Literature Review: Score-Based Generative Modeling through Stochastic Differential Equations (Song et al. 2021)

PhD Students: Parley Ruogu Yang & Georgios Batzolis
Postdoc: Dr Christian Etmann
Big boss: Prof Carola-Bibiane Schönlieb

UNIVERSITY OF CAMBRIDGE
Faculty of Mathematics

30 March 2021

Key references:

- Anderson, Brian D O (1982). *Reverse-time diffusion equation models*, Stochastic Process. Appl., 12(3): pp.313–326
- Sarkka, Simo and Arno Solin (2019). *Applied stochastic differential equations*, CUP
- Song, Yang et al. (2021). *Score-Based Generative Modeling through Stochastic Differential Equations*, arXiv:2011.13456v2

Other literature are mentioned with arXiv ref attached, as we proceed.

Figure 14: Extended intermediate samples from annealed Langevin dynamics for CelebA.

Figure: Song and Ermon (2020). See arXiv:1907.05600

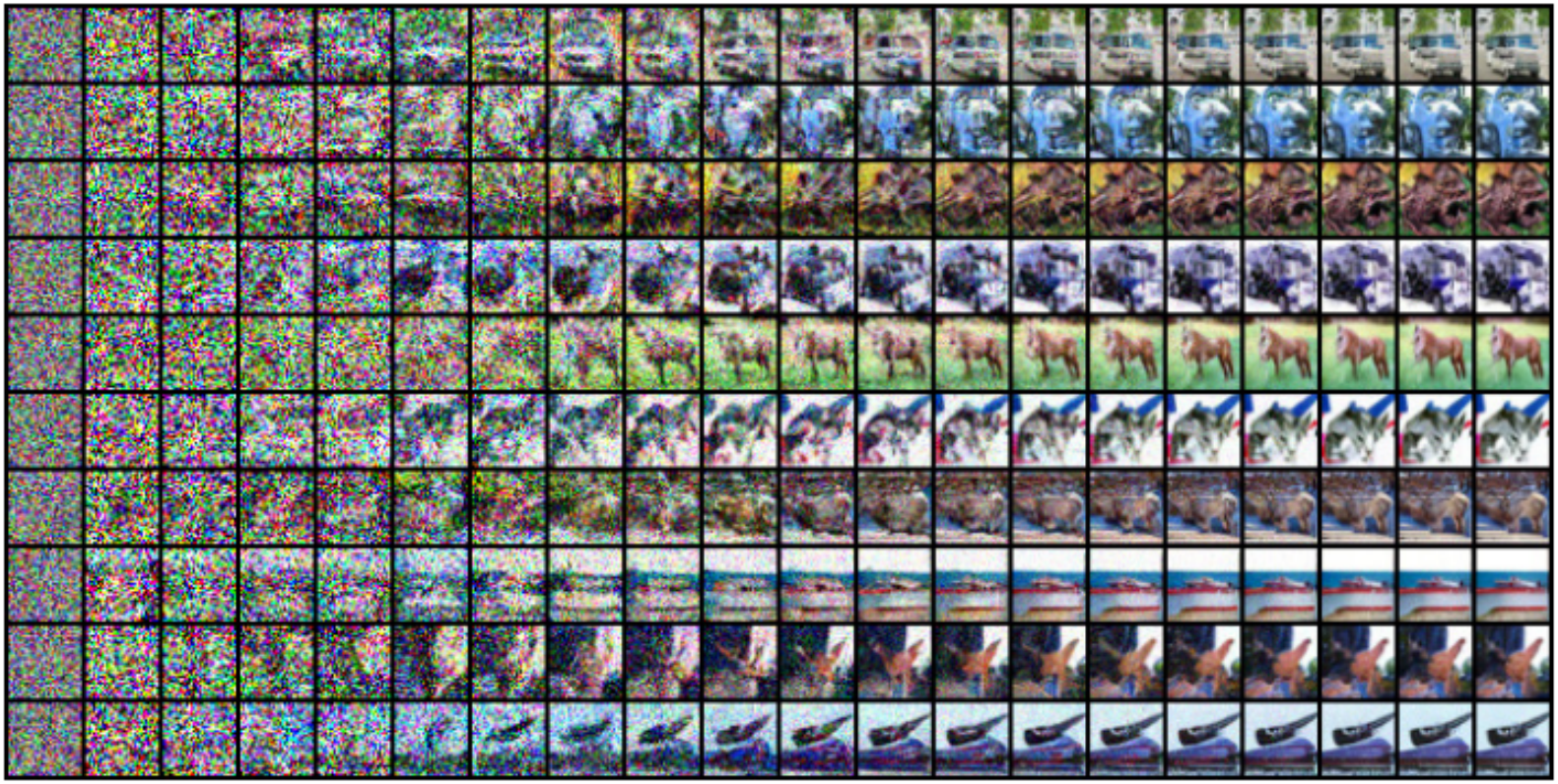


Figure 15: Extended intermediate samples from annealed Langevin dynamics for CelebA.

Figure: Song and Ermon (2020). See arXiv:1907.05600

A mathematical introduction

Consider $x_1, ..., x_n$ drawn from $p_0(\mathbb{R}^d)$ where $d >> n$ and p_0 unknown. Ultimate aim: know more about p_0 and be able to draw from a similar distribution.

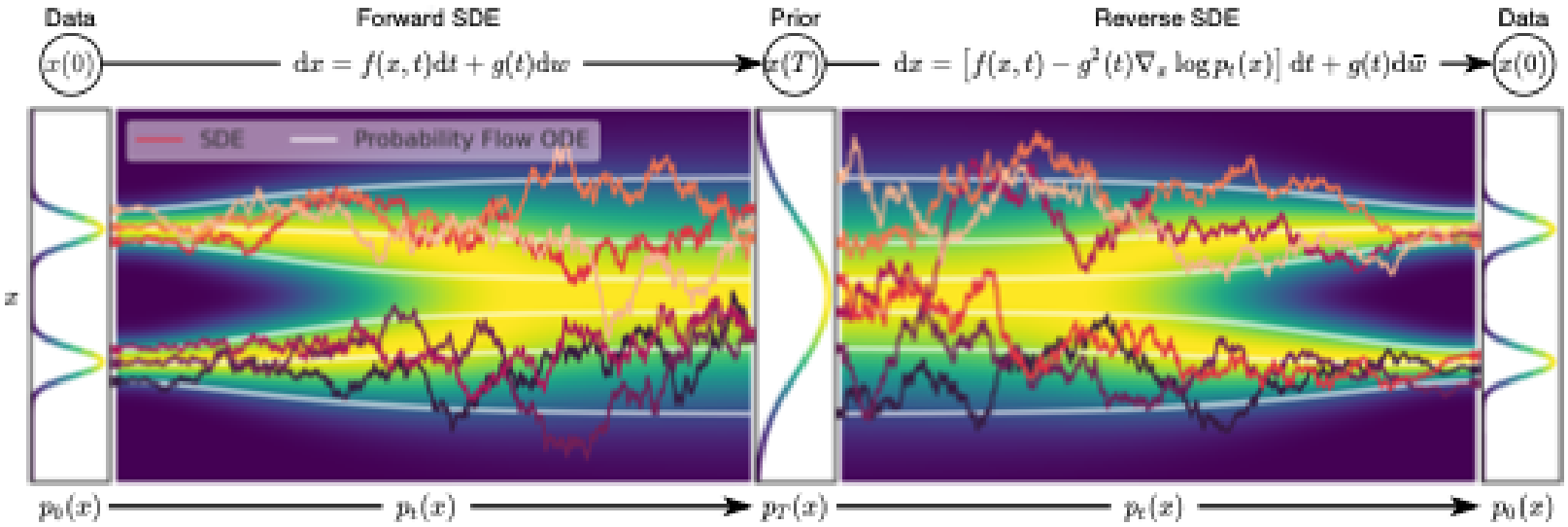


Figure 2: **Overview of score-based generative modeling through SDEs**. We can map data to a noise distribution (the prior) with an SDE (Section 3.1), and reverse this SDE for generative modeling (Section 3.2). We can also reverse the associated probability flow ODE (Section 4.3), which yields a deterministic process that samples from the same distribution as the SDE. Both the reverse-time SDE and probability flow ODE can be obtained by estimating the score $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ (Section 3.3).

Figure: Song et al. (2021).

Whiteboard: further introduction

A factory-like concept: $(D, \mathit{Methods}) \mapsto \tilde{p_0}$
Guarantee from the theory: if D drawn from p_0, with Anderson (1982) and LLN, $\tilde{p_0} \longrightarrow p_0$ as $|D| \to \infty$

Whiteboard: Practical implementation (Forward)

Forward SDE:

$$\hat{p_0}(x_1(0), ..., x_n(0)) \mapsto \hat{p_T}(x_1(T), ..., x_n(T))$$

Design the method such that $\hat{p}(x(T)|x(0)) \approx N(0, I)$, e.g. Song et al. (2021, p.16)

Whiteboard: Practical implementation (Backward)

Then for a desired number of samples m, draw iid samples
$y_1(T), ..., y_m(T) \sim N(0, I)$
and get iid $y_1(0), ..., y_m(0) \sim \tilde{p_0}$

Utilise **the design** to facilitate backward SDE such that

$$\hat{p_T}(y_1(T), ..., y_m(T)) \mapsto \tilde{p_0}(y_1(0), ..., y_m(0))$$

Key takeaways

- Reverse SDE (Anderson 1982)
- More about $p(x(t)|x(0))$ (Sarkka and Solin 2019)

Whiteboard: Definition of reverse SDE

- 'Reverse white-noise'
- Meaning of reverse-time Ito equation

Anderson Theorem in the context of scalar $g : [0, T] \to \mathbb{R}$

Consider forward Ito

$$dx = f(x, t)dt + g(x, t)dw$$

Theorem. *Let x_t be the process described by (3.3), and suppose $f(\cdot, \cdot)$ and $g(\cdot, \cdot)$ are such as to guarantee the existence of the probability density $p(x_t, t)$ for $t_0 \leqslant t \leqslant T$ as a smooth and unique solution of its associated Kolmogorov equation. Suppose further that an r-vector process $\bar{w}_t$ is defined by $\bar{w}_{t_0} = 0$ and*

$$\mathrm{d}\bar{w}_t^k = \mathrm{d}w_t^k + \frac{1}{p(x_t, t)} \sum_j \frac{\partial}{\partial x_t^j} [p(x_t, t) g^{jk}(x_t, t)] \, \mathrm{d}t, \tag{3.10}$$

and that the forward Kolmogorov equation associated with the joint process $(x_t, \bar{w}_t)$ yields a smooth and unique solution in $t > t_0$ for $p(x_t, \bar{w}_t, t)$ and in $t > s \geqslant t_0$ for $p(x_t, \bar{w}_t, t \mid \bar{w}_s, s)$. Then

(i) *x_t and $\bar{w}_t - \bar{w}_s$ are independent for all $t \geqslant s \geqslant t_0$.*

(ii) *With $\bar{\mathcal{A}}_t$ the minimal σ-algebra with respect to which x_s for $s \geqslant t$ and $\bar{w}_s$ for $s \geqslant t$ are measurable, conditions (3.4) and (3.5) hold.*

(iii) *A reverse time model for x_t is defined by*

$$\mathrm{d}x_t = \bar{f}(x_t, t) \, \mathrm{d}t + g(x_t, t) \, \mathrm{d}\bar{w}_t \tag{3.11}$$

where

$$\bar{f}^i(x_t, t) = f^i(x_t, t) - \frac{1}{p(x_t, t)} \sum_{j,k} \frac{\partial}{\partial x_t^j} [p(x_t, t) g^{ik}(x_t, t) g^{jk}(x_t, t)]. \tag{3.12}$$

Figure: Anderson (1982)

Key desire

We want

$$dx = [f(x,t) - g(t)^2 \nabla_x \log(p_t(x))]dt + g(t)d\overline{w}$$

All we need now is

- $\nabla_x \log(p_t(x))$
- Alternative: $\nabla_x \log(p(x(t)|x(0)))$

Whiteboard: more about $p(x(t)|x(0))$

Handover slide

Training Objective

Goal: Estimate $\nabla_{\mathbf{x}(t)} \log p_t(\mathbf{x}(t))$ using a neural network $s_\theta(\mathbf{x}(t), t)$

Train a score-based model $s_\theta(\mathbf{x}(t), t)$ by minimising:

$$\theta^* = \arg\min_\theta \mathbb{E}_t \left\{ \lambda(t) \mathbb{E}_{\mathbf{x}(0)} \mathbb{E}_{\mathbf{x}(t)|\mathbf{x}(0)} \left[\left\| s_\theta(\mathbf{x}(t), t) - \nabla_{\mathbf{x}(t)} \log p_{0t}(\mathbf{x}(t)|\mathbf{x}(0)) \right\|_2^2 \right] \right\}$$

where $\lambda : [0, T] \to \mathbb{R}_{>0}$

With sufficient data and model capacity: $s_{\theta^*}(\mathbf{x}(t), t) \to \nabla_{\mathbf{x}(t)} \log p_t(\mathbf{x}(t))$

How to choose $\lambda(t)$?

In this paper: empirical choice $\lambda(t) \propto 1/\mathbb{E}[\|\nabla_x \log p_{0t}(x(t)|x(0))\|_2^2]$

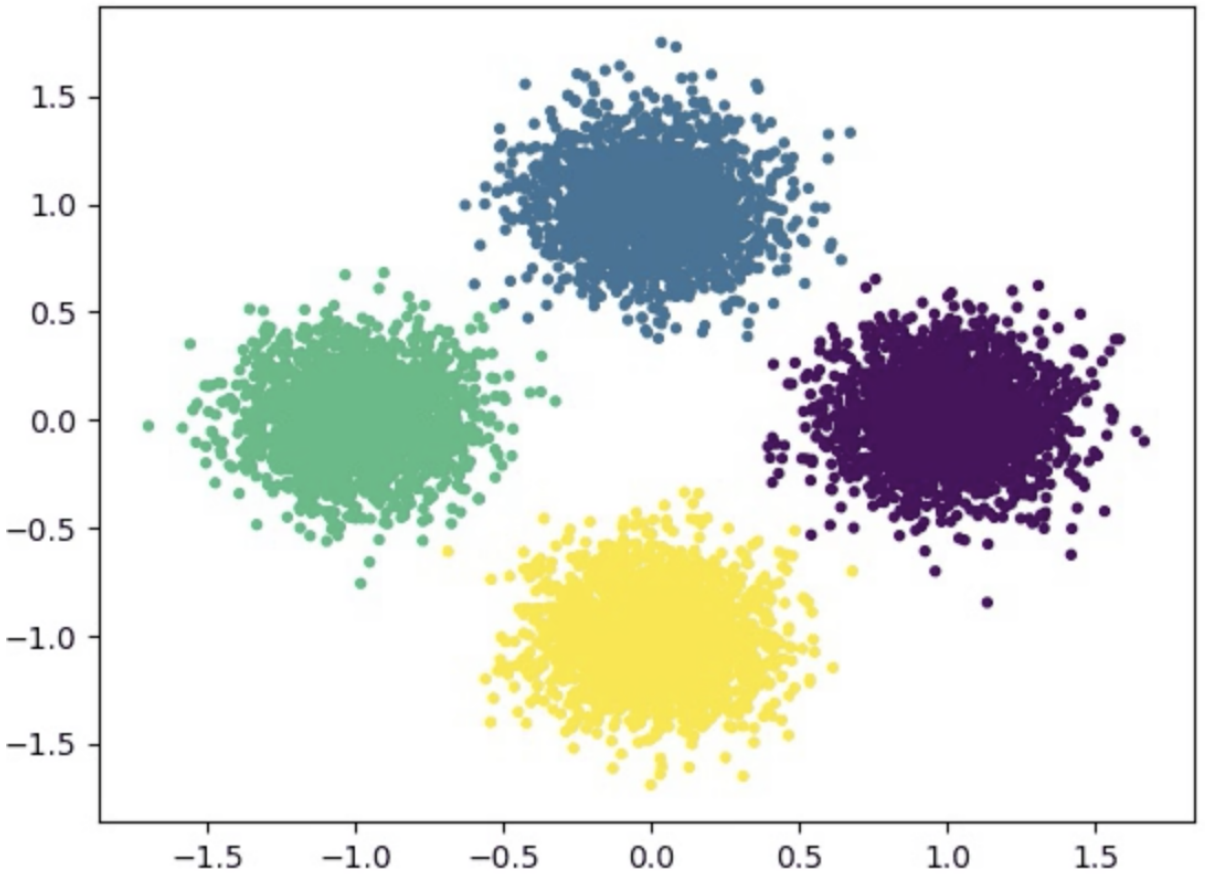
In the follow-up paper: $\lambda(t) = g(t)^2$

Theorem 1. *Consider two continuous distributions p and q over $\mathbb{R}^D$. Let $\{\mathbf{x}(t)\}_{t\in[0,T]}$ be a stochastic process defined by the SDE in Eq. (1). We use p_t and q_t to denote the distributions of $\mathbf{x}(t)$ when $\mathbf{x}(0) \sim p$ and $\mathbf{x}(0) \sim q$ respectively. Assuming $\log p_t(\mathbf{x})$ and $\log q_t(\mathbf{x})$ are smooth functions which have at most polynomial growth at infinity, we have*

$$D_{\mathrm{KL}}(p \parallel q) = \frac{1}{2}\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t(\mathbf{x})}[g(t)^2 \,\|\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) - \nabla_{\mathbf{x}} \log q_t(\mathbf{x})\|_2^2]\,\mathrm{d}t. \tag{5}$$

DEMO

Model a simple 2D synthetic distribution: equal mixture of 4 Gaussians



DEMO

VE SDE:

$$d\mathbf{x} = \sigma(t)d\mathbf{w}$$

where $\sigma(t) = \sigma_{min}(\frac{\sigma_{max}}{\sigma_{min}})^t\sqrt{2\log(\frac{\sigma_{max}}{\sigma_{min}})},\ t \in (\epsilon, 1]$

$\sigma_{min} = 0.01 \quad \& \quad \sigma_{max} = 3.5$

Perturbation kernel :

$$p_{0t}(\mathbf{x}(t)|\mathbf{x}(0)) = \mathcal{N}(\mathbf{x}(t); \mathbf{x}(0), \sigma_{min}^2\left(\frac{\sigma_{max}}{\sigma_{min}}\right)^{2t}\mathbb{I})$$

"Prior" Distribution: $p_{\mathbf{x}(1)}(\mathbf{x}(1)) \simeq \mathcal{N}(\mathbf{0}, \sigma_{max}^2\mathbb{I})$

Future Work

Further exploration of the applications of score-based modeling in:

- Inverse problems, e.g. $p(y|x)$
- Finance applications, e.g. $p(\mathbb{R}^d)$ with large d and large high-order momentums