Statistical Methods

Samy Braik

We denote by p the target distribution and q an easy-to-sample distribution, for example a centered Gaussian.

Diffusion

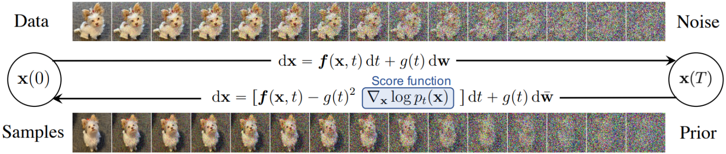
Let $X_0 \sim p$. We want to add noise until we reach pure noise, and denoise it afterward. We choose an horizon of time $T \in \mathbb{N}^*$ and a noise schedule $\beta : [0, T] \to \mathbb{R}^*$, continuous and non decreasing.

Forward process

$$d\overrightarrow{X}_t = \frac{-\beta(t)}{2\sigma^2}\overrightarrow{X}_t dt + \sqrt{\beta(t)}dB_t, \quad \overrightarrow{X}_0 \sim p$$

Backward process

$$d\overleftarrow{X}_t = \left(\frac{\beta(T-t)}{2\sigma^2}\overleftarrow{X}_t + \beta(T-t)\nabla \log p_{T-t}\left(\overleftarrow{X}_t\right)\right)dt$$
$$+ \sqrt{\beta(T-t)}dB_t, \quad \overleftarrow{X}_0 \sim p_T$$



We learn the score by using score-matching techniques

Score matching

$$\mathcal{L}_{\mathsf{score}}(\theta) = \mathbb{E}\left[\left\| s_\theta\left(\tau, \overrightarrow{X}_\tau\right) - \log p_\tau\left(\overrightarrow{X}_\tau | X_0\right)\right\|^2\right]$$

Plug it in the backward process and generate by discretizing the dynamics.

Let $X_0 \sim q$ and $f : \mathbb{R}^d \to \mathbb{R}^d$ an invertible and differentiable function an set $X_1 := f(X_0)$ such that $X_1 \sim p$.
We can write the density of X_1 as

$$p_{X_1} = p_{X_0}(f^{-1}(x_1)) \left| \det \frac{\partial f^{-1}}{\partial x_1}(x_1) \right| \tag{1}$$

$$= p_{X_0}(f^{-1}(x_1)) \left| \det \frac{\partial f}{\partial x_0}(f^{-1}(x_1)) \right|^{-1} \tag{2}$$

We can then write the log-likelihood as

$$\log p_{X_1}(x_1) = \log p_{X_0}(f^{-1}(x_1)) - \log \left| \det \frac{\partial f}{\partial x_0}(f^{-1}(x_1)) \right| \tag{3}$$

Normalizing flow

Let $X_0 \sim q$ and $X_1 \sim p$. We want to learn f_θ such that $X_1 \simeq f_\theta(X_0) = Z \sim p_Z$. To do that, we set a structure on f_θ, with $f_1, \ldots, f_k$ simpler function (all parametrized by θ) such that

$$f_\theta = f_1 \circ f_2 \circ \ldots \circ f_k$$

We determine f_θ by minimizing

$$\mathcal{L}_{\mathsf{NF}}(\theta) = \mathbb{E}\left[-\log p_Z(f_\theta(x)) - \log \left|\det \frac{\partial f_\theta}{\partial x}(x)\right|\right]$$

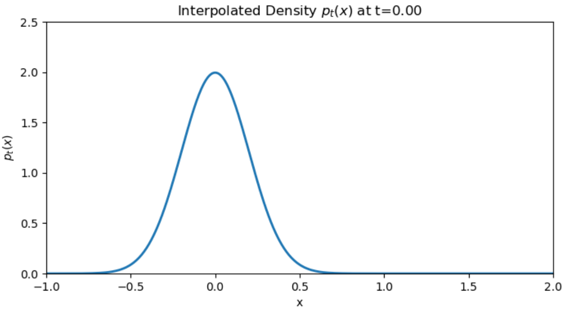
We start by defining a probability density path

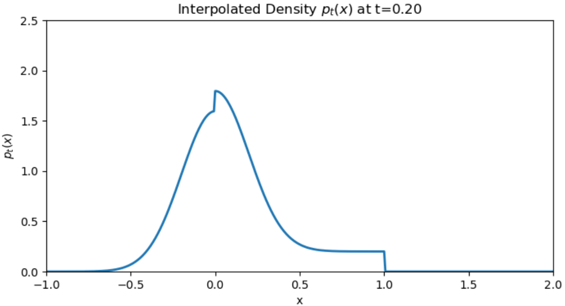
Probability density path

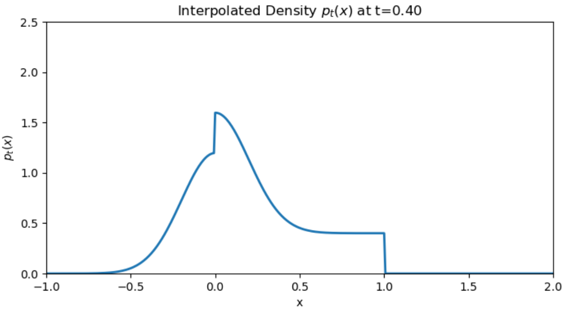
A probability path $p : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ meaning that for each time t, p_t is density function i.e. $\int p_t(x)dx = 1$.

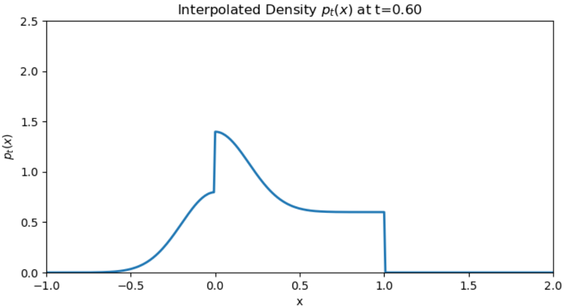
A simple example of such a path is a path p interpolating two density p_0 and p_1 with $p_t = tp_1 + (1-t)p_0$

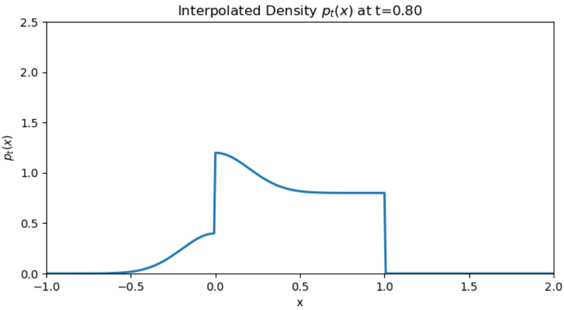
Figure: A probability path interpolating $\mathcal{N}(0, 0.2)$ and $\mathcal{U}([0, 1])$

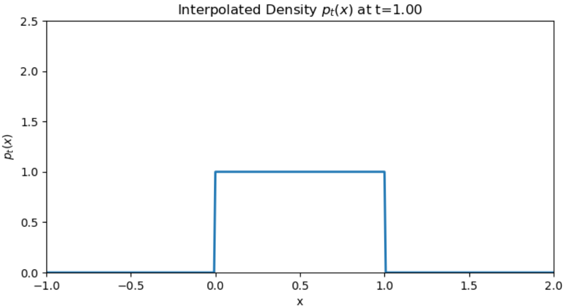

Interpolated Density $p_t(x)$ at t=0.00
$p_t(x)$
x


Interpolated Density $p_t(x)$ at t=0.20
$p_t(x)$
x


Interpolated Density $p_t(x)$ at t=0.40
$p_t(x)$
x
2.5
2.0
1.5
1.0
0.5
0.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


Interpolated Density $p_t(x)$ at t=0.60
$p_t(x)$
x
2.5
2.0
1.5
1.0
0.5
0.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


Interpolated Density $p_t(x)$ at t=0.80
2.5
2.0
1.5
1.0
0.5
0.0
$p_t(x)$
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
x


Interpolated Density $p_t(x)$ at t=1.00
2.5
2.0
1.5
1.0
0.5
0.0
$p_t(x)$
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
x

Flow matching

The framework is as follow :
We define a probability path p_t interpolating from $p_0 = q$ to $p_1 = 1$. Then we learn a velocity field v_t^θ generating the path p_t by minimizing the flow matching loss

$$\mathcal{L}_{\mathsf{FM}}(\theta) := \mathbb{E}[\|v_t^\theta(X_t) - \dot{X}_t\|^2] + c \tag{4}$$

Finally, we can sample from p_1 by solving the ODE (**??**) with the learned velocity field v_t^θ and the initial condition $X_0 \sim q$.

Comparison

Models	Pros	Cons
Diffusion	1.2	
Normalizing flow	Exact density estimation	Computationaly inter
Flow matching	Exact density estimation Simulation free training	test
Kernel estimator	Flexible Easy to exploit	Slow rate of converg Hard to evaluate at new d Hard to choose tuning pa