非平衡熱力学を用いた教師なしディープラーニング

目次

- 概要
- 1 はじめに
  - 1.1 拡散確率モデル
  - 1.2 他の研究との関係
- 2 アルゴリズム
  - 2.1 前進軌道
  - 2.2 逆軌道
  - 2.3 モデル確率
  - 2.4 学習
    - 2.4.1 拡散レート βt の設定
  - 2.5 分布の乗算と事後分布の計算
    - 2.5.1 修正マージナル分布
    - 2.5.2 修正拡散ステップ
    - 2.5.3 r $left(x^{(t)}$ $right)$ の適用
    - 2.5.4 r $left(x^{(t)}$ $right)$ の選択
  - 2.6 逆プロセスのエントロピー
- 3 実験
  - 3.1 玩具問題
    - 3.1.1 ロールケーキ
    - 3.1.2 バイナリーハートビート分布
  - 3.2 画像
    - 3.2.1 データセット
      - MNIST
      - CIFAR-10
      - Dead Leaf Images
      - Bark Texture Images
- 4 結論
- 謝辞
- A 条件付きエントロピー境界の導出
- B 対数尤度の下界
  - B.1 $p(X^{(T)})$ のエントロピー
  - B.2 $t = 0$ におけるエッジ効果の除去
  - B.3 事後 $q(x^{(t-1)}|x^{(0)})$ で書き換える
  - B.4 KLダイバージェンスとエントロピーで書き直す
- C 摂動ガウス遷移
- D 実験内容
  - D.1 玩具問題
    - D.1.1 ロールケーキ
    - D.1.2 バイナリーハートビート分布
  - D.2 画像
    - D.2.1 構造
      - 読み取り
      - 時間依存性
      - 平均と分散

    - マルチスケール畳み込み
    - 緻密なレイヤー
- 参照

概要

機械学習における中心的な問題は、学習、サンプリング、推論、評価が解析的あるいは計算量的に扱いやすい、非常に柔軟な確率分布族を用いて複雑なデータセットをモデル化することである。ここでは、柔軟性と扱いやすさの両方を同時に達成するアプローチを開発する。非平衡統計物理学に着想を得た本質的なアイデアは、反復的な順拡散過程を通じて、データ分布の構造を系統的にゆっくりと破壊することである。その後、データの構造を復元する逆拡散過程を学習することで、柔軟性が高く扱いやすいデータの生成モデルが得られる。このアプローチにより、何千もの層や時間ステップを持つ深い生成モデルの確率を迅速に学習、サンプリング、評価することができ、また学習したモデルの下での条件付き確率や事後確率を計算することができる。さらに、このアルゴリズムのリファレンス実装をオープンソースで公開している。

1 はじめに

歴史的に、確率モデルは「扱いやすさ」と「柔軟性」という相反する2つの目的のトレードオフに悩まされてきた。扱いやすいモデルは、解析的に評価することができ、データ（例えば、ガウスやラプラス）に簡単に適合させることができる。しかし、これらのモデルでは、豊富なデータセットの構造を適切に記述することはできない。一方、柔軟なモデルは、任意のデータの構造に適合するように成形することができる。例えば、柔軟な分布 $p(x) = \frac{\phi(x)}{Z}$ をもたらす任意の（非負）関数 $\varphi(x)$ でモデルを定義することができる。しかし、この正規化定数を計算するのは一般的に困難である。このような柔軟なモデルの評価、学習、またはサンプルの抽出には、通常、非常に高価なモンテカルロ処理が必要である。

例えば、平均場理論とその展開 [1] [2]、変分ベイズ [3]、対比発散 [4] [5]、最小確率フロー [6] [7]、最小KL縮約 [8]、適切なスコアリングルール[9] [10]、スコアマッチング [11]、擬尤度[12]、ルーピー確率伝播法 [13]、その他多数。ノンパラメトリック手法 [14] も非常に効果的である（ノンパラメトリック手法は、扱いやすいモデルと柔軟なモデルの間をスムーズに移行するものと見なすことができる。例えば、ノンパラメトリックのガウス混合モデルは、単一のガウスを使って少量のデータを表現するが、無限のデータを無限のガウスの混合として表現することができる）。

1.1 拡散確率モデル

我々は、

1. モデル構造の柔軟性、
2. 厳密なサンプリング、
3. 事後分布を計算するための他の分布との簡単な乗算、
4. モデルの対数尤度や個々の状態の確率を安価に評価すること

を可能にする、確率モデルを定義する新しい方法を提示する。

これは非平衡統計物理学 [15] や逐次モンテカルロ法 [16] で使われている考え方である。我々は、単純な既知の分布（例えばガウス分布）を拡散過程を用いてターゲット（データ）分布に変換する生成マルコフ連鎖を構築する。このマルコフ連鎖を使って、別の方法で定義されたモデルを近似的に評価するのではなく、マルコフ連鎖の終点として確率モデルを明示的に定義する。拡散連鎖の各ステップは解析的に評価可能な確率を持つので、完全な連鎖も解析的に評価できる。

この枠組みにおける学習は、拡散過程に対する小さな摂動を推定することを含む。小さな摂動を推定することは、単一の非解析的正規化可能なポテンシャル関数で完全な分布を明示的に記述するよりも扱いやすい。さらに、拡散過程はどのような滑らかな対象分布に対しても存在するため、この方法は任意の形式のデータ分布を捉えることができる。

この拡散確率モデルの有用性を、2次元のロールケーキ、2値配列、手書き数字（MNIST）、およびいくつかの自然画像（CIFAR-10、樹皮、枯葉）のデータセットに対して高対数尤度モデルを学習することにより実証する。

1.2 他の研究との関係

wake-sleep アルゴリズム [17] [18] は、推論と生成確率モデルを相互に学習させるというアイデアを導入した。このアプローチは、いくつかの例外 [19] [20] はあるものの、20年近くほとんど未解明のままであった。最近、この考え方を発展させる研究が爆発的に増えている。[21] [22] [23] [24] では、変分学習と推論アルゴリズムが開発され、柔軟な生成モデルと潜在変数に対する事後分布を、互いに直接学習できるようになった。

これらの論文の変分境界は、我々の訓練目的や [19:1] の先行研究で用いられているものと類似している。しかし、我々の動機とモデルの形式は全く異なっており、本論文ではこれらの手法と比較して以下のような違いや利点を残している。

1. 我々は、変分ベイズ法からではなく、物理学、準静的過程、アニールされた重要度サンプリングからのアイデアを用いてフレームワークを開発する。
2. 学習した分布と他の確率分布（例えば、事後分布を計算するための条件付き分布）を簡単に掛け合わせる方法を示す。
3. 推論モデルと生成モデルの間の目的が非対称であるため、変分推論法では推論モデルの学習が特に困難であるという問題に対処する。逆（生成）過程も同じ関数形を持つように、順（推論）過程を単純な関数形に制限する。
4. 私たちは、ほんの一握りのレイヤーではなく、何千ものレイヤー（または時間ステップ）でモデルを訓練する。
5. 各レイヤー（または時間ステップ）におけるエントロピー生成の上界と下界を提供する。

確率モデルを訓練するための関連技術（以下に要約）は数多くあり、生成モデルのための柔軟性の高いフォームを開発したり、確率的軌道を訓練したり、ベイジアンネットワークの反転を学習したりする。Reweighted wake-sleep [25] は、オリジナルのウェイクスリープ・アルゴリズムの拡張と改良された学習ルールを開発している。生成確率ネットワーク [26] [27] は、マルコフ・カーネルを訓練して、その均衡分布をデータ分布に一致させる。ニューラル自己回帰分布推定量 [28]（およびそのリカレント [29] とディープ [30] 拡張）は、結合分布を各次元にわたる扱いやすい条件付き分布の列に分解する。敵対的ネットワーク [31] は、生成されたサンプルを真のデータと区別しようとする分類器に対して、生成モデルを学習する。[32] における同様の目的は、マージン的に独立したユニットを持つ表現への双方向写像を学習することである。[33] [34] では、単純な階乗密度関数を持つ潜在表現に対して、双射的決定性写像が学習される。[35] では、ベイジアンネットワークに対して確率的逆行列が学習される。条件付きガウススケール混合 (MCGSMs) [36] は、一連の因果近傍に依存するパラメータを持つガウススケール混合を用いてデータセットを記述する。さらに、単純な潜在分布からデータ分布への柔軟な生成マッピングを学習する重要な研究がある。初期の例としては、ニューラルネットワークを生成モデルとして導入した [37] や、潜在空間からデータ空間への確率多様体マッピングを学習した[38]がある。敵対的ネットワークと MCGSMs に対して実験的に比較する。

物理学からの関連するアイデアには、機械学習ではアニールされた重要度サンプリング（AIS）[16:1] として知られるヤジンスキー方程式 [15:1] があり、これは正規化定数の比率を計算するために、ある分布を別の分布にゆっくりと変換するマルコフ連鎖を使用する。[39] では、AISは順方向ではなく逆方向の軌道を使っても実行できることが示されている。フォッカー・プランク方程式の確率的実現であるランジュヴァン・ダイナミクス [40] は、任意の目標分布を平衡とするガウス拡散過程を定義する方法を示している。[41] では、Fokker-Planck 方程式が確率最適化に用いられている。最後に、Kolmogorov の前進・後退方程式 [42] は、多くの前進拡散過程について、逆拡散過程も同じ関数形で記述できることを示している。

2 アルゴリズム

我々の目標は、どんな複雑なデータ分布も単純で扱いやすい分布に変換する順方向（または推論）拡散過程を定義し、生成モデル分布を定義するこの拡散過程の有限時間反転を学習することである（図1）。まず、前方への推論拡散プロセスについて説明する。次に、逆生成拡散プロセスを学習し、確率評価に使用する方法を示す。また、逆プロセスのエントロピー境界を導出し、学習された分布がどのように任意の2番目の分布と乗算できるかを示す（例えば、画像のインペイントやノイズ除去の際に事後分布を計算するために行われるようなもの）。

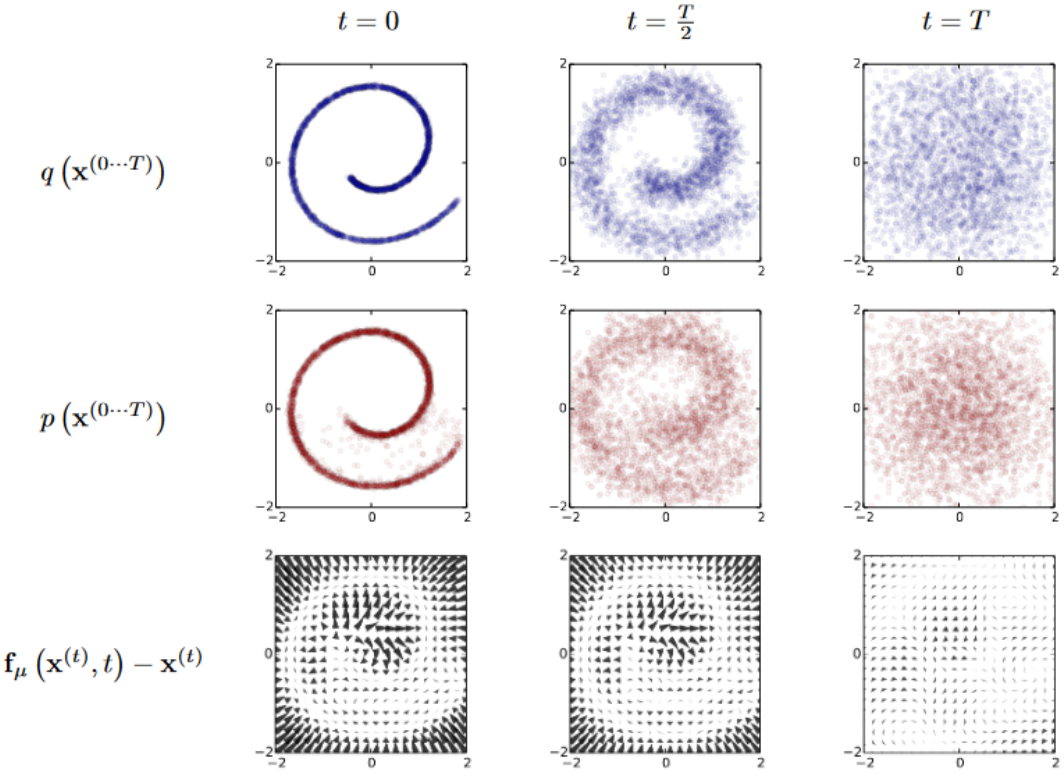

図1：**提案するモデリングフレームワークの2次元のロールケーキデータでの学習**
上段は前方軌道 $q\left(x^{(0\cdots T)}\right)$ からのタイムスライス。データ分布（左）はガウス拡散を受け、徐々に恒等共分散ガウス分布（右）に変化する。中段は学習された逆軌跡 $p\left(x^{(0\cdots T)}\right)$ からの対応するタイムスライスを示す。恒等共分散ガウス分布（右）は、平均と共分散関数を学習したガウス拡散過程を経て、徐々にデータ分布（左）に戻る。下段は同じ逆拡散過程のドリフト項 $f_\mu\left(x^{(t)}, t\right) - x^{(t)}$ を示す。

2.1 前進軌道

データ分布 $q\left(x^{(0)}\right)$ にラベルを付ける。このデータ分布は、$\pi(y)$ に対してマルコフ拡散カーネル $T_\pi(y|y';\beta)$ を繰り返し適用することで、$\pi(y)$ が徐々にお行儀の良い（解析的に扱いやすい）分布に変換される（β は拡散率）。

$$\pi(y) = \int dy' T_\pi(y|y';\beta)\pi(y) \tag{1}$$

$$q\left(x^{(t)}|x^{(t-1)}\right) = T_\pi\left(x^{(t)}|x^{(t-1)};\beta_t\right) \tag{2}$$

従って、データ分布から出発して T 回の拡散ステップを実行することに対応する順方向軌道は、以下のようになる。

$$q\left(x^{(0\cdots T)}\right) = q\left(x^{(0)}\right)\prod_{t=1}^{T} q\left(x^{(t)}|x^{(t-1)}\right) \tag{3}$$

以下に示す実験では、$q\left(x^{(t)}|x^{(t-1)}\right)$ は、恒等共分散を持つガウス分布へのガウス拡散か、独立二項分布への二項拡散のいずれかに対応する。表App.1にガウス分布と二項分布の拡散カーネルを示す。

2.2 逆軌道

生成分布は、同じ軌跡を記述するように学習されるが、逆に、

$$p\left(x^{(T)}\right) = \pi\left(x^{(T)}\right) \tag{4}$$

$$p\left(x^{(0\cdots T)}\right) = p\left(x^{(T)}\right)\prod_{t=1}^{T} p\left(x^{(t-1)}|x^{(t)}\right) \tag{5}$$

ガウス拡散と二項拡散の両方について、連続拡散（小さなステップサイズ β の極限）では、拡散過程の反転は前進過程と同じ関数形を持つ[42:1]。$q\left(x^{(t)}|x^{(t-1)}\right)$ はガウス(二項)分布であり、β_t が小さければ $q\left(x^{(t-1)}|x^{(t)}\right)$ もガウス(二項)分布となる。軌跡が長ければ長いほど、拡散率 β を小さくすることができる。

学習時には、ガウス拡散カーネルであれば平均と共分散、二項カーネルであればビットフリップ確率のみを推定すればよい。表App.1に示すように、$f_\mu\left(x^{(t)},t\right)$ と $f_\Sigma\left(x^{(t)},t\right)$ はガウスの逆マルコフ遷移の平均と共分散を定義する関数であり、$f_b\left(x^{(t)},t\right)$ は二項分布のビットフリップ確率を提供する関数である。このアルゴリズムを実行するための計算コストは、これらの関数のコストにタイムステップ数を掛けたものになる。本論文のすべての結果において、これらの関数を定義するために多層パーセプトロンが使用されている。しかし、ノンパラメータの手法も含め、幅広い回帰や関数フィッティングの手法が適用できるだろう。

2.3 モデル確率

生成モデルがデータに割り当てる確率は以下のように表現される。

$$p\left(x^{(0)}\right)=\int dx^{(1\cdots T)}p\left(x^{(0\cdots T)}\right) \tag{6}$$

直観的には、この積分は難解である。しかし、アニールされた重要度サンプリングと Jarzynski の等式からヒントを得て、代わりに順方向と逆方向の軌跡の相対確率を評価し、順方向の軌跡を平均する。

$$\begin{aligned}p\left(x^{(0)}\right)&=\int dx^{(1\cdots T)}p\left(x^{(0\cdots T)}\right)\frac{q\left(x^{(1\cdots T)}|x^{(0)}\right)}{q\left(x^{(1\cdots T)}|x^{(0)}\right)} &(7)\\&=\int dx^{(1\cdots T)}q\left(x^{(1\cdots T)}|x^{(0)}\right)\frac{p\left(x^{(0\cdots T)}\right)}{q\left(x^{(1\cdots T)}|x^{(0)}\right)} &(8)\\&=\int dx^{(1\cdots T)}q\left(x^{(1\cdots T)}|x^{(0)}\right)p\left(x^{(T)}\right)\prod_{t=1}^{T}\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t)}|x^{(t-1)}\right)} &(9)\end{aligned}$$

これは、前方軌道 $q\left(x^{(1\cdots T)}|x^{(0)}\right)$ のサンプルを平均化することで高速に評価できる。β が無限小の場合、軌道上の順方向分布と逆方向分布は同一にすることができる（2.2節）。もし両者が同一であれば、上の積分を正確に評価するために必要なのは $q\left(x^{(1\cdots T)}|x^{(0)}\right)$ からのサンプル1つだけである。これは統計物理学における準静的過程の場合に相当する [43] [44]。

2.4 学習

学習はモデルの対数尤度

$$\begin{aligned}L&=\int dx^{(0)}q\left(x^{(0)}\right)\log p\left(x^{(0)}\right) &(10)\\&=\int dx^{(0)}q\left(x^{(0)}\right)\log\left[\int dx^{(1\cdots T)}q\left(x^{(1\cdots T)}|x^{(0)}\right)p\left(x^{(T)}\right)\prod_{t=1}^{T}\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t)}|x^{(t-1)}\right)}\right] &(11)\end{aligned}$$

を最大化することになり、これは Jensen の不等式

$$L\geq\int dx^{(0\cdots T)}q\left(x^{(0\cdots T)}\right)\log\left[p\left(x^{(T)}\right)\prod_{t=1}^{T}\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t)}|x^{(t-1)}\right)}\right] \tag{12}$$

によって下界が与えられる。付録Bにあるように、我々の拡散軌道では、エントロピーとKLダイバージェンスが解析的に計算できる

$$\begin{aligned}&L\geq K &(13)\\&K=-\sum_{t=2}^{T}\int dx^{(0)}dx^{(t)}q\left(x^{(0)},x^{(t)}\right)D_{KL}\left(q\left(x^{(t-1)}|x^{(t)},x^{(0)}\right)||p\left(x^{(t-1)}|x^{(t)}\right)\right)+H_q\left(X^{(T)}|X^{(0)}\right)-H_q\left(X^{(1)}|X^{(0)}\right)-H_p\left(X^{(T)}\right) &(14)\end{aligned}$$

に帰着する。この境界の導出は、変分ベイズ法における対数尤度の境界の導出と類似している。

2.3節のように、順方向と逆方向の軌道が同一であれば、準静的過程に相当し、式13の不等式は等式になる。

学習は、この対数尤度の下界

$$\hat{p}\left(x^{(t-1)}|x^{(t)}\right)=\underset{p\left(x^{(t-1)}|x^{(t)}\right)}{\arg\max}\ K$$

を最大化する逆マルコフ遷移を見つけることからなる。ガウス拡散と二項拡散の具体的な推定対象を表App.1に示す。

このように、確率分布を推定するタスクは、ガウシアン列の平均と共分散を設定する（またはベルヌーイ試行列の状態フリップ確率を設定する）関数に回帰を実行するタスクに縮小された。

2.4.1 拡散レート βt の設定

前進軌道における β_t の選択は、学習済みモデルの性能にとって重要である。AISでは、中間分布の適切なスケジュールは、対数分割関数の推定精度を大幅に向上させることができる[45]。熱力学では、平衡分布間を移動する際のスケジュールが、どれだけの自由エネルギーが失われるかを決定する[43:1] [44:1]。

ガウス拡散の場合、K に対する勾配上昇によって前方拡散スケジュール $\beta_{2\ldots T}$ を学習する。最初のステップの分散 β_1 は、オーバーフィッティングを防ぐために小さな定数に固定されている。凍結ノイズ (frozen noise) を使用することで、$q\left(x^{(1\cdots T)}|x^{(0)}\right)$ からのサンプルの $\beta_{1\ldots T}$ への依存性が明示される。[21:1]のように、ノイズは追加の補助変数として扱われ、パラメータに関して K の偏導関数を計算する間、一定に保たれる。

二項拡散の場合、離散的な状態空間は、凍結ノイズを伴う勾配上昇を不可能にする。代わりに、拡散ステップごとに元の信号の一定割合 $\frac{1}{T}$ を消去するように、順拡散スケジュール $\beta_{1\ldots T}$ を選択し、$\beta_t = (T-t+1)^{-1}$ の拡散率を得る。

2.5 分布の乗算と事後分布の計算

信号ノイズ除去や欠損値の推論を行うために事後分布を計算するようなタスクは、モデル分布 $p\left(x^{(0)}\right)$ と第2の分布、または境界付き正関数 $r\left(x^{(0)}\right)$ の掛け算を必要とし、新しい分布 $\tilde{p}\left(x^{(0)}\right) \propto p\left(x^{(0)}\right) r\left(x^{(0)}\right)$ を生成する。

分布の乗算は、変分オートエンコーダ、GSNs、NADEs、ほとんどのグラフィカルモデルなど、多くの技術にとってコストがかかり困難である。しかし、拡散モデルの下では、第2の分布は拡散過程の各ステップに対する小さな摂動として扱うか、あるいは多くの場合、各拡散ステップに正確に掛け合わせることができるので、これは簡単である。図3と図5は、自然画像のノイズ除去とインペインティングを行うために拡散モデルを使用することを示している。以下のセクションでは、拡散確率モデルの文脈における分布の乗算方法について説明する。

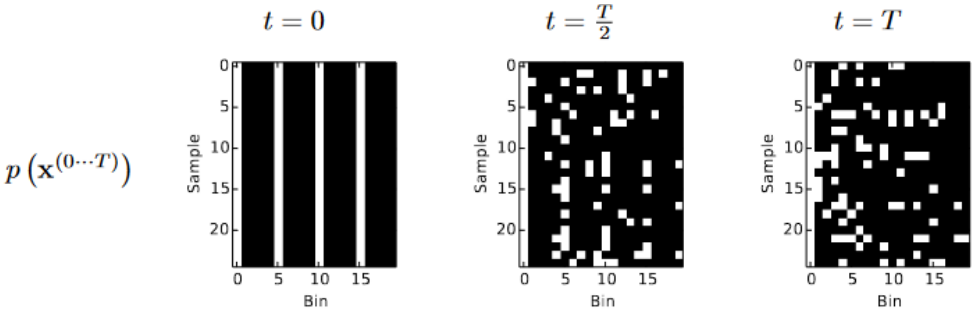

図2：**二項拡散による二値系列学習**
2項拡散モデルは、5ビンごとにパルスが発生する2値の心拍データで学習された。生成されたサンプル（左）は学習データと同じ。サンプリング手順は、独立した二項ノイズ（右）で初期化され、学習されたビットフリップ確率で、二項拡散プロセスによってデータ分布に変換される。各行は独立したサンプルを含む。可視化を容易にするため、すべてのサンプルは最初の列でパルスが発生するようにシフトされている。生のシーケンスデータでは、最初のパルスは最初の5つのビンに一様に分布している。

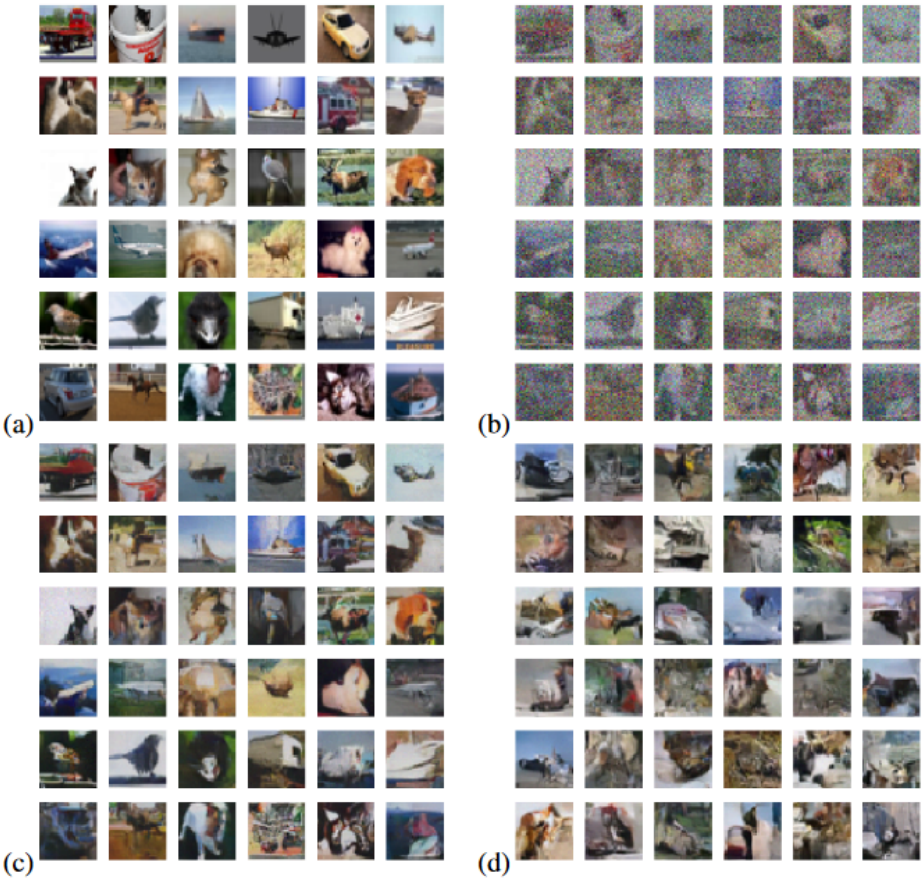

図3：**CIFAR-10 [46] データセットで学習した提案フレームワーク**
(a)がホールドアウトデータの例（学習データと同様）、(b)が分散 1 のガウスノイズ $(\mathrm{SNR}=1)$ で汚染されたホールドアウトデータ、(c)が(b)の画像を条件として、ノイズ除去画像に対する事後分布からサンプリングして生成したノイズ除去画像、(d)が拡散モデルによって生成されたサンプル。

2.5.1 修正マージナル分布

まず、$\tilde{p}\left(x^{(0)}\right)$ を計算するために、各中間分布に対応する関数 $r\left(x^{(t)}\right)$ を掛ける。分布やマルコフ遷移の上にあるチルダ記号は、その分布やマルコフ遷移がこのように変更された軌道に属することを示す。$\tilde{p}\left(x^{(0\cdots T)}\right)$ は、分布 $\tilde{p}\left(x^{(T)}\right)=\frac{1}{\tilde{Z}_T}p\left(x^{(T)}\right)r\left(x^{(T)}\right)$ から始まり、中間分布のシーケンス

$$\tilde{p}\left(x^{(t)}\right)=\frac{1}{\tilde{Z}_t}p\left(x^{(t)}\right)r\left(x^{(t)}\right) \tag{16}$$

を進む修正逆軌跡であり、$\tilde{Z}_t$ は t 番目の中間分布の正規化定数である。

2.5.2 修正拡散ステップ

逆拡散過程のマルコフ・カーネル $p\left(x^{(t)}|x^{(t+1)}\right)$ は平衡条件

$$p\left(x^{(t)}\right)=\int dx^{(t+1)}p\left(x^{(t)}|x^{(t+1)}\right)p\left(x^{(t+1)}\right) \tag{17}$$

に従う。我々は、摂動マルコフ・カーネル $\tilde{p}\left(x^{(t)}|x^{(t+1)}\right)$ が摂動分布の平衡条件

$$\tilde{p}\left(x^{(t)}\right)=\int dx^{(t+1)}\tilde{p}\left(x^{(t)}|x^{(t+1)}\right)\tilde{p}\left(x^{(t+1)}\right) \tag{18}$$

$$\frac{p\left(x^{(t)}\right)r\left(x^{(t)}\right)}{\tilde{Z}_t}=\int dx^{(t+1)}\tilde{p}\left(x^{(t)}|x^{(t+1)}\right)\frac{p\left(x^{(t+1)}\right)r\left(x^{(t+1)}\right)}{\tilde{Z}_{t+1}} \tag{19}$$

$$p\left(x^{(t)}\right)=\int dx^{(t+1)}\tilde{p}\left(x^{(t)}|x^{(t+1)}\right)\frac{\tilde{Z}_t r\left(x^{(t+1)}\right)}{\tilde{Z}_{t+1}r\left(x^{(t)}\right)}p\left(x^{(t+1)}\right) \tag{20}$$

に従うことを望む。式20は以下の式が成り立つときに成り立つ。

$$\tilde{p}\left(x^{(t)}|x^{(t+1)}\right) = p\left(x^{(t)}|x^{(t+1)}\right)\frac{\tilde{Z}_{t+1}r\left(x^{(t)}\right)}{\tilde{Z}_t r\left(x^{(t+1)}\right)} \tag{21}$$

式21は正規化確率分布に対応しない可能性があるため、$\tilde{Z}_t\left(x^{(t+1)}\right)$ を正規化定数とし、$\tilde{p}\left(x^{(t)}|x^{(t+1)}\right)$ を対応する正規化分布

$$\tilde{p}\left(x^{(t)}|x^{(t+1)}\right) = \frac{1}{\tilde{Z}_t\left(x^{(t+1)}\right)}p\left(x^{(t)}|x^{(t+1)}\right)r\left(x^{(t)}\right) \tag{22}$$

とする。

ガウスの場合、分散が小さいため、各拡散ステップは $r\left(x^{(t)}\right)$ に対して非常に鋭いピークを持つのが一般的である。これは、$\frac{r\left(x^{(t)}\right)}{r\left(x^{(t+1)}\right)}$ が $p\left(x^{(t)}|x^{(t+1)}\right)$ に対する小さな摂動として扱えることを意味する。ガウスに対する小さな摂動は平均に影響するが、正規化定数には影響しないので、この場合、式21と式22は等価である（付録C）。

2.5.3 $r\left(x^{(t)}\right)$ の適用

$r\left(x^{(t)}\right)$ が十分に滑らかであれば、逆拡散カーネル $p\left(x^{(t)}|x^{(t+1)}\right)$ に対する小さな摂動として扱うことができる。この場合、$\tilde{p}\left(x^{(t)}|x^{(t+1)}\right)$ は $p\left(x^{(t)}|x^{(t+1)}\right)$ と同じ関数形を持つが、ガウスカーネルの場合は平均が摂動され、二項カーネルの場合はフリップ率が摂動される。摂動された拡散カーネルは表App.1に示されており、付録Cでガウシアンについて導かれている。

$r\left(x^{(t)}\right)$ が閉じた形でガウス（または二項）分布と掛け合わせることができるなら、閉じた形で逆拡散カーネル $p\left(x^{(t)}|x^{(t+1)}\right)$ と直接掛け合わせることができる。これは、図5のインペインティングの例のように、$r\left(x^{(t)}\right)$ がある座標の部分集合に対するデルタ関数で構成されている場合に適用される。

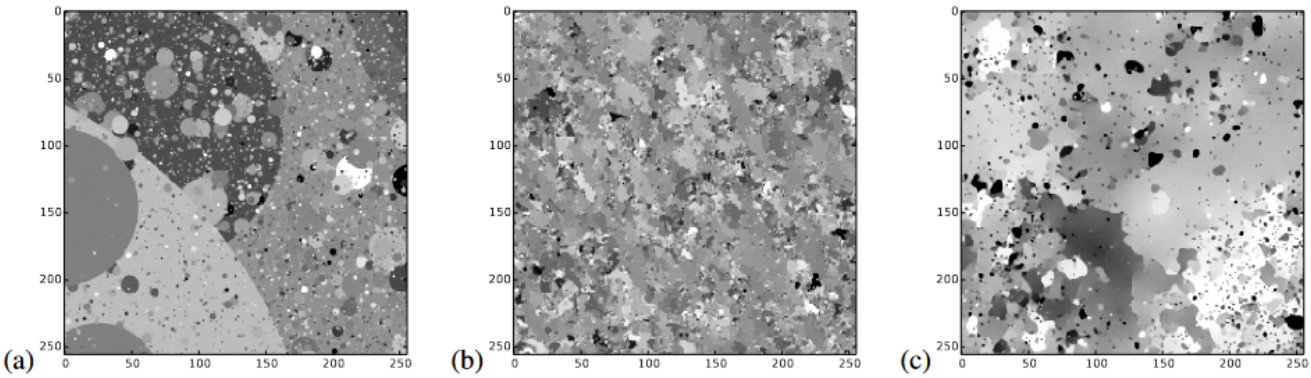

図4：**枯葉画像で学習した提案フレームワーク** [47] [48]

(a)が学習画像の例、(b)が同じデータで学習した以前の最新自然画像モデル[49]のサンプル、(c)が拡散モデルによって生成されたサンプル。かなり一貫したオクルージョン関係を示し、オブジェクトのサイズにマルチスケール分布を示し、特に小さいスケールでは円のようなオブジェクトを生成することに注意したい。表2に示すように、拡散モデルはテストセットで最も高い対数尤度を持つ。

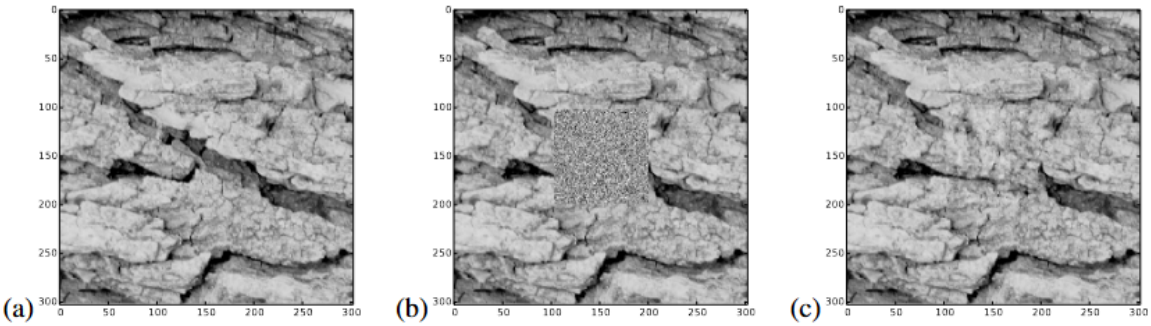

図5：**インペインティング**

(a)は [50] の樹皮画像であり、(b)は同じ画像で、中央の 100×100 ピクセルの領域を等方性ガウスノイズで置き換えたもの。これは逆軌跡の初期化 $\tilde{p}\left(x^{(T)}\right)$ である。(c)中央の 100×100 の領域は、樹皮の画像で学習した拡散確率モデルを用いて、画像の残りの部分を条件として欠損領域上

の事後分布からサンプリングすることにより、塗りつぶされている。例えば、塗りつぶされた領域の左側に入るクラックのような、長距離の空間構造に注目してほしい。$r\left(x^{(0)}\right)$ は既知のデータに対してはデルタ関数に、欠損データに対しては定数に設定した。

Dataset	K	$K - L_{null}$
Swiss Roll	2.35 bits	6.45 bits
Binary Heartbeat	-2.414 bits/seq.	12.024 bits/seq.
Bark	-0.55 bits/pixel	1.5 bits/pixel
Dead Leaves	1.489 bits/pixel	3.536 bits/pixel
CIFAR-10[3]	5.4 ± 0.2 bits/pixel	11.5 ± 0.2 bits/pixel
MNIST	See table 2	

表1：**各学習モデルのホールドアウト集合で計算された対数尤度の下界 K**
式12を参照。右の列は、等方性ガウス分布または独立二項分布に対する改善度である。L_{null} は $\pi\left(x^{(0)}\right)$ の対数尤度である。Binary Heartbeatを除くすべてのデータセットは、尤度の対数を計算する前に、分散が 1 になるように定数でスケーリングされた。

Model	*Log Likelihood*
Dead Leaves	
MCGSM	1.244 bits/pixel
Diffusion	**1.489 bits/pixel**
MNIST	
Stacked CAE	174 ± 2.3 bits
DBN	199 ± 2.9 bits
Deep GSN	309 ± 1.6 bits
Diffusion	$\mathbf{317 \pm 2.7}$ **bits**
Adversarial net	325 ± 2.9 bits
Perfect model	349 ± 3.3 bits

表2：**他のアルゴリズムとの対数尤度の比較**
枯葉画像は、[36:1]と同じ学習データとテストデータを用いて評価した。MNIST の対数尤度は[31:1]の Parzen-window コードを用いて推定され、その値はビット単位で与えられている。完璧なモデルエントリーは、学習データのサンプルに Parzen コードを適用して計算された。

2.5.4 $r\left(x^{(t)}\right)$ の選択

通常、$r\left(x^{(t)}\right)$ は軌跡の過程でゆっくりと変化するように選択されるべきである。本稿の実験では、これを一定とした。

$$r\left(x^{(t)}\right) = r\left(x^{(0)}\right)$$

もう一つの便利な選択肢は $r\left(x^{(t)}\right) = r\left(x^{(0)}\right)^{\frac{T-t}{T}}$ である。この2番目の選択では、$r\left(x^{(t)}\right)$ は逆方向軌道の開始分布には寄与しない。これによって、逆軌跡のための $\tilde{p}\left(x^{(T)}\right)$ からの初期サンプルの描画は簡単なままであることが保証される。

2.6 逆プロセスのエントロピー

順方向過程は既知であるので、逆方向の軌跡の各ステップの条件付きエントロピーの上界と下界を導くことができ、したがって対数尤度の上界と下界を導くことができる。

$$\begin{aligned} & H_q\left(X^{(t)}|X^{(t-1)}\right) + H_q\left(X^{(t-1)}|X^{(0)}\right) - H_q\left(X^{(t)}|X^{(0)}\right) \\ \leq & H_q\left(X^{(t-1)}|X^{(t)}\right) \\ \leq & H_q\left(X^{(t)}|X^{(t-1)}\right) \end{aligned} \tag{24}$$

ここで、上界も下界も $q\left(x^{(1\cdots T)}|x^{(0)}\right)$ にのみ依存し、解析的に計算できる。導出は付録Aに記載されている。

3 実験

様々な連続データセットとバイナリデータセットで拡散確率モデルを学習する。次に、学習済みモデルからのサンプリングと欠損データのインペインティングを実証し、モデルの性能を他の手法と比較する。すべての場合において、目的関数と勾配は Theano [51] を使って計算された。モデルのトレーニングは、CIFAR-10 を除き、SFO [52] を使用した。CIFAR-10 の結果は、アルゴリズムのオープンソース実装と最適化のための

RMSprop を使用した。我々のモデルによって得られる対数尤度の下界は、表1のすべてのデータセットについて報告されている。Blocks [53] を利用したアルゴリズムのリファレンス実装は、https://github.com/Sohl-Dickstein/Diffusion-Probabilistic-Modelsで利用可能である。

3.1 玩具問題

3.1.1 ロールケーキ

$f_\mu\left(x^{(t)}, t\right)$ と $f_\Sigma\left(x^{(t)}, t\right)$ を生成するために放射基底関数ネットワークを用いて、2次元のロールケーキ分布の拡散確率モデルを構築した。図1に示すように、ロールケーキ分布の学習に成功した。詳細は付録D.1.1項を参照のこと。

3.1.2 バイナリーハートビート分布

拡散確率モデルは、多層パーセプトロンを用いて、逆軌跡のベルヌーイ率 $f_b\left(x^{(t)}, t\right)$ を生成し、長さ 20 の単純な二値系列で学習され、5番目の時間ビンごとに 1 が発生し、残りのビンは 0 である。真の分布の下での対数尤度は、シーケンスあたり $\log_2\left(\frac{1}{5}\right) = -2.322$ ビットである。図2と表1を見ればわかるように、学習はほぼ完璧だった。詳細は付録D.1.2項を参照。

3.2 画像

ガウス拡散確率モデルをいくつかの画像データセットで学習した。これらの実験に共通するマルチスケール畳み込みアーキテクチャは、付録D.2.1項で説明され、図D.1に示されている。

3.2.1 データセット

MNIST

簡単なデータセットで先行研究との直接比較を可能にするため、MNIST digits [54] で学習した。[55] [26:1] [31:2] に対する対数尤度を表2に示す。MNIST モデルのサンプルを付録図 App.1 に示す。我々の訓練アルゴリズムは、漸近的に一貫した対数尤度の下界を提供する。しかし、これまでに報告された連続 MNIST 対数尤度に関する結果のほとんどは、モデルサンプルから計算された Parzen-window ベースの推定値に依存している。この比較のために、我々は [31:3] で公開された Parzen-window コードを用いて MNIST の対数尤度を推定した。

CIFAR-10

CIFAR-10チャレンジデータセット[46:1]のトレーニング画像に確率モデルを当てはめた。学習済みモデルのサンプルを図3に示す。

Dead Leaf Images

Dead Leaf Images [47:1] [48:1] は、スケール上のべき乗則分布から描かれた、重層的なオクルージョン円からなる。これらは解析的に扱いやすい構造を持っているが、自然画像の統計的な複雑さの多くを捉えているため、自然画像モデルの説得力のあるテストケースとなる。表2と図4に示すように、我々は枯葉データセットで最先端の性能を達成した。

Bark Texture Images

[50:1]の樹皮テクスチャ画像(T01-T04)を用いて確率モデルを学習した。このデータセットでは、図5でモデルの事後分布からのサンプルを使って欠損データの大きな領域をインペイントすることで、事後分布から評価または生成することが簡単であることを実証している。

4 結論

我々は、確率の厳密なサンプリングと評価を可能にする確率分布をモデル化するための新しいアルゴリズムを導入し、困難な自然画像データセットを含む様々なトイデータセットと実データセットでその有効性を実証した。これらのテストのそれぞれについて、我々は同様の基本アルゴリズムを使用し、我々の手法が様々な分布を正確にモデル化できることを示した。既存の密度推定技術のほとんどは、扱いやすく効率的であり続けるためにモデル化能力を犠牲にしなければならず、サンプリングや評価には非常にコストがかかることが多い。我々のアルゴリズムの核心は、データをノイズ分布に写像するマルコフ拡散連鎖の反転を推定することである。ステップ数を大きくすると、各拡散ステップの反転分布は単純になり、推定が容易になる。その結果、どのようなデータ分布にも適合する学習が可能でありながら、学習、正確なサンプリング、評価が扱いやすく、条件分布と事後分布を簡単に操作できるアルゴリズムが完成した。

謝辞

Lucas Theis、Subhaneil Lahiri、Ben Poole、Diederik P. Kingma、Taco Cohen、Philip Bachman、Aäron van den Oordには非常に有益な議論を、Ian Goodfellow には Parzen-window のコードを提供していただいた。Jascha Sohl-Dickstein に資金を提供してくれた Khan AcademyとOffice of

Naval Research に感謝し、Surya Ganguli に資金を提供してくれた Office of Naval Research と Burroughs-Wellcome 財団、Sloan 財団、James S. McDonnell 財団に感謝する。

A 条件付きエントロピー境界の導出

逆軌跡のステップの条件付きエントロピー $H_q\left(X^{(t-1)}|X^{(t)}\right)$ は以下のようになる。

$$H_q\left(X^{(t-1)},X^{(t)}\right) = H_q\left(X^{(t)},X^{(t-1)}\right) \tag{25}$$

$$H_q\left(X^{(t-1)}|X^{(t)}\right) + H_q\left(X^{(t)}\right) = H_q\left(X^{(t)}|X^{(t-1)}\right) + H_q\left(X^{(t-1)}\right) \tag{26}$$

$$H_q\left(X^{(t-1)}|X^{(t)}\right) = H_q\left(X^{(t)}|X^{(t-1)}\right) + H_q\left(X^{(t-1)}\right) - H_q(X^{(t)}) \tag{27}$$

エントロピー変化の上限は、$\pi(y)$ が最大エントロピー分布であることを観察することで構築できる。これは、二項分布の場合は無条件で成立し、ガウスの場合は分散 1 の訓練データで成立する。ガウシアンの場合、以下の等式が成り立つためには、訓練データは単位ノルムになるようにスケーリングされなければならない。白色化する必要はない。上限は次のように導かれる。

$$H_q\left(X^{(t)}\right) \geq H_q\left(X^{(t-1)}\right) \tag{28}$$

$$H_q\left(X^{(t-1)}\right) - H_q\left(X^{(t)}\right) \leq 0 \tag{29}$$

$$H_q\left(X^{(t-1)}|X^{(t)}\right) \leq H_q\left(X^{(t)}|X^{(t-1)}\right) \tag{30}$$

エントロピーの差の下界は、マルコフ連鎖のステップを増やしても、その連鎖の初期状態について利用可能な情報は増えないので、初期状態の条件付きエントロピーは減らないことを観察することで確立できる。

$$H_q\left(X^{(0)}|X^{(t)}\right) \geq H_q\left(X^{(0)}|X^{(t-1)}\right) \tag{31}$$

$$H_q\left(X^{(t-1)}\right) - H_q\left(X^{(t)}\right) \geq H_q\left(X^{(0)}|X^{(t-1)}\right) + H_q\left(X^{(t-1)}\right) - H_q\left(X^{(0)}|X^{(t)}\right) - H_q\left(X^{(t)}\right) \tag{32}$$

$$H_q\left(X^{(t-1)}\right) - H_q\left(X^{(t)}\right) \geq H_q\left(X^{(0)},X^{(t-1)}\right) - H_q\left(X^{(0)},X^{(t)}\right) \tag{33}$$

$$H_q\left(X^{(t-1)}\right) - H_q\left(X^{(t)}\right) \geq H_q\left(X^{(t-1)}|X^{(0)}\right) - H_q\left(X^{(t)}|X^{(0)}\right) \tag{34}$$

$$H_q\left(X^{(t-1)}|X^{(t)}\right) \geq H_q\left(X^{(t)}|X^{(t-1)}\right) + H_q\left(X^{(t-1)}|X^{(0)}\right) - H_q\left(X^{(t)}|X^{(0)}\right) \tag{35}$$

これらの式を組み合わせることで、1ステップの条件付きエントロピー

$$H_q\left(X^{(t)}|X^{(t-1)}\right) \geq H_q\left(X^{(t-1)}X^{(t)}\right) \geq H_q\left(X^{(t)}|X^{(t-1)}\right) + H_q\left(X^{(t-1)}|X^{(0)}\right) - H_q\left(X^{(t)}|X^{(0)}\right) \tag{36}$$

を束縛することができる。上界も下界も条件付き前進軌道 $q\left(x^{(1\cdots T)}|x^{(0)}\right)$ にのみ依存し、解析的に計算することができる。

B 対数尤度の下界

対数尤度の下界は以下の式で表される。

$$L \geq K \tag{37}$$

$$K = \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log\left[p\left(x^{(T)}\right)\prod_{t=1}^{T}\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t)}|x^{(t-1)}\right)}\right] \tag{38}$$

B.1 $p\left(X^{(T)}\right)$ のエントロピー

$p\left(X^{(T)}\right)$ からの寄与を取り除き、エントロピーとして書き直すことができる。

$$K = \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \sum_{t=1}^{T} \log\left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(T)}|x^{(t-1)}\right)}\right] + \int dx^{(T)} q\left(x^{(T)}\right) \log p\left(x^{(T)}\right) \tag{39}$$

$$= \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \sum_{t=1}^{T} \log\left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(T)}|x^{(t-1)}\right)}\right] + \int dx^{(T)} q\left(x^{(T)}\right) \log \pi\left(x^{T}\right) \tag{40}$$

設計上、$\pi\left(x^{(t)}\right)$ に対するクロスエントロピーは拡散カーネルの下では一定であり、$p\left(x^{(T)}\right)$ のエントロピーに等しい。したがって、以下のようになる。

$$K = \sum_{t=1}^{T} \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log\left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t)}|x^{(t-1)}\right)}\right] - H_p\left(x^{(T)}\right) \tag{41}$$

B.2 $t = 0$ におけるエッジ効果の除去

エッジ効果を避けるため、逆方向の軌跡の最終ステップは、対応する順方向の拡散ステップと同じになるように以下のように設定した。

$$p\left(x^{(0)}|x^{(1)}\right) = q\left(x^{(1)}|x^{(0)}\right) \frac{\pi\left(x^{(0)}\right)}{\pi\left(x^{(1)}\right)} = T_\pi\left(x^{(0)}|x^{(1)};\beta_1\right) \tag{42}$$

次に、この等価性を利用して、和の最初の時間ステップである

$$K = \sum_{t=2}^{T} \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log\left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t)}|x^{(t-1)}\right)}\right] + \int dx^{(0)} dx^{(1)} q\left(x^{(0)}, x^{(1)}\right) \log\left[\frac{q\left(x^{(1)}|x^{(0)}\right)\pi\left(x^{(0)}\right)}{q\left(x^{(1)}|x^{(0)}\right)\pi\left(x^{(1)}\right)}\right] - H_p\left(X^{(T)}\right) \tag{43}$$

$$= \sum_{t=2}^{T} \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log\left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t)}|x^{(t-1)}\right)}\right] - H_p\left(x^{(T)}\right) \tag{44}$$

の寄与を除去する。ここで再び、設計によって、$-\int dx^{(t)} q\left(x^{(t)}\right) \log \pi\left(x^{(t)}\right) = H_p(X^{(T)})$ がすべての t に対して定数であるという事実を利用した。

B.3 事後 $q\left(x^{(t-1)}\middle|x^{(0)}\right)$ で書き換える

前進軌道はマルコフ過程のため、以下のようになる。

$$K = \sum_{t=2}^{T} \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log\left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t)}|x^{(t-1)}, x^{(0)}\right)}\right] - H_p\left(X^{(T)}\right) \tag{45}$$

ベイズの法則を使えば、これを事後値と前方軌道からのマージナルで書き直すことができる。

$$K = \sum_{t=2}^{T} \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log\left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right)} \frac{q\left(x^{(t-1)}|x^{(0)}\right)}{q\left(x^{(t)}|x^{(0)}\right)}\right] - H_p\left(X^{(T)}\right) \tag{46}$$

B.4 KLダイバージェンスとエントロピーで書き直す

そして、いくつかの項が条件付きエントロピーであることを確認する。

$$K = \sum_{t=2}^{T} \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log\left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right)}\right] + \sum_{t=2}^{T} \left[H_q\left(X^{(t)}|X^{(0)}\right) - H_q\left(X^{(t-1)}|X^{(0)}\right)\right] - H_p\left(X^{(T)}\right) \tag{47}$$

$$= \sum_{t=2}^{T} \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log\left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right)}\right] + H_q\left(X^{(T)}|X^{(0)}\right) - H_q\left(X^{(1)}|X^{(0)}\right) - H_p\left(X^{(T)}\right) \tag{48}$$

最後に、確率分布の対数比をKLダイバージェンスに変換する。

$$\begin{aligned} K = &- \sum_{t=2}^{T} \int dx^{(0)} dx^{(t)} q\left(x^{(0)}, x^{(t)}\right) D_{KL}\left(q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right) || p\left(x^{(t-1)}|x^{(t)}\right)\right) \\ &+ H_q\left(X^{(T)}|X^{(0)}\right) - H_q\left(X^{(1)}|X^{(0)}\right) - H_p\left(X^{(T)}\right) \end{aligned} \tag{49}$$

エントロピーは解析的に計算でき、KLダイバージェンスは $x^{(0)}$ と $x^{(t)}$ が与えられれば解析的に計算できることに注意されたい。

C 摂動ガウス遷移

$\tilde{p}(x^{(t-1)}|x^{(t)})$ を計算する。表記を簡単にするために、$\mu = f_\mu\left(x^{(t)}, t\right)$ 、$\Sigma = f_\Sigma\left(x^{(t)}, t\right)$ 、$y = x^{(t-1)}$ とする。

$$\tilde{p}\left(y|x^{(t)}\right) \propto p\left(y|x^{(t)}\right) r(y) \tag{50}$$

$$= \mathcal{N}\left(y; \mu, \Sigma\right) r(y) \tag{51}$$

これをエネルギー関数で書き直すと、$E_r(y) = -\log r(y)$ とすると以下のようになる。

$$\tilde{p}\left(y|x^{(t)}\right) \propto \exp[-E(y)] \tag{52}$$

$$E(y) = \frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu) + E_r(y) \tag{53}$$

もし $E_r(y)$ が $\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)$ に対して滑らかであれば、μ 周りのテイラー展開を使って近似することができる。1つの十分条件は、$E_r(y)$のヘシアンの固有値が Σ^{-1} の固有値よりずっと小さいことである。ここで $g = \left.\frac{\partial E_r(y')}{\partial y'}\right|_{y'=\mu}$ とすると以下のようになる。

$$E_r(y) \approx E_r(\mu) + (y-\mu)g \tag{54}$$

これを全エネルギーに突っ込むと以下のようになる。

$$E(y) \approx \frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu) + (y-\mu)^T g + \text{constant} \tag{55}$$

$$= \frac{1}{2}y^T\Sigma^{-1}y - \frac{1}{2}y^T\Sigma^{-1}\mu - \frac{1}{2}\mu^T\Sigma^{-1}y + \frac{1}{2}y^T\Sigma^{-1}\Sigma g + \frac{1}{2}g^T\Sigma\Sigma^{-1}y + \text{constant} \tag{56}$$

$$= \frac{1}{2}(y-\mu+\Sigma g)^T\Sigma^{-1}(y-\mu+\Sigma g) + \text{constant} \tag{57}$$

これはガウシアンに相当する。

$$\tilde{p}(y|x^{(t)}) \approx \mathcal{N}(y; \mu - \Sigma g, \Sigma) \tag{58}$$

元の形式論に置き換えると以下のようになる。

$$\tilde{p}\left(x^{(t-1)}|x^{(t)}\right) \approx \mathcal{N}\left(x^{(t-1)}; f_\mu\left(x^{(t)}, t\right) + f_\Sigma\left(x^{(t)}, t\right) \left.\frac{\partial \log r\left(x^{(t-1)'}\right)}{\partial x^{(t-1)'}}\right|_{x^{(t-1)'}=f_\mu\left(x^{(t)},t\right)}, f_\Sigma\left(x^{(t)}, t\right)\right) \tag{59}$$

D 実験内容

D.1 玩具問題

D.1.1 ロールケーキ

二次元のスイスロール分布の確率論的モデルが構築された。生成モデル $p(x^{(0\cdots T)})$ は、同一性共分散ガウス分布で初期化されたガウス拡散の40時間ステップで構成されている。1つの隠れ層と16の隠れユニットを持つ（正規化された）放射基底関数ネットワークが、平均と共分散関数 $f_\mu(x^{(t)}, t)$ と、逆軌跡の対角 $f_\Sigma(x^{(t)}, t)$ を生成するために学習された。各関数の一番上の読み出し層は、各時間ステップで独立して学習されたが、他のすべての層については、すべての時間ステップと両方の関数にわたって重みが共有された。最上層の出力 $f_\Sigma(x^{(t)}, t)$ をシグモイドに通し、0と1の間に制限した。図1を見ればわかるように、スイスロール分布の学習に成功した。

D.1.2 バイナリーハートビート分布

長さ20の単純な2値系列で確率モデルを学習し、5番目の時間ビンごとに1が発生し、残りのビンは0である。生成モデルは、データと同じ平均活性を持つ独立二項分布 $\left(p(x_i^{(T)} = 1) = 0.2\right)$ で初期化された二項拡散の2000時間ステップで構成されている。シグモイド非線形を持つ多層パーセプトロンが学習され、20の入力ユニットと50ユニットずつの3つの隠れ層が、逆軌跡のベルヌーイレート $f_b(x^{(t)}, t)$ を生成した。一番上の読み出し層は時間ステップごとに独立して学習されたが、それ以外の層はすべての時間ステップで重みが共有された。最上層の出力は、0と1の間に制限す

るためにシグモイドに通された。図2からわかるように、ハートビート分布の学習は成功した。真の生成過程での対数尤度は $log_2(\frac{1}{5}) = -2.322$ ビット/シーケンスとなる。図2と表1を見ればわかるように、学習はほぼ完璧だった。

D.2 画像

D.2.1 構造

読み取り

すべての場合において、畳み込みネットワークは、各画像ピクセル i に対して出力 $y_i \in \mathcal{R}^{2J}$ のベクトルを生成するために使用された。

時間依存性

畳み込み出力 y^μ は、時間依存の「バンプ」関数の和における画素ごとの重み付け係数として使用され、画素iごとに出力 $z_i^\mu \in \mathcal{R}$ を生成する。

$$z_i^\mu = \sum_{j=1}^{J} y_{ij}^\mu g_j(t) \tag{60}$$

バンプ関数は、$\tau_j \in (0, T)$ をバンプ中心、w をバンプ中心の間隔とする

$$g_j(t) = \frac{\exp\left(-\frac{1}{2w^2}\left(t-\tau_j\right)^2\right)}{\sum_{k=1}^{J}\exp\left(-\frac{1}{2w^2}\left(t-\tau_k\right)^2\right)} \tag{61}$$

で構成される。z^Σ は同じ方法で生成されるが、y^Σ を使用する。すべての画像実験では、$T = 500$ を使用した樹皮データセットを除き、タイムステップ数 $T = 1000$ を使用した。

平均と分散

最後に、これらの出力を組み合わせて、各画素 i の拡散平均と分散予測を生成する。

$$\Sigma_{ii} = \sigma\left(z_i^\Sigma + \sigma^{-1}(\beta_t)\right) \tag{62}$$
$$\mu_i = \left(x_i - z_i^\mu\right)\left(1 - \Sigma_{ii}\right) + z_i^\mu \tag{63}$$

ここで、Σ と μ の両方は、順拡散カーネル $T_\pi\left(x^{(t)}|x^{(t-1)};\beta_t\right)$ の周りの摂動としてパラメータ化され、z_i^μ は、$p\left(x^{(t-1)}|x^{(t)}\right)$ を何度も適用した結果として生じる平衡分布の平均である。Σ は対角行列に制限される。

マルチスケール畳み込み

具体的には、学習データの長距離依存性やマルチスケール依存性を発見し、利用することである。しかし、ネットワークの出力は各ピクセルの係数のベクトルであるため、ダウンサンプルされた特徴マップではなく、完全な解像度の特徴マップを生成することが重要である。そこで、以下のステップからなるマルチスケールコンボリューションレイヤーを定義する。

1. ミーンプーリングを行って画像を複数のスケールにダウンサンプリングする。ダウンサンプリングは2の累乗で行う。
2. 各スケールでコンボリューションを実行する。
3. すべてのスケールをフル解像度にアップサンプリングし、得られた画像を合計する。
4. soft relu $\left(\log[1+\exp(\cdot)]\right)$ からなるポイントワイズ非線形変換を行う。

最初の3つの線形演算の合成は、アップサンプリングによってもたらされるブロッキングアーティファクトまでは、マルチスケールコンボリューションカーネルによる畳み込みに似ている。このマルチスケール畳み込みを実現する方法は[56]で述べられている。

緻密なレイヤー

密な（画像ベクトル全体に作用する）層とカーネル幅-1の畳み込み（各ピクセルの特徴ベクトルに別々に作用する）層は同じ形式を共有する。これらは線形変換と、それに続く $\tanh$ 非線形性で構成される。

参照

1. T, P. Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. J. Phys. A: Math. Gen. 15 1971, 1982. ↩
2. Tanaka, T. Mean-field theory of Boltzmann machine learning. Physical Review Letters E, January 1998. ↩

3. Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. Machine learning, 37(2):183–233, 1999. ↩
4. Welling, M. and Hinton, G. A new learning algorithm for mean field Boltzmann machines. Lecture Notes in Computer Science, January 2002. ↩
5. Hinton, G. E. Training products of experts by minimizing contrastive divergence. Neural Computation, 14(8): 1771–1800, 2002. ↩
6. Sohl-Dickstein, J., Battaglino, P. B., and DeWeese, M. R. Minimum Probability Flow Learning. International Conference on Machine Learning, 107(22):11–14, November 2011b. ISSN 0031-9007. doi: 10.1103/PhysRevLett.107.220601. ↩
7. Sohl-Dickstein, J., Battaglino, P., and DeWeese, M. New Method for Parameter Estimation in Probabilistic Models: Minimum Probability Flow. Physical Review Letters, 107(22):11–14, November 2011a. ISSN 0031-9007. doi: 10.1103/PhysRevLett.107.220601. ↩
8. Lyu, S. Unifying Non-Maximum Likelihood Learning Objectives with Minimum KL Contraction. Advances in Neural Information Processing Systems 24, pp. 64–72, 2011. ↩
9. Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477):359–378, 2007. ↩
10. Parry, M., Dawid, A. P., Lauritzen, S., and Others. Proper local scoring rules. The Annals of Statistics, 40(1):561–592, 2012. ↩
11. Hyv¨arinen, A. Estimation of non-normalized statistical models using score matching. Journal of Machine Learning Research, 6:695–709, 2005. ↩
12. Besag, J. Statistical Analysis of Non-Lattice Data. The Statistician, 24(3), 179-195, 1975. ↩
13. Murphy, K. P., Weiss, Y., and Jordan, M. I. Loopy belief propagation for approximate inference: An empirical study. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pp. 467–475. Morgan Kaufmann Publishers Inc., 1999. ↩
14. Gershman, S. J. and Blei, D. M. A tutorial on Bayesian nonparametric models. Journal of Mathematical Psychology, 56(1):1–12, 2012. ↩
15. Jarzynski, C. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. Physical Review E, January 1997. ↩ ↩
16. Neal, R. Annealed importance sampling. Statistics and Computing, January 2001. ↩ ↩
17. Hinton, G. E. The wake-sleep algorithm for unsupervised neural networks). Science, 1995. ↩
18. Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. The helmholtz machine. Neural computation, 7(5):889–904, 1995. ↩
19. Sminchisescu, C., Kanaujia, A., and Metaxas, D. Learning joint top-down and bottom-up processes for 3D visual inference. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, volume 2, pp. 1743–1752. IEEE, 2006. ↩ ↩
20. Kavukcuoglu, K., Ranzato, M., and LeCun, Y. Fast inference in sparse coding algorithms with applications to object recognition. arXiv preprint arXiv:1010.3467, 2010. ↩
21. Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. International Conference on Learning Representations, December 2013. ↩ ↩
22. Gregor, K., Danihelka, I., Mnih, A., Blundell, C., and Wierstra, D. Deep AutoRegressive Networks. arXiv preprint arXiv:1310.8499, October 2013. ↩
23. Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. Proceedings of the 31st International Conference on Machine Learning (ICML-14), January 2014. ↩
24. Ozair, S. and Bengio, Y. Deep Directed Generative Autoencoders. arXiv:1410.0630, October 2014. ↩
25. Bornschein, J. and Bengio, Y. Reweighted Wake-Sleep. International Conference on Learning Representations, June 2015. ↩
26. Bengio, Y. and Thibodeau-Laufer, E. Deep generative stochastic networks trainable by backprop. arXiv preprint arXiv:1306.1091, 2013. ↩ ↩
27. Yao, L., Ozair, S., Cho, K., and Bengio, Y. On the Equivalence Between Deep NADE and Generative Stochastic Networks. In Machine Learning and Knowledge Discovery in Databases, pp. 322–336. Springer, 2014. ↩
28. Larochelle, H. and Murray, I. The neural autoregressive distribution estimator. Journal of Machine Learning Research, 2011. ↩
29. Uria, B., Murray, I., and Larochelle, H. RNADE: The real-valued neural autoregressive density-estimator. Advances in Neural Information Processing Systems, 2013a. ↩
30. Uria, B., Murray, I., and Larochelle, H. A Deep and Tractable Density Estimator. arXiv:1310.1757, pp. 9, October 2013b. ↩
31. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Nets. Advances in Neural Information Processing Systems, 2014. ↩ ↩ ↩ ↩
32. Schmidhuber, J. Learning factorial codes by predictability minimization. Neural Computation, 1992. ↩
33. Rippel, O. and Adams, R. P. High-Dimensional Probability Estimation with Deep Density Models. arXiv:1410.8516, pp. 12, February 2013. ↩
34. Dinh, L., Krueger, D., and Bengio, Y. NICE: Non-linear Independent Components Estimation. arXiv:1410.8516, pp. 11, October 2014. ↩
35. Stuhlm¨uller, A., Taylor, J., and Goodman, N. Learning stochastic inverses. Advances in Neural Information Processing Systems, 2013. ↩
36. Theis, L., Hosseini, R., and Bethge, M. Mixtures of conditional Gaussian scale mixtures applied to multiscale image representations. PloS one, 7(7):e39857, 2012. ↩ ↩
37. MacKay, D. Bayesian neural networks and density networks. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 1995. ↩
38. Bishop, C., Svens´en, M., and Williams, C. GTM: The generative topographic mapping. Neural computation, 1998. ↩
39. Burda, Y., Grosse, R. B., and Salakhutdinov, R. Accurate and Conservative Estimates of MRF Log-likelihood using Reverse Annealing. arXiv:1412.8566, December 2014. ↩

40. Langevin, P. Sur la th´eorie du mouvement brownien. CR Acad. Sci. Paris, 146(530-533), 1908. ↩
41. Suykens, J. and Vandewalle, J. Nonconvex optimization using a Fokker-Planck learning machine. In 12th European Conference on Circuit Theory and Design, 1995. ↩
42. Feller, W. On the theory of stochastic processes, with particular reference to applications. In Proceedings of the [First] Berkeley Symposium on Mathematical Statistics and Probability. The Regents of the University of California, 1949. ↩ ↩
43. Spinney, R. and Ford, I. Fluctuation Relations : A Pedagogical Overview. arXiv preprint arXiv:1201.6381, pp. 3–56, 2013. ↩ ↩
44. Jarzynski, C. Equalities and inequalities: irreversibility and the second law of thermodynamics at the nanoscale. Annu. Rev. Condens. Matter Phys., 2011. ↩ ↩
45. Grosse, R. B., Maddison, C. J., and Salakhutdinov, R. Annealing between distributions by averaging moments. In Advances in Neural Information Processing Systems, pp. 2769–2777, 2013. ↩
46. Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Computer Science Department University of Toronto Tech. Rep., 2009. ↩ ↩
47. Jeulin, D. Dead leaves models: from space tesselation to random functions. Proc. of the Symposium on the Advances in the Theory and Applications of Random Sets, 1997. ↩ ↩
48. Lee, A., Mumford, D., and Huang, J. Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. International Journal of Computer Vision, 2001. ↩ ↩
49. Theis, L., van den Oord, A., and Bethge, M. A note on the evaluation of generative models. arXiv preprint arXiv:1511.01844, 2015. ↩
50. Lazebnik, S., Schmid, C., and Ponce, J. A sparse texture representation using local affine regions. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(8):1265–1278, 2005. ↩ ↩
51. Bergstra, J. and Breuleux, O. Theano: a CPU and GPU math expression compiler. Proceedings of the Python for Scientific Computing Conference (SciPy), 2010. ↩
52. Sohl-Dickstein, J., Poole, B., and Ganguli, S. Fast largescale optimization by unifying stochastic gradient and quasi-Newton methods. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 604–612, 2014. ↩
53. van Merri¨enboer, B., Chorowski, J., Serdyuk, D., Bengio, Y., Bogdanov, D., Dumoulin, V., and Warde-Farley, D. Blocks and Fuel. Zenodo, May 2015. doi: 10.5281/zenodo.17721. ↩
54. LeCun, Y. and Cortes, C. The MNIST database of handwritten digits. 1998. ↩
55. Bengio, Y., Mesnil, G., Dauphin, Y., and Rifai, S. Better Mixing via Deep Representations. arXiv preprint arXiv:1207.4404, July 2012. ↩
56. Barron, J. T., Biggin, M. D., Arbelaez, P., Knowles, D. W., Keranen, S. V., and Malik, J. Volumetric Semantic Segmentation Using Pyramid Context Features. In 2013 IEEE International Conference on Computer Vision, pp. 3448–3455. IEEE, December 2013. ISBN 978-1-4799-2840-8. doi: 10.1109/ICCV.2013.428. ↩