol{\mu}_z | \boldsymbol{z}\right] = \boldsymbol{z} + \boldsymbol{\Sigma}_z \nabla_{\boldsymbol{z}} \log p(\boldsymbol{z})$$

さっきまでの結果に流用

真の分布 $q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\,\mathbf{I})$

Tweedieの公式を適用
(x_tから$\sqrt{\bar{\alpha}_t}x_0$を推定するテイ)

$$\mathbb{E}\left[\boldsymbol{\mu}_{x_t} | \boldsymbol{x}_t\right] = \boldsymbol{x}_t + (1-\bar{\alpha}_t)\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$$

$$\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 = \boldsymbol{x}_t + (1-\bar{\alpha}_t)\nabla \log p(\boldsymbol{x}_t)$$

$$\therefore \boldsymbol{x}_0 = \frac{\boldsymbol{x}_t + (1-\bar{\alpha}_t)\nabla \log p(\boldsymbol{x}_t)}{\sqrt{\bar{\alpha}_t}}$$

© NTT Communications Corporation All Rights Reserved.

スコア推定としての解釈

前ページ
Tweedieの公式より
以下に代入

$$\therefore \boldsymbol{x}_0 = \frac{\boldsymbol{x}_t + (1-\bar{\alpha}_t)\nabla \log p(\boldsymbol{x}_t)}{\sqrt{\bar{\alpha}_t}}$$

$$-\sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

Diffusion ModelのLoss (時間が揃ってて嬉しい)

$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$の平均

$$\boldsymbol{\mu}_q(\boldsymbol{x}_t,\boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}$$

$$= \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla \log p(\boldsymbol{x}_t)$$

$p_\theta\ (\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$の平均

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)}{1-\bar{\alpha}_t}$$

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t) = \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}}\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)$$

良い感じに式変形

$\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)$と定数などを
適当に置き換え

先述と同様にLoss, KL最小化は二乗誤差 $\left|\left|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t) - \nabla \log p(\boldsymbol{x}_t)\right|\right|^2$の最小化に帰着する。

よって、
Diffusion Modelの学習は**スコア推定の問題**としても解釈できる (※スコア関数も、NNでモデリングする)

© NTT Communications Corporation All Rights Reserved.

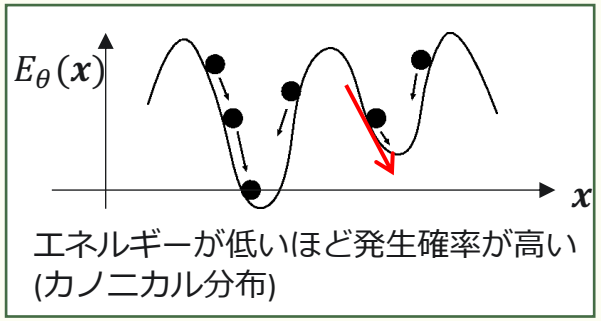
(再掲)生成モデル(エネルギーモデル,スコアベースモデル)

- 生成モデルの系譜 (知ってる人向けの説明)
  - ”Likelihood-Based” (原文ママ)
    - パラメトリックな確率モデルを考えて、データに対する尤度を最大化してフィッティング
      - Energy Based Models
        - よく統計力学とかでありそうな分布形、$p(\boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{\theta}}}\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))$を仮定して尤度Lossで$E_{\boldsymbol{\theta}}(\boldsymbol{x})$を学習
        - 統計力学の知見を流用しつつ、尤度に登場する期待値(規格化定数 $Z_{\boldsymbol{\theta}}$)計算を近似とサンプリングで上手く回避
          - $Z_{\boldsymbol{\theta}}$を求めず分布 $p(\boldsymbol{x})$ からサンプリングする方法 (MCMC+ランジュバンダイナミクス)
            - $\boldsymbol{x}^{t+1} = \boldsymbol{x}^{t} + \eta\, \nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{\omega}$　($\boldsymbol{\omega}$: ガウスノイズ, η: ハイパーパラメタ)
          - ※ スコア $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log(\frac{1}{Z_{\boldsymbol{\theta}}}\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))) = \nabla_{\boldsymbol{x}} \{\log(\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x})) - \log(\mathrm{Z}_{\boldsymbol{\theta}})\} = -\nabla_{\boldsymbol{x}} E_{\boldsymbol{\theta}}(\boldsymbol{x})$
    - Score Based Models
      - 発想：$E_{\boldsymbol{\theta}}(\boldsymbol{x})$ を介さずに直接 $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})$ だけ学習したらいいのでは？？
        - スコアマッチング：
          - データからスコアをフィッティングするLossを計算するための工夫(変形)

Fisher Divergence の最小化問題 →

$$\mathrm{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\frac{1}{2}\left|\left|\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x}) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\right|\right|_2^2\right] = \mathrm{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\frac{1}{2}\left|\left|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\right|\right|_2^2 + \mathrm{tr}\left(\nabla_{\boldsymbol{x}}\, \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right]$$

ここがわからなくてもOK



エネルギーが低いほど発生確率が高い
(カノニカル分布)

筆者曰く、Diffusion Modelは尤度ベース、スコアベースどちらとしても解釈可能とのこと (楽しみ)

© NTT Communications Corporation All Rights Reserved.

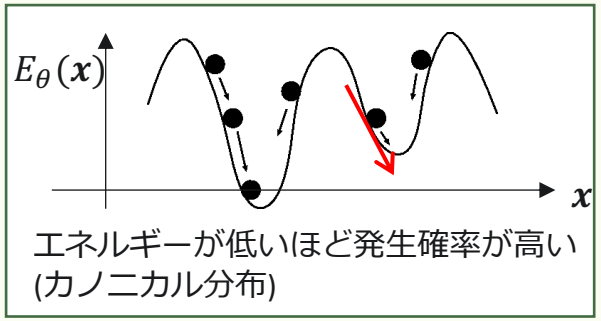
(再掲)生成モデル(エネルギーモデル,スコアベースモデル)

- 生成モデルの系譜 (知ってる人向けの説明)
  - ”Likelihood-Based” (原文ママ)
    - パラメトリックな確率モデルを考えて、データに対する尤度を最大化してフィッティング
      - Energy Based Models
        - よく統計力学とかでありそうな分布形、$p(\boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{\theta}}}\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))$を仮定して尤度Lossで$E_{\boldsymbol{\theta}}(\boldsymbol{x})$を学習
        - 統計力学の知見を流用しつつ、尤度に登場する期待値(規格化定数 $Z_{\boldsymbol{\theta}}$)計算を近似とサンプリングで上手く回避
          - $Z_{\boldsymbol{\theta}}$を求めず分布 $p(\boldsymbol{x})$ からサンプリングする方法 (MCMC+ランジュバンダイナミクス)
            - $\boldsymbol{x}^{t+1} = \boldsymbol{x}^{t} + \eta\, \nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{\omega}$　($\boldsymbol{\omega}$: ガウスノイズ, η: ハイパーパラメタ)
          - ※ スコア $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log(\frac{1}{Z_{\boldsymbol{\theta}}}\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))) = \nabla_{\boldsymbol{x}} \{\log(\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x})) - \log(Z_{\boldsymbol{\theta}})\} = -\nabla_{\boldsymbol{x}} E_{\boldsymbol{\theta}}(\boldsymbol{x})$
    - Score Based Models
      - 発想：$E_{\boldsymbol{\theta}}(\boldsymbol{x})$ を介さずに直接 $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})$ だけ学習したらいいのでは？？
        - スコアマッチング：
          - データからスコアをフィッティングするLossを計算するための工夫(変形)

Fisher Divergence の最小化問題 →

$$\mathrm{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\frac{1}{2}\left|\left|\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x}) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\right|\right|_2^2\right] = \mathrm{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\frac{1}{2}\left|\left|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\right|\right|_2^2 + \mathrm{tr}\left(\nabla_{\boldsymbol{x}}\, \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right]$$

ここがわからなくてもOK



エネルギーが低いほど発生確率が高い
(カノニカル分布)

筆者曰く、Diffusion Modelは尤度ベース、スコアベースどちらとしても解釈可能とのこと (楽しみ)

© NTT Communications Corporation All Rights Reserved.

(再掲)生成モデル(エネルギ

ランジュバンダイナミクスが
逆拡散過程と対応。

また以下のような式も導出できる

$$\boldsymbol{x}_0 = \frac{\boldsymbol{x}_t + (1-\bar{\alpha}_t)\nabla \log p(\boldsymbol{x}_t)}{\sqrt{\bar{\alpha}_t}} = \frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0}{\sqrt{\bar{\alpha}_t}}$$

$$\therefore (1-\bar{\alpha}_t)\nabla \log p(\boldsymbol{x}_t) = -\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0$$

$$\nabla \log p(\boldsymbol{x}_t) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_0$$

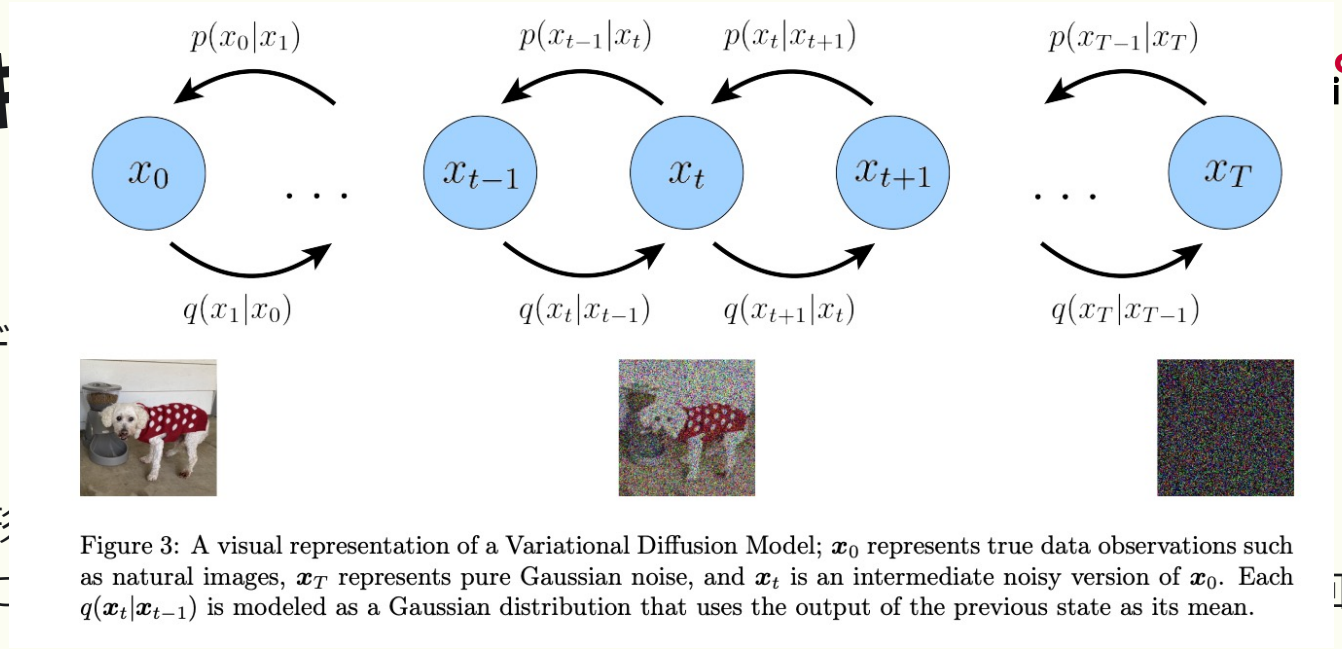


Figure 3: A visual representation of a Variational Diffusion Model; $\boldsymbol{x}_0$ represents true data observations such as natural images, $\boldsymbol{x}_T$ represents pure Gaussian noise, and $\boldsymbol{x}_t$ is an intermediate noisy version of $\boldsymbol{x}_0$. Each $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ is modeled as a Gaussian distribution that uses the output of the previous state as its mean.

- $Z_{\boldsymbol{\theta}}$を求めず分布 $p(\boldsymbol{x})$ からサンプリングする方法 (MCMC+ランジュバンダイナミクス)
  - $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \eta\, \nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{\omega}$　($\boldsymbol{\omega}$: ガウスノイズ, η: ハイパーパラメタ)

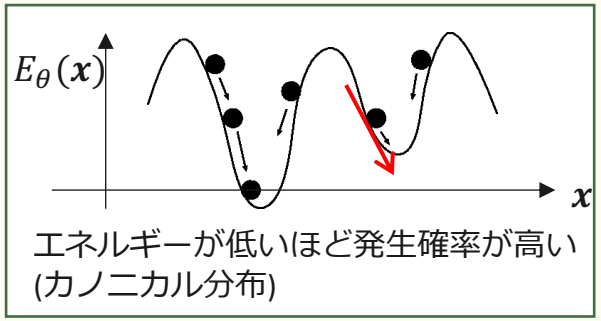
- ※ スコア $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log(\frac{1}{Z_{\boldsymbol{\theta}}} \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))) = \nabla_{\boldsymbol{x}} \{\log(\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x})) - \log(Z_{\boldsymbol{\theta}})\} = -\nabla_{\boldsymbol{x}} E_{\boldsymbol{\theta}}(\boldsymbol{x})$

- Score Based Models
  - 発想：$E_{\boldsymbol{\theta}}(\boldsymbol{x})$ を介さずに直接 $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})$ だけ学習したらいいのでは？？
    - スコアマッチング：
      - データからスコアをフィッティングするLossを計算するための工夫(変形)

Fisher Divergence の最小化問題

$$\mathrm{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\frac{1}{2}\left\|\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\right\|_2^2\right] = \mathrm{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\frac{1}{2}\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\right\|_2^2 + \mathrm{tr}\left(\nabla_{\boldsymbol{x}}\, \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right]$$

ここがわからなくてもOK



エネルギーが低いほど発生確率が高い
(カノニカル分布)

筆者曰く、Diffusion Modelは尤度ベース、スコアベースどちらとしても解釈可能とのこと (楽しみ)

© NTT Communications Corporation All Rights Reserved.

スコア推定としての解釈 → Score Based Models

まとめ

- 先ほどのスコア推定としての解釈結果$\left|\left|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \nabla \log p(\boldsymbol{x}_t)\right|\right|^2$を利用して、Diffusion ModelをScore Based Models として再解釈可能である
- Score Based Modelsのランジュバンダイナミクスと逆拡散過程(デコーダ)が対応している
- Score Based Modelsと同様、ランジュバンダイナミクスを使ったサンプリングも可能である
  - つまり言い換えると、デコーダ$p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t\right)$を繰り返し適用したサンプリングが可能である
- 連続時間のダイナミクスとかも考えられるらしい、Neural ODEとの組み合わせとか
- あとは解説しなかったですが条件付きサンプリングとか(DALL-Eとかで使われてるやつですね)

素朴な疑問
- じゃあここで言うエネルギーってなんやろ？？
  - データの尤度でしょうか…？？

© NTT Communications Corporation All Rights Reserved.

参考) ランジュバンダイナミクス

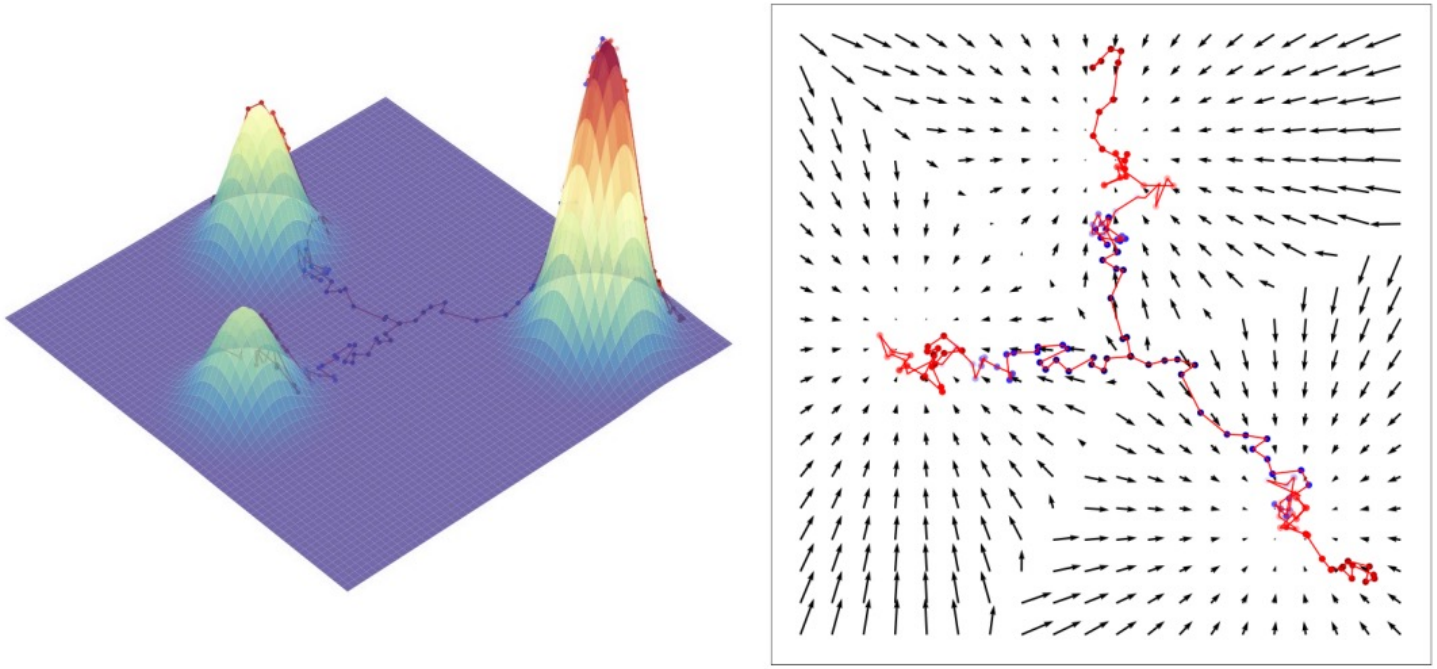


Figure 6: Visualization of three random sampling trajectories generated with Langevin dynamics, all starting from the same initialization point, for a Mixture of Gaussians. The left figure plots these sampling trajectories on a three-dimensional contour, while the right figure plots the sampling trajectories against the ground-truth score function. From the same initialization point, we are able to generate samples from different modes due to the stochastic noise term in the Langevin dynamics sampling procedure; without it, sampling from a fixed point would always deterministically follow the score to the same mode every trial.

© NTT Communications Corporation All Rights Reserved.