Stochastic Localization and Diffusions

27 May 2025

Overview

- Diffusions are a successful technique to sample from high-dimensional distributions
- Stochastic localization is a unifying framework for sampling that generalises diffusion
- Stochastic localisation has a wider design space as compared to diffusions which only denoise Gaussians

Sampling from μ

We want to generate

$$x^* \sim \mu(dx) \quad \text{given} \quad \mu \in \mathcal{P}(\mathbb{R}^n),$$

- Non log-concave
- High-dimensional, $n \geq 100$.

Diffusion Model: OU Process

The Ornstein-Uhlenbeck process is commonly used in practice

$$dZ_s = -Z_s ds + \sqrt{2} dB_s$$

is distributed (conditioned on $Z_0 = x$)

$$Z_s \overset{d}{=} e^{-s}x + \sqrt{1 - e^{-2s}}G, \quad G \sim (0, I_n) \perp\!\!\!\perp x.$$

μ_s^Z converges exponentially fast to $\mu_\infty^Z = N(0, I_n)$.

Solving the reverse process

For the reverse process, with $\bar{Y}_0 \sim N(0, I_n)$,

$$d\bar{Y}_t = -\frac{1+t}{t(1+t)}\bar{Y}_t dt + \frac{1}{\sqrt{t(1+t)}} m(\sqrt{t(1+t)}\bar{Y}_t; t)dt + \frac{1}{\sqrt{t(1+t)}} dB_t,$$

where

$$m(y; t) = \mathbb{E}[x \mid tx + \sqrt{t}G = y], \quad (x, G) \sim \mu \otimes N(0, I_n)$$

Stochastic Localization

Goal: Sample $x^* \sim \mu$.

Intuition: Generate a stochastic process in $\mathcal{P}(\mathbb{R}^n)$ such that at each time $t \in [0, \infty)$, the random probability measure μ_t satisfies

- As $t \to \infty$, $\mu_t \Rightarrow \delta_{x^*}$

Alternatively: Think of the process as

1. Sample $x^* \sim \mu$
2. Observation Process: $(Y_t)_{t \geq 0}$ is a noisy observation of x^* which becomes 'more informative' as t increases
3. $\mu_t(x \in \cdot) = \mathbb{P}[x \in \cdot \mid Y_t]$

Simple Example of Stochastic Localization (Isotropic Gaussian)

Consider Y_t which is Gaussian defined as

$$Y_t = tx^* + W_t, \quad W_t \sim N(0, t).$$

Result in Stochstic Processes

Suppose μ has finite moment. Then, the process $(Y_t)_{t \geq 0}$ defined above is the unique solution of

$$dY_t = m(Y_t; t)dt + dB_t,$$

where $Y_0 = 0$, $(B_t)_{t \geq 0}$ is a standard BM and

$$m(y; t) = \mathbb{E}[x \mid tx + \sqrt{t}G = y], \quad (x, G) \sim \mu \otimes N(0, I_n).$$

Connection to OU Process

OU Process:

$$d\bar{Y}_t = -\frac{1+t}{t(1+t)}\bar{Y}_t dt + \frac{1}{\sqrt{t(1+t)}} m(\sqrt{t(1+t)}\bar{Y}_t; t)dt + \frac{1}{\sqrt{t(1+t)}} dB_t.$$

Isotropic Gaussian Stochastic Localisation:

$$dY_t = m(Y_t; t)dt + dB_t, \quad Y_0 = 0.$$

Connection:

$$Y_t = \sqrt{t(1+t)}\bar{Y}_t.$$

Why Can They Be The Same?

We have a process

$$m_t(y) = \mathbb{E}[x \mid y], \quad \frac{1}{t}y = x + \frac{1}{\sqrt{t}}g, \quad g \sim N(0, I_n),$$

$$m_t(\cdot) = \underset{\phi:\mathbb{R}^n \to \mathbb{R}^n}{\arg\min}\, \mathbb{E}[\|\phi(y) - x\|_2^2].$$

Why Can They Be The Same?

We have a process

$$m_t(y) = \mathbb{E}[x \mid y], \quad \frac{1}{t}y = x + \frac{1}{\sqrt{t}}g, \quad g \sim N(0, I_n),$$

$$m_t(\cdot) = \operatorname*{arg\,min}_{\phi:\mathbb{R}^n \to \mathbb{R}^n} \mathbb{E}[\|\phi(y) - x\|_2^2].$$

Remark

If we have an optimal denoiser for Gaussian noise, we have a sampler!

Estimate $m_t(\cdot)$ from data

$\mathbf{m}_t(\cdot)$

$$\begin{aligned} &\text{minimise } \mathbb{E}[\|\phi(y) - x\|_2^2] \\ &\text{subj. to } \phi : \mathbb{R}^n \to \mathbb{R}^n \text{ measurable.} \end{aligned}$$

Estimate $m_t(\cdot)$ from data

$\mathbf{m}_t(\cdot)$

$$\text{minimise } \mathbb{E}[\|\phi(y) - x\|_2^2]$$
$$\text{subj. to } \phi : \mathbb{R}^n \to \mathbb{R}^n \text{ measurable.}$$

Assume we have data $x_1, x_2, \cdots, x_N \sim_{iid} \mu$

$\widehat{\mathbf{m}}_t(\cdot)$: generate $y_1, y_2, \cdots, y_N$

$$\text{minimise } \frac{1}{N}\sum_{i=1}^{N} \|\phi(y_i) - x_i\|_2^2$$
$$\text{subj. to } \phi \in \mathcal{F} \text{ (function class).}$$

For example, if $x_1, \cdots, x_N$ are images, then

$$\begin{aligned} &\text{minimise } \frac{1}{N}\sum_{i=1}^{N} \|\phi(y_i) - x_i\|_2^2 \\ &\text{subj. to } \phi \in \text{ CNN} \end{aligned}$$

For example, if $x_1, \cdots, x_N$ are images, then

$$\begin{aligned} &\text{minimise } \frac{1}{N}\sum_{i=1}^{N} \|\phi(y_i) - x_i\|_2^2 \\ &\text{subj. to } \phi \in \text{ CNN} \end{aligned}$$

Diffusions!

Why Sampling Scheme is Important?

- Consider a mixture of 2 Gaussians in the form of

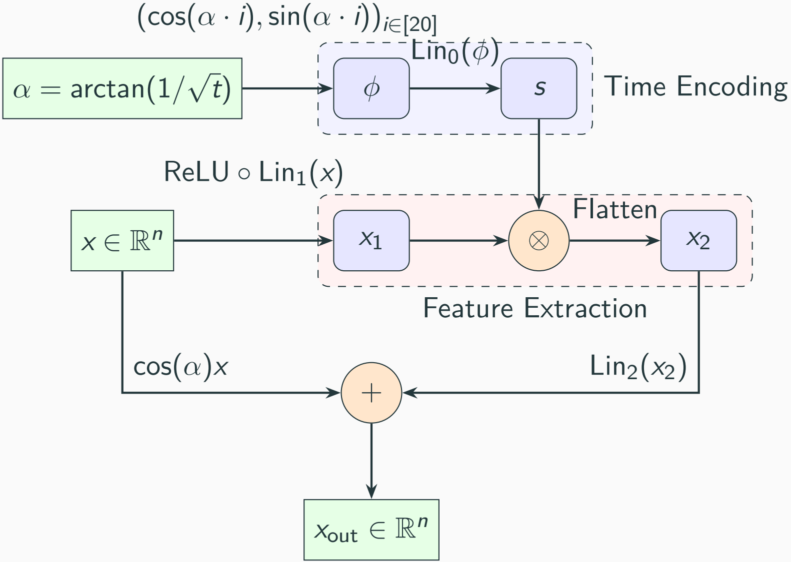
$$\mu = pN(a_1, I_n) + (1-p)N(a_2, I_n),$$

  where p is the weight, $a_1, a_2 \in \mathbb{R}^n$ are the means
- Assuming $pa_1 + (1-p)a_2 = 0$, we can rewrite it as

$$\mu = p \cdot N((1-p)a, I_n) + (1-p) \cdot N(-pa, I_n),$$

  where $a = a_1 - a_2$

2-Layer Fully Connected Denoiser Architecture



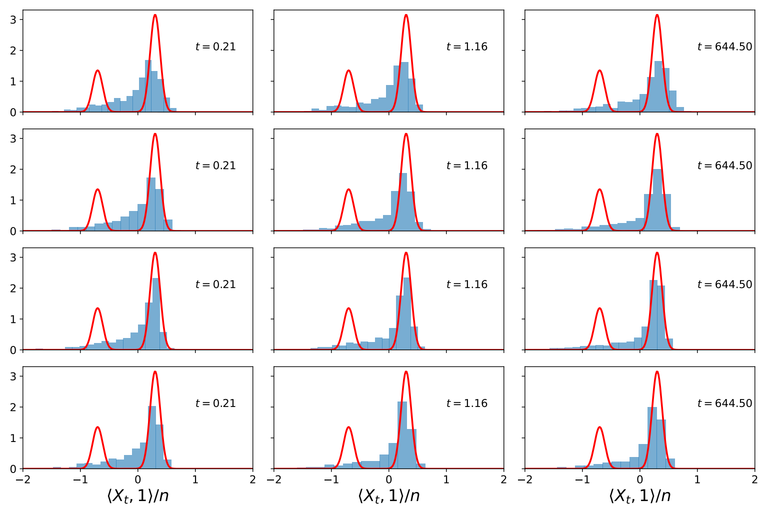


Figure 1: Sampling from the trained network and projecting onto α.

A Slightly Different Model

- v is the principle eigenvector of the covariance of the dataset X
- $p =$ fraction of datapoints such that $\langle x, v \rangle \geq 0$
- 2 models m_+, m_- trained with the same architecture as before
- m_+ trained on $x's$ such that $\langle x, v \rangle \geq 0$, and m_- trained of $x's$ with $\langle x, v \rangle < 0$
- Sample from $pm_+ + (1 - p)m_-$

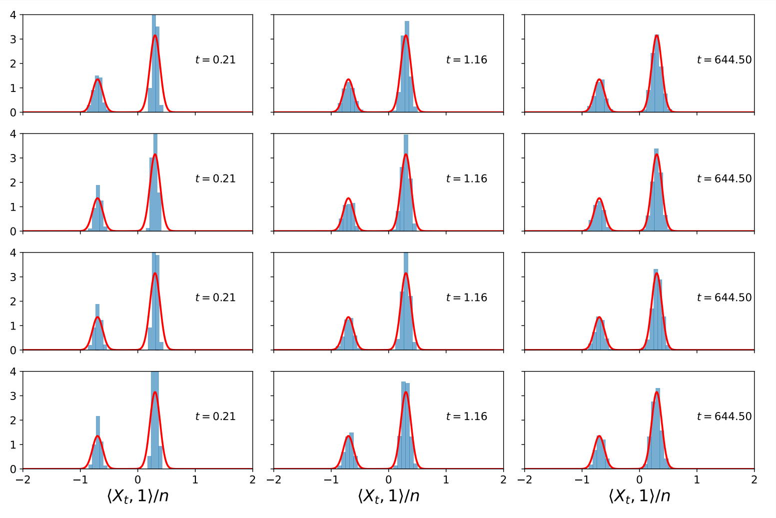


Figure 2: Sampling from a mixture of 2 trained network and projecting onto α.

Why the Difference?

The posterior mean

$$m(y;t) = \frac{y}{1+t} + \phi(a, y; t),$$

where ϕ can be well-approximated by a mixture of 2 ReLU network with 1 hidden layer. On the other hand, when $n \to \infty$, the posterior mean becomes sensitivity in the direction of a, which causes the accuracy of the 1 model to be less efficient.

General Stochastic Localization

Given $x \sim \mu$, let $(Y_t)_{t \in I}$ be a sequence of random variables indexed by $I \subset [0, \infty)$.

Definition

Observation Process $(Y_t)_{t \in I}$ is an observation process with respect to x if for each integer k and every $t_1 < t_2 < \cdots < t_k \in I$, the sequence of random variables $x, Y_{t_k}, Y_{t_{k-1}}, \cdots, Y_{t_1}$ forms a Markov chain. i.e.

$$\mathbb{P}[Y_{t_{i-1}} \in \cdot \mid x, Y_{t_i}, \cdots, Y_{t_k}] = \mathbb{P}[Y_{t_{i-1}} \in \cdot \mid Y_{t_i}].$$

Definition

Stochastic Localization process (scheme) Given an observation process $(Y_t)_{t\in I}$, the stochastic localization process $(\mu_t)_{t\in I}$ is defined to be

$$\mu_t(\cdot) = \mathbb{P}[x \in \cdot \mid Y_t].$$

1. We assume that the whole path $(Y_t)_{t\in I}$ gives complete information about x. In other words, for any $A \subset \mathbb{R}^n$,

$$\mu_\infty(A) := \mathbb{P}[x \in A \mid Y_t, t \in I] \in \{0, 1\}$$

2. $\lim_{t\to\infty} \mu_t(A)$ exists almost surely by Levy's martingale convergence theorem
3. Since $\mu_\infty(A) \in \{0, 1\}$ for all A, then $\mu_\infty(A) = 1_{x\in A}$

Constructing the Algorithm

Remark

Since $x, Y_{t_k}, Y_{t_{k-1}}, \cdots, Y_0$ forms a Markov chain, so is the reverse sequence $Y_0, Y_{t_1}, \cdots, Y_{t_k}, x$.

Consequently, there is transition probabilities $\mathbb{P}_{t,t'}[y \mid A] = \mathbb{P}[Y_{t'} \in A \mid Y_t = y]$.

1. Discretize the time index set to $I_m = (t_0, t_1, \cdots, t_m)$
2. Construct approximate kernels $\hat{\mathbb{P}}_{t_k,t_{k+1}}[y_k \mid \cdot] \approx \mathbb{P}_{t_k,t_{k+1}}[y_k \mid \cdot\,]$
3. For each $k \in [m]$, sample

$$y_{k+1} \sim \hat{\mathbb{P}}_{t_k,t_{k+1}}[y_k \mid \cdot\,]$$

Examples of Sampling Schemes (TBC)

1. $Y_t = tx^* + W_t$
2. $Y_t = \int_0^t Q(s)x^* ds + \int_0^t Q(s)^{1/2} dW_s$
3. For each $i \in [n]$, let $T_i \sim \text{Unif}\left([0,1]\right)$ and set

$$Y_{t,i} = \begin{cases} x_i & \text{if } t \geq T_i \\ * & \text{if } t < T_i \end{cases}$$

4. If $x \in \{\pm 1\}^n$, let $Y_t = x \odot Z_t$, where $\odot$ is the Hadamard product and $(Z_t)_{t\in[0,1]}$ is a suitable noise process in $\{\pm 1\}^n$

5. Fix matrix $A \in \mathbb{R}^{m \times n}$, $Y_t = tAx + B_t$
6. Suppose $x \in \mathbb{R}^n_{\geq 0}$, let $Y_t \in \mathbb{N}^n$ have coordinates conditionally independent given x, and $(Y_{t,k})_{t \geq 0}|_x \sim PPP(x_k dt)$ is a Poisson Point Process with rate x_k