Deep Unsupervised Learning using Nonequilibrium Thermodynamics

12311207 董文芃

2025 年 6 月 3 日

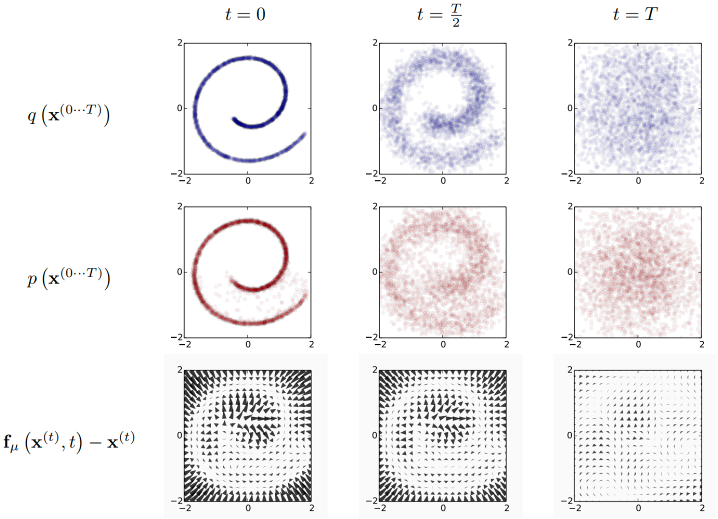
0.1. Background

- Probabilistic models suffer from a tradeoff between two conflicting objectives: **tractability** and **flexibility**.
- *Tractable models*: can be analytically evaluated and easily fit to data(e.g. a Gaussian or Laplace), but unable to aptly describe structure in rich datasets.
- *Flexible models*: can be molded to fit structure in arbitrary data, but evaluating, training, or drawing samples from such flexible models typically requires a very expensive Monte Carlo process.
- Inspired by non-equilibrium statistical physics, we develop an approach that simultaneously achieves **both** flexibility and tractability.
  - 1.Systematically and slowly destroy structure in a data distribution through an iterative forward diffusion process.
  - 2.Learn a reverse diffusion process that restores structure in data, yielding a highly flexible and tractable generative model of the data.

0.2. Intro

- Method:
  - We build a generative Markov chain which converts a simple known distribution (e.g. a Gaussian) into a target (data) distribution using a diffusion process. (explicitly define the probabilistic model as the endpoint of the Markov chain.)
  - Since each step in the diffusion chain has an analytically evaluable probability, the full chain can also be analytically evaluated.
  - Estimating small perturbations is more tractable than explicitly describing the full distribution with a single, non-analytically-normalizable, potential function.
  - Since a diffusion process exists for any smooth target distribution, this method can capture data distributions of arbitrary form
- Goal:
  - Define a forward diffusion process which converts any complex data distribution into a simple, tractable distribution
  - Learn a finite-time reversal of this diffusion process which defines our generative model distribution
- Advantages:
  - extreme flexibility in model structure
  - exact sampling
  - (easy multiplication with other distributions, e.g. in order to compute a posterior)
  - the model log likelihood, and the probability of individual states, to be cheaply evaluated

e.g. Swiss Roll



1.1 Process

1. Forward trajectory

The data distribution $q\left(x^{(0)}\right)$ is gradually converted into a well-behaved (analytically tractable) distribution $\pi(y)$ by repeated application of a Markov diffusion kernel $T_\pi(y|y';\beta)$ for $\pi(y)$, where β is the diffusion rate,

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

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

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

2. Reverse trajectory

The generative distribution will be trained to describe the same trajectory, but in reverse,

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

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

Feller, 1949. For continuous diffusion (small step size β) the reversal of the diffusion process has the identical functional form as the forward process

1.2. Model Probability

The probability the generative model assigns to the data is

$$p\left(\mathbf{x}^{(0)}\right) = \int d\mathbf{x}^{(1\cdots T)} p\left(\mathbf{x}^{(0\cdots T)}\right)$$

But the integral is intractable.
Inspired by **annealed importance sampling** and the **Jarzynski equality**, we instead compute:

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

$$= \int dx^{(1\cdots T)} q\left(x^{(1\cdots T)} \mid x^{(0)}\right) \frac{p\left(x^{(0\cdots T)}\right)}{q\left(x^{(1\cdots T)} \mid x^{(0)}\right)} \tag{7}$$

$$= \int dx^{(1\cdots T)} q\left(x^{(1\cdots T)} \mid x^{(0)}\right) \cdot \frac{p\left(x^{(T)}\right) \prod_{t=1}^{T} p\left(x^{(t-1)} \mid x^{(t)}\right)}{\prod_{t=1}^{T} q\left(x^{(t)} \mid x^{(t-1)}\right)}.$$

This can be evaluated rapidly by averaging over samples from the forward trajectory $q\left(x^{(1\cdots T)} \mid x^{(0)}\right)$.

1.3. Training

We want to maximize $p(x_0)$ when $x_0 \sim q(x_0)$, i.e. to maximize model log likelihood (equivalently, minimize the cross entropy) $E_{x_0 \sim q(x_0)}[log(p(x_0))]$

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

This reduces to $L \geq K$ (ELBO in variational inference)

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

Where the entropies and KL divergences can be analytically computed.

1.3. Training (cont.)

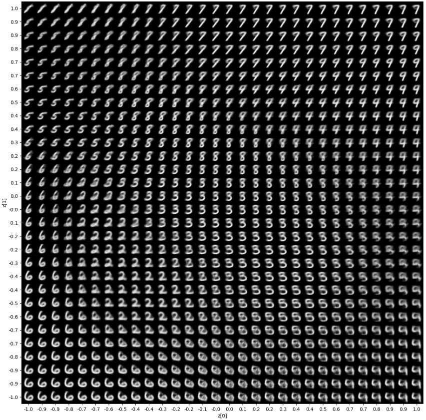
- Training consists of finding the reverse Markov transitions which maximize this lower bound on the log likelihood

$$\hat{p}\left(x^{(t-1)} \mid x^{(t)}\right) = \underset{p\left(x^{(t-1)} \mid x^{(t)}\right)}{\operatorname{argmax}} K$$

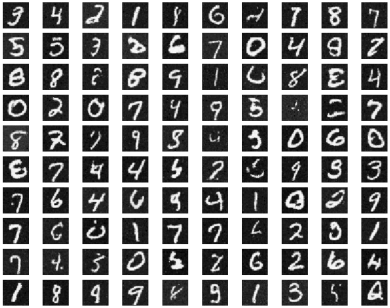
  Thus, the task of estimating a probability distribution has been reduced to the task of performing regression on the functions which set the mean and covariance of a sequence of Gaussians
- Remaning: Setting The Diffusion Rate β_t
  - **Gaussian diffusion**: learn the forward diffusion $\beta_{2\ldots T}$ by gradient ascent on K. The variance β_1 of the first step is fixed to a small constant to prevent overfitting. The dependence of samples from $q\left(x^{(1\cdots T)}|x^{(0)}\right)$ on $\beta_{1\ldots T}$ is made explicit by using *frozen noise* – as in the noise is treated as an additional auxiliary variable, and held constant while computing partial derivatives of K with respect to the parameters.
  - **Binomial diffusion**: Discrete state space, gradient ascent with frozen noise impossible. Instead choose the forward diffusion schedule $\beta_{1\ldots T}$ to erase a constant fraction $\frac{1}{T}$ of the original signal per diffusion step, yielding a diffusion rate of $\beta_t = (T - t + 1)^{-1}$.
  - Recent experiments suggest that it is just as effective for Gaussian diffusion to instead use the same fixed t schedule as for binomial diffusion

analytically tractable distribution	$\pi\left(x^{(T)}\right) = \begin{cases} \mathcal{N}\left(x^{(T)}; \mathbf{0}, \mathbf{I}\right) & \text{(Gauss)} \\ \mathcal{B}\left(x^{(T)}; 0.5\right) & \text{(Bin)} \end{cases}$
Forward diffusion kernel	$q\left(x^{(t)} \mid x^{(t-1)}\right) = \begin{cases} \mathcal{N}\left(x^{(t)}; x^{(t-1)}\sqrt{1-\beta_t}, \mathbf{I}\beta_t\right) & \text{(Gauss)} \\ \mathcal{B}\left(x^{(t)}; x^{(t-1)}(1-\beta_t) + 0.5\beta_t\right) & \text{(Bin)} \end{cases}$
Reverse diffusion kernel	$p\left(x^{(t-1)} \mid x^{(t)}\right) = \begin{cases} \mathcal{N}\left(x^{(t-1)}; f_\mu\left(x^{(t)}, t\right), f_\Sigma\left(x^{(t)}, t\right)\right) & \text{(Gauss)} \\ \mathcal{B}\left(x^{(t-1)}; f_b\left(x^{(t)}, t\right)\right) & \text{(Bin)} \end{cases}$
Training targets	$\begin{cases} f_\mu\left(x^{(t)}, t\right), f_\Sigma\left(x^{(t)}, t\right), \beta_{1\cdots T} & \text{(Gauss)} \\ f_b\left(x^{(t)}, t\right), \beta_{1\cdots T} & \text{(Bin)} \end{cases}$
Forward distribution	$q\left(x^{(0\cdots T)}\right) = q\left(x^{(0)}\right) \prod_{t=1}^{T} q\left(x^{(t)} \mid x^{(t-1)}\right)$
Reverse distribution	$p\left(x^{(0\cdots T)}\right) = \pi\left(x^{(T)}\right) \prod_{t=1}^{T} p\left(x^{(t-1)} \mid x^{(t)}\right)$
Log likelihood	$\mathcal{L} = \int \mathrm{d}x^{(0)} q\left(x^{(0)}\right) \log p\left(x^{(0)}\right)$
Lower bound on log likelihood	$\mathcal{K} = -\sum_{t=2}^{T} \mathbb{E}_{q\left(x^{(0)}, x^{(t)}\right)}\left[D_{\mathrm{KL}}\left(q\left(x^{(t-1)} \mid x^{(t)}, x^{(0)}\right) \,\|\, p\left(x^{(t-1)} \mid x^{(t)}\right)\right)\right] + H_q\left(X^{(T)} \mid X^{(0)}\right) - H_q\left(X^{(1)} \mid X^{(0)}\right) - H_p\left(X^{(T)}\right)$

e.g. MNIST



(a) VAE



(b) DIFFUSION

unlike many MNIST sample figures, diffusion-generated are true samples rather than the mean of the Gaussian or binomial distribution from which samples would be drawn.

Thank you for your listening.