

MAT2000 project: The mathematics of Score-Based Generative Models

Supervision: Håkon Hoel

December 10, 2025

1 Background and outline of the problem

Generative artificial intelligence has recently seen a surge of interest with the introduction of diffusion models. While popularly known for generating images, the theoretical foundation of these models rests heavily on the theory of Stochastic Differential Equations (SDEs) and time-reversal of diffusion processes. The seminal work by Song et al. [1] unifies these methods into a single framework known as Score-Based Generative Modeling (SBGM).

The core idea is to define a *forward* SDE that gradually perturbs a data distribution $X_t \sim p_t(x)$ towards a known prior distribution $\pi(x)$ (typically a standard Gaussian) over a time interval $t \in [0, T)$ (ideally with $T = \infty$). This process is given by:

$$d\mathbf{X}_t = \mathbf{f}(\mathbf{X}_t, t)dt + g(t)d\mathbf{W}_t,$$

where $\mathbf{f}$ is a drift coefficient, g is a diffusion coefficient and $\mathbf{W}_t$ is a d -dimensional standard Brownian motion. As $T \rightarrow \infty$, the distribution of $\mathbf{X}_T$, p_T approaches $\pi(x)$.

The generative process is realized by solving the *reverse-time* SDE. A remarkable result from Anderson [2] states that the reverse process is also a diffusion process, governed by:

$$d\mathbf{X}_t = [\mathbf{f}(\mathbf{X}_t, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{X}_t)]dt + g(t)d\mathbf{B}_t,$$

where time flows backwards from T to 0, and $d\mathbf{B}_t$ is another standard, d -dimensional Brownian motion (independent of $\mathbf{W}$). The critical term here is the *score function*, $\nabla_{\mathbf{x}} \log p_t(\mathbf{X}_t)$, which represents the vector field pointing towards high-density regions of the distribution at time t .

In practical applications, this score function is unknown and is approximated using neural networks (score matching). However, for understanding the theoretical properties of these generative models, one can study cases where the score is known analytically or approximated via simple numerical methods. This allows for the study of the generative process as a pure numerical integration problem of SDEs, separating the stochastic analysis errors from the function approximation errors.

2 Project goals

The project is divided into theoretical study and numerical experimentation.

1. **Theoretical Foundations:** derive the reverse-time SDE formulation as introduced by Anderson [2] and utilized by Song et al. [1]. Specifically, detail the relationship between the forward diffusion process, the Fokker-Planck equation, and the resulting score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$.
2. **Implementation of a Toy Model:** Implement a numerical solver for the reverse SDE in Matlab, Julia, or Python. To isolate numerical errors from approximation errors, focus on a target distribution p_{data} defined as a **Gaussian Mixture Model (GMM)** in $\mathbb{R}^d$ for some $d \leq 10$. In this setting, derive and implement the *exact* analytical score function, removing the need for neural network training.
3. **Numerical Analysis and Convergence:** Use the Euler-Maruyama scheme to generate samples from the reverse process. Quantify the quality of the generated distribution by estimating the **Weak Error** or the **Kullback-Leibler (KL) divergence** between the sampled distribution and the ground truth. Produce convergence plots showing error as a function of the time-step size Δt and possibly also of the value of cutoff time T .

References

- [1] Song, Yang, et al. "Score-Based Generative Modeling through Stochastic Differential Equations." *International Conference on Learning Representations (ICLR)*, 2021.
- [2] Anderson, Brian DO. "Reverse-time diffusion equation models." *Stochastic Processes and their Applications* 12.3 (1982): 313-326.
- [3] Hyvärinen, Aapo. "Estimation of non-normalized statistical models by score matching." *Journal of Machine Learning Research* 6 (2005): 695-709.
- [4] Higham, Desmond, and Peter Kloeden. *An introduction to the numerical simulation of stochastic differential equations*. Society for Industrial and Applied Mathematics, 2021.