Malliavin-Bismut Score Function: Linear Case

Ehsan Mirafzali
Daniele Venturi
Razvan Marinescu

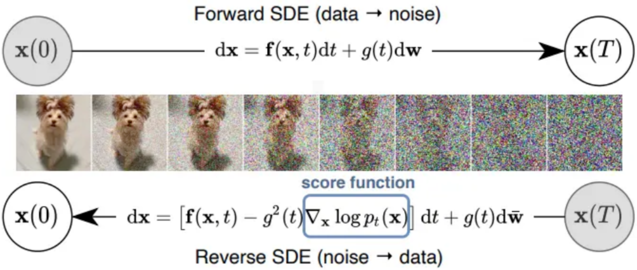
Department of Computer Science
Department of Applied Mathematics
University of California, Santa Cruz

Outline

- Introduction to Score-Based Diffusion Models
- Time Reversal of SDEs
- The Score Function
- The Fokker-Planck Equation
- Motivations for Nonlinear Diffusion Models
- Current Score Matching Techniques
- Limitations of Current Methods
- Malliavin Calculus
- Bismut Formula
- Malliavin-Bismut Framework for Linear SDEs
- Experiments and Results
- Discussion and Conclusions

Diffusion Models: Concept and Inspiration

- Inspired by nonequilibrium thermodynamics (Sohl-Dickstein et al., 2015).
- **Forward Process**: Gradually transforms structured data $x_0 \sim p_{\text{data}}(x)$ into noise.
- Perturbations mimic physical systems transitioning from order to disorder over time.
- **Reverse Process**: Reconstructs original data distribution by learning to denoise.
- Goal: Generate high-quality samples (e.g., images, audio, video) from noise.
- The forward process is modeled by a stochastic differential equation (SDE), where B_t is a standard Brownian motion defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$.



Diffusion Models: Discrete Diffusion (DDPM)

- Denoising Diffusion Probabilistic Models (Ho et al., 2020).
- **Forward Process**: Discrete steps with Gaussian noise:

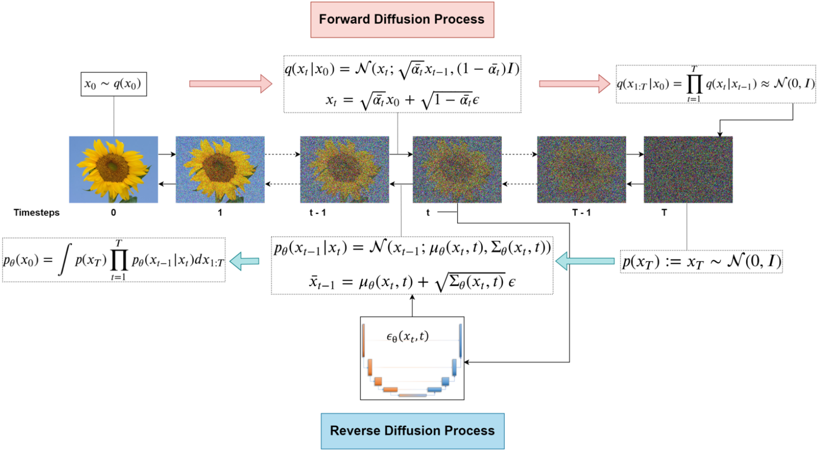
$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

- β_t: Noise schedule, $0 < \beta_t < 1$, increases over $t = 1, \ldots, T$, chosen such that $\prod_{t=1}^{T}(1-\beta_t) > 0$.
- **Reverse Process**: Learned Markov chain:

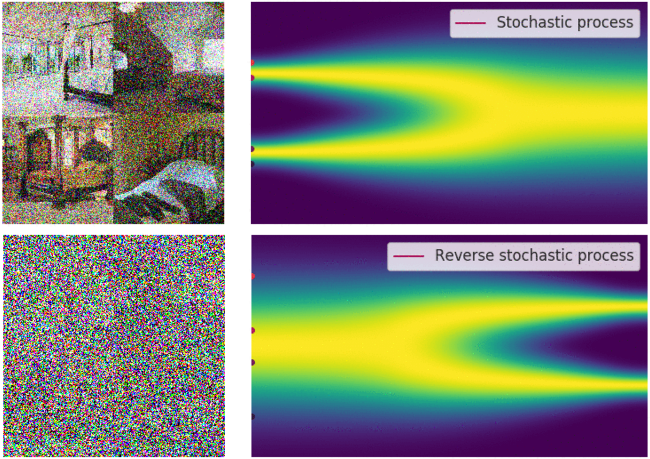
$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(t))$$

- Trains neural network to predict μ_θ, effectively denoising.
- We can also estimate the noise ϵ_θ instead of mean.

Denoising Diffusion Probabilistic Models



Forward and Reverse Processes



Diffusion Models: Applications

- **Image Synthesis**:
  - High-resolution images (Song et al., 2021).
  - Example: 256x256 images with fine details (edges, textures).
  - Photorealistic generation, style transfer, inpainting.
  - Models: DALL-E, Stable Diffusion, Imagen, Stable Diffusion 3.
- **Audio Generation**:
  - Realistic waveforms (Kong et al., 2020).
  - Example: Speech synthesis with natural harmonics.
  - Music generation, sound effects, audio restoration.
  - Models: WaveGrad, DiffWave.
- **Text-to-Image Synthesis**:
  - Generating images from textual descriptions.
  - Example: ”A cat painting in the style of Van Gogh.”
  - Models: DALL-E 2, Midjourney
- **Video Generation**:
  - Generating coherent video sequences.
  - Example: Short clips with smooth motion.
  - Models: Video Diffusion Models (VDM), Make-A-Video, Phenaki, Sora.
- Success hinges on accurate score function $\nabla_x \log p_t(x)$.

Diffusion Models: Applications

- **Medical Imaging**:
  - Synthetic medical images, quality enhancement.
  - Example: Improved MRI or CT scan resolution.
  - Anomaly detection, data augmentation.
- **Molecular Generation**:
  - Drug discovery, molecular docking.
  - Example: Diffusion-based docking (e.g., DiffDock).
  - Models: DiffDock, GeoDiff, AlphaFold (protein folding inspiration).
- **Weather Forecasting**:
  - Precipitation nowcasting, climate modelling.
- **Financial Modelling**:
  - Synthetic time-series data, risk assessment.
- **Other Domains**:
  - Robotics: Motion planning with diffusion policies.
  - NLP: Text generation (e.g., Diffusion-LM).
  - Gaming: Procedural content generation (DI-PCG).

Midjourney

In the style of Japanese anime, imagine an advertisement for ”9540” sneakers featuring a girl with white hair and light brown eyes walking on a zebra crossing. She is holding her coffee in one hand while trying to pass people who are walking quickly. The background features tall buildings. Her feet are wearing high-top canvas shoes that are primarily orange in color. A man dressed in a black suit stands next to her, watching. The illustration has a dynamic feel, reminiscent of detailed character illustrations

Diffusion Models: Mathematical Framework

- **Forward SDE**:

$$dx_t = f(t, x_t)\,dt + g(t)\,dB_t, \quad x_0 \sim p_{\text{data}}(x)$$

- $f(t, x_t)$: Drift (deterministic evolution), assumed Lipschitz continuous in x uniformly in t to ensure a unique strong solution (Øksendal, 2003, Theorem 5.2.1).
- $g(t)$: Diffusion coefficient (noise scale), continuous and bounded, B_t: Standard Brownian motion in $\mathbb{R}^d$.
- **Reverse SDE**:

$$dx_t = [f(t, x_t) - g(t)^2 \nabla_x \log p_t(x_t)]\,dt + g(t)\,d\tilde{B}_t$$

- $\nabla_x \log p_t(x_t)$: Score function, critical for reversing noise, exists if p_t is C^1 and positive.
- $\tilde{B}_t$: Reverse-time Brownian motion, defined via time reversal on $[0, T]$.

Limitations of Linear Diffusion

- Smooth distributions to Gaussian, losing complex structures (e.g., multimodality).
- Can't capture nonlinear dynamics (e.g., chaos, saturation).
- State-independent noise misses multiplicative effects (e.g., finance).
- Fixed diffusion path limits adaptability.

Motivations: Advantages of Nonlinear Diffusion Models

- Enhanced expressivity: models complex, non-Gaussian marginal distributions (*e.g.*, $f(x) = -x^3$, $g(x) = 1$.)
- Adaptation to data geometry: captures complex manifold structures (*e.g.*, $f(x) = -x|x|$, $g(x) = \sqrt{|x|}$ adapts to curvature)
- Improved generative modelling for intricate distributions, utilised in advanced models like Latent Diffusion Models

Challenges in Nonlinearity

- Nonlinear Fokker-Planck lacks closed-form solutions.
- Example: $f(x) = -x^3$ requires numerical or probabilistic methods.
- Need advanced tools: Lie groups, Malliavin calculus, etc.
- Sets stage for Malliavin-Bismut framework.

Aim: Use Mallavin Calculus to help learn non-linear Diffusion Models

- **Theorem**: $\partial_k \log p(y) = -\mathbb{E}[\delta(u_k)|F = y]$ (Bismut-type formula).
- **Why Malliavin Calculus?**:
  - Handles nonlinear diffusions and manifold geometries.
  - Computes score functions probabilistically, bypassing explicit densities.
  - Flexible as long as Malliavin derivatives are well-defined.
- **Bridging to Machine Learning**:
  - Rigorous foundation for score estimation.
  - Unified framework for general dynamics (linear, nonlinear, manifolds).
  - Practical tools via Malliavin calculus for ML applications.
- **Aim of Our Work**:
  - Build a rigorous, flexible framework for diffusion models.
  - Enable any dynamics with Malliavin calculus as the backbone.
  - Enhance machine learning models with theoretical advances.

Methods

Diffusion Models: Continuous vs. Discrete

Discrete (DDPM)

- Finite steps, predefined β_t
- Example: $x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t}\epsilon$, where $\alpha_t = \prod_{s=1}^{t}(1 - \beta_s)$, $\epsilon \sim \mathcal{N}(0, I)$

Continuous

- SDE-based, $f(t, x)$ and $g(t)$ flexible
- Advantages: Analytical tractability, customisable noise schedules
- Challenges: Requires stochastic calculus (stochastic integrals)

Time Reversal of SDEs: Concept

- **Forward**: $dx_t = f(t, x_t)\, dt + g(t)\, dB_t$.
- **Reverse**:

$$dx_t = [-f(T-t, x_t) + g(T-t)^2 \nabla_x \log p_{T-t}(x_t)]\, dt + g(T-t)\, d\tilde{B}_t$$

- Enables sampling: Noise $\rightarrow$ Data.
- Relies on accurate score estimation.
- The reverse process is Markovian, with transition densities governed by the Kolmogorov equations.

The Score Function: Definition and Intuition

- **Definition**: $s(x,t) = \nabla_x \log p_t(x)$, where $p_t(x)$ is the density of x_t in $L^1(\mathbb{R}^d)$.
- **Intuition**: Gradient of log-density, points to higher probability.
- Gaussian case: $p_t(x) = \mathcal{N}(\mu_t, \Sigma_t)$,

$$s(x,t) = -\Sigma_t^{-1}(x - \mu_t)$$

- Example: 1D, $\mu_t = 0$, $\Sigma_t = 1$, $s(x,t) = -x$.

The Score Function: Role in Reverse Process

- Guides reverse SDE:

$$dx_t = [f(t, x_t) - g(t)^2 s(x, t)]\, dt + g(t)\, d\tilde{B}_t$$

- Corrects drift to align x_t with $p_t(x)$.
- Example: VP SDE, $f = -\frac{1}{2}\beta(t)x$, $g = \sqrt{\beta(t)}$.
- Critical for generative sampling from noise.

The Score Function: Estimation Challenges

- Unknown $p_t(x)$ requires score estimation.
- Methods: Score Matching, DSM, SSM (next section).
- Challenge: Singularity in $\gamma^{-1}(t)$ as $t \to 0$.
- Example: VP SDE instability near initial time.
- The singularity arises from the Malliavin matrix $\gamma(t)$ having eigenvalues $\to 0$, requiring $\det \gamma(t) > 0$ almost surely for invertibility.

Score Matching: Overview

- Introduced by Hyvärinen (2005) for unnormalised statistical models.
- **Objective**: Minimise the Fisher divergence via the score matching objective:

$$J(\theta) = \frac{1}{2}\mathbb{E}_{x\sim\text{data}}\left[\|\nabla_x \log p_\theta(x) \ - \nabla_x \log p(x)\|^2\right]$$

- Avoids computing the partition function using integration by parts.
- They obtain a Laplacian-based estimator:

$$\mathbb{E}[\|\nabla_x \log p_\theta(x)\|^2 + 2\text{tr}(\nabla_x^2 \log p_\theta(x))]$$

- Impractical for high-dimensional data (e.g., images, audio) without approximations.

Sliced Score Matching: Objective

- Introduced by Song et al. (2019): A scalable method to estimate score functions by projecting gradients onto random vectors $\mathbf{v}$.
- **Objective**:

$$J_{\text{SSM}}(\theta) = \mathbb{E}_{x\sim p_{\text{data}},\mathbf{v}\sim\mathcal{N}(0,I)}\left[\frac{1}{2}\left(\mathbf{v}^\top\nabla_x \log p_\theta(x)\right)^2 + \mathbf{v}^\top\nabla_x^2 \log p_\theta(x)\mathbf{v}\right]$$

- **Intuition**: Approximates the score matching objective $\mathbb{E}[\|\nabla_x \log p_\theta(x)\|^2 + 2\text{tr}(\nabla_x^2 \log p_\theta(x))]$ using random projections, making it computationally efficient.
- Uses Hutchinson's trace estimator: $\mathbb{E}[\mathbf{v}^\top\nabla_x^2 \log p_\theta(x)\mathbf{v}] = \text{tr}(\nabla_x^2 \log p_\theta(x))$, reducing complexity from $O(d^2)$ to $O(d)$.
- Random vectors $\mathbf{v} \sim \mathcal{N}(0, I)$ enable Monte Carlo estimation of the expectation.
- **Pros**: Scales to high dimensions (e.g., $d = 10^6$).
- **Cons**: The estimator has Monte Carlo variance due to random projections.

Denoising Score Matching: Objective

- Introduced by Vincent (2011): Perturbs data x with a noise kernel $q_\sigma(\tilde{x}|x)$.
- **Idea**: Match the model's score on perturbed data to the perturbation kernel's score, approximating the original score matching objective
- **Objective**:

$$J_{\text{DSM}}(\theta) = \mathbb{E}_{x \sim p_{\text{data}}} \mathbb{E}_{\tilde{x} \sim q_\sigma(\cdot|x)} \left[\| \nabla_{\tilde{x}} \log p_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x) \|^2 \right]$$

- For Gaussian noise: $q_\sigma(\tilde{x}|x) = \mathcal{N}(\tilde{x}; x, \sigma^2 I)$, so:

$$\nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x) = -\frac{\tilde{x} - x}{\sigma^2}$$

- Computational advantage: Avoids Hessian computation, scaling linearly as $O(d)$ per sample.

Limitations: Score Matching Techniques

- **High Computational Cost of Score Matching:** It requires computing second-order derivatives or using trace estimators, which is expensive in high dimensions.
- **Challenges with Sliced Score Matching:** This method introduces approximation errors and does not easily handle the time-dependent score functions in diffusion models and lacks the proper conditioning.
- **Limitations of Denoising Score Matching:** It relies on knowing the transition probability, which is often difficult to obtain in nonlinear diffusion models.

- **Solution: Malliavin Calculus ...**

Malliavin Calculus: Historical Context

- Developed by Paul Malliavin in the 1970s to investigate the regularity properties of solutions to hypoelliptic partial differential equations (PDEs), a class of equations where solutions can be smooth even if the coefficients lack full ellipticity.
- Motivation: To establish conditions ensuring that the probability density function p_F of a functional $F(B_t)$ of Brownian motion B_t (e.g., $F(B_t) = \int_0^t B_s \, ds$, the time integral of Brownian motion) is smooth and differentiable, rather than merely continuous or singular.
- Stochastic Partial Differential Equations (SPDEs): Offers powerful tools to prove the existence of solutions and assess their smoothness, critical for modelling random phenomena in physics and engineering.
- Finance: Applied in option pricing, especially in advanced models incorporating stochastic volatility (e.g., Heston model) or discontinuous jumps (e.g., Lévy processes), enhancing pricing accuracy.
- Introduced the Malliavin derivative D, an operator that generalises differentiation to functionals defined on Wiener space (the space of continuous functions representing Brownian motion paths). The derivative DF of a functional F takes values in $L^2([0, T])$, the space of square-integrable functions over $[0, T]$, enabling calculus-based methods in stochastic analysis.

Malliavin Calculus: Wiener Space Definition

- $\Omega = C_0([0, \infty); \mathbb{R})$:
  - Continuous paths $\omega : [0, \infty) \to \mathbb{R}$ with $\omega(0) = 0$.
  - A Polish space (separable and completely metrisable), ideal for supporting the Wiener measure.
- Wiener measure $\mathbb{P}$:
  - A probability measure on Ω defined on the Borel σ-algebra generated by the topology of uniform convergence on compact sets.
  - The coordinate process $B_t(\omega) = \omega(t)$ is a Brownian motion.
  - Uniquely determined by the finite-dimensional distributions of B_t, consistent with Kolmogorov's extension theorem.
- Cameron-Martin space H_{CM}:
  - Subspace of Ω: absolutely continuous h with $\dot{h} \in L^2([0, \infty); \mathbb{R})$.
  - Inner product: $\langle h, g \rangle_{H_{CM}} = \int_0^\infty \dot{h}(t)\dot{g}(t)\, dt$.
- Cameron-Martin theorem:
  - For $h \in H_{CM}$, the shifted measure $\mathbb{P}_h(A) = \mathbb{P}(A - h)$ is equivalent to $\mathbb{P}$ (mutually absolutely continuous, quasi-invariant).
  - For $h \notin H_{CM}$, $\mathbb{P}_h$ and $\mathbb{P}$ are singular (mutually exclusive).
  - H_{CM} is a Hilbert space, central to Malliavin calculus.

Malliavin Calculus: Smooth Functionals

- Let $H = L^2([0,\infty);\mathbb{R})$, the space of square-integrable functions.
- **Definition**: A smooth functional is of the form $F = f(B(h_1), \dots, B(h_n))$, where:
  - $f \in C_b^\infty(\mathbb{R}^n)$ (smooth with bounded derivatives),
  - $h_i \in H$.
- $B(h_i) = \int_0^\infty h_i(t)\, dB_t$, the Wiener integral, a Gaussian random variable in $L^2(\Omega, \mathbb{P})$.
- These functionals are dense in $L^2(\Omega, \mathbb{P})$, forming a basis for Malliavin operators.

Malliavin Calculus: Malliavin Derivative Definition

- For a smooth functional $F = f(B(h_1), \ldots, B(h_n))$, with $h_i \in H = L^2([0,\infty);\mathbb{R})$:

$$D_t F = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(B(h_1), \ldots, B(h_n)) h_i(t)$$

- $DF : \Omega \to H$, where $H = L^2([0,\infty);\mathbb{R})$, measures sensitivity to perturbations in B_t.
- Example: For $F = B_T$, $D_t F = 1_{[0,T]}(t)$, which belongs to H.
- D is a Fréchet derivative in the directions of H_{CM}, well-defined if $f \in C_b^1(\mathbb{R}^n)$.

Malliavin Calculus: Skorokhod Integral

- **Adjoint**: $\mathbb{E}[F\delta(u)] = \mathbb{E}[\langle DF, u\rangle_H]$, $u \in$ Dom(δ).
- For $u_t = \sum F_j h_j(t)$,

$$\delta(u) = \sum F_j B(h_j) - \langle DF_j, h_j\rangle_H$$

- δ extends the Itô integral to $L^2(\Omega; H)$, with Dom(δ) dense in $L^2(\Omega; H)$.

Malliavin Calculus: Malliavin Matrix

- For $F = (F^1, \dots, F^m)$,

$$\gamma_F = (\langle DF^i, DF^j \rangle_H)_{i,j}$$

- **Nondegeneracy**: $P(\det \gamma_F > 0) = 1$ implies F has a density in $C^\infty(\mathbb{R}^m)$.
- 1D: $\gamma_F = \int_0^T (D_t F)^2 \, dt$.

Malliavin Calculus: Density Regularity

- Malliavin's criterion: $F \in \mathbb{D}^{\infty}$, $\mathbb{E}[|\det \gamma_F|^{-p}] < \infty$ for some $p > 1$.
- Implies p_F is smooth (Nualart, 2018).
- Key to score function derivation.

Bismut Formula: Historioal Development

- **Jean-Michel Bismut (1980s)**:
  - Introduced the Bismut formula for the gradient of the heat semigroup on Riemannian manifolds, expressed via stochastic processes.
  - Linked stochastic analysis to differential geometry, aiding large deviation principles and index theorems.
- **Elworthy-Li (1994)**:
  - Extended the formula to stochastic flows driven by stochastic differential equations (SDEs), using Malliavin calculus.
  - Applied it to derivatives of expectations for diverse diffusion processes and functionals.
- **"Bismut-type" Formulae**:
  - Refers to extensions of Bismut's work, distinct from the original heat kernel context.
  - Used in fields like financial sensitivity analysis (e.g., option pricing Greeks).
- **Probabilistic Representation**:
  - Computes the gradient of expectations, e.g., $\nabla \mathbb{E}[\varphi(X_t)]$, where X_t is a diffusion process solving an SDE.
  - **Bismut Formula (Simplified)**:

    $$\nabla \mathbb{E}[\varphi(X_t)] = \mathbb{E}\left[\varphi(X_t) \cdot W_t\right]$$

    where W_t is a stochastic weight derived from Malliavin calculus or the first variation process and φ is a functional.
  - **Key Idea**: Expresses sensitivity to initial conditions probabilistically, avoiding explicit density calculations.

Bismut Formula: Covering Vector Fields

- **Definition**: For a random vector $F = (F_1, \ldots, F_m)$, covering vector fields $u_k \in L^2(\Omega; H)$ (for $k = 1, \ldots, m$) satisfy:

  $$\langle DF_i, u_k \rangle_H = \delta_{i,k} \quad \text{(1 if } i = k\text{, 0 otherwise)}$$

  where DF_i is the Malliavin derivative of F_i, and H is the Cameron-Martin space of perturbation directions.
- **Intuition**: Think of u_k as "arrows" in the space of Brownian paths. Each u_k perturbs the noise so that only the k-th component of F changes, "covering" all directions like a coordinate system. This lets us measure how F varies in each direction independently.
- **Purpose**: They enable the Bismut formula to compute gradients:

  $$\partial_k \mathbb{E}[\varphi(F)] = \mathbb{E}[\varphi(F)\delta(u_k)]$$

  linking deterministic derivatives to stochastic integrals.
- **Example**: For $F = X_T$, the solution to an SDE at time T, $u_k = \sum_{j=1}^{m} (\gamma_{X_T}^{-1})_{k,j} DX_{T,j}$, where $\gamma_{X_T} = (\langle DX_{T,i}, DX_{T,j} \rangle_H)_{i,j}$ is the Malliavin covariance matrix.
- **Properties**: The $u_k \in L^2(\Omega; H)$ ensure that the map $DF : H \to \mathbb{R}^m$ is surjective, which holds when the Malliavin covariance matrix γ_F is invertible, allowing perturbations in all directions of F's range.

Bismut-Type Formula

- **Theorem**: $\partial_k \log p(y) = -\mathbb{E}[\delta(u_k)|F = y]$.
- Probabilistic expression for score.
- **How can we arrive at a computable formula for this expression?**
  - Pick a suitable covering vector field u_k.
  - Reduce the Skorokhod integral $\delta(u_k)$ to an Itô integral for computational tractability.
  - Rewrite Malliavin derivatives DF in terms of variation processes, derived from the SDE.

SDE and First Variation Process

- Consider the linear SDE:

$$dX_t = b(t)X_t\,dt + \sigma(t)\,dB_t$$

  where $X_t \in \mathbb{R}^m$, B_t is a standard Brownian motion in $\mathbb{R}^d$, $\sigma(t) \in \mathbb{R}^{m\times d}$, $b(t) \in \mathbb{R}^{m\times m}$, and $X_0 \sim p_{\text{data}}$. The drift term $b(t)X_t\,dt$ is linear in X_t.
- The first variation process $Y_t = \frac{\partial X_t}{\partial x_0}$ satisfies:

$$dY_t = \partial_x[b(t)X_t]Y_t\,dt + \partial_x\sigma(t)Y_t\,dB_t$$

  Since $\sigma(t)$ is independent of X_t, $\partial_x\sigma(t) = 0$, and $\partial_x[b(t)X_t] = b(t)$, reducing it to the ODE:

$$dY_t = b(t)Y_t\,dt, \quad Y_0 = I_m$$

- This becomes:

$$\frac{dY_t}{dt} = b(t)Y_t$$

  with solution:

$$Y_t = \exp\left(\int_0^t b(s)\,ds\right)$$

  assuming $b(t)$ commutes if matrix-valued.
- Example: If $b(t) = -I_m$, then:

$$Y_t = e^{-t}I_m$$

- Properties: Y_t is continuous in t, invertible, and bounded in $L^\infty([0,T])$ if $b(t)$ is integrable.

Malliavin-Bismut: Malliavin Matrix Derivation

- The Malliavin matrix is defined as:

$$\gamma_{X_T} = \int_0^T D_r X_T (D_r X_T)^\top \, dr$$

  where $D_r X_T$ is the Malliavin derivative of X_T, showing its sensitivity to Brownian motion perturbations at time r.

- For a linear SDE $dX_t = b(t) X_t \, dt + \sigma(t) \, dB_t$:

$$D_r X_T = Y_T Y_r^{-1} \sigma(r)$$

  with $Y_t = \exp\left(\int_0^t b(s) \, ds \right)$, the first variation process.

- Substituting and simplifying:

$$\gamma_{X_T} = Y_T \left(\int_0^T Y_r^{-1} \sigma(r) \sigma(r)^\top (Y_r^{-1})^\top \, dr \right) Y_T^\top$$

  resembling the covariance structure of X_T's Malliavin derivatives.

Malliavin-Bismut: Covering Vector Field Construction

- $u_k(t) = \sum_{j=1}^{m} \gamma_{X_T}^{-1}(k,j) D_t X_T^j$.
- Verification: $\langle DX_T^i, u_k \rangle_H = \delta_{i,k}$.
- Ensures directional sensitivity.

Malliavin-Bismut: Skorokhod to Itô Reduction

- Since $u_k(t)$ is adapted (due to deterministic coefficients),

$$\delta(u_k) = \int_0^T u_k(t)\, dB_t$$

- Simplify: $\delta(u_k) = [\gamma_{X_T}^{-1}(X_T - Y_T X_0)]_k$.
- The reduction holds in $L^2(\Omega)$ as u_k is $\mathcal{F}_t$-adapted, aligning with the Itô integral's definition.

Malliavin-Bismut: Score Function Formula

- Bismut formula: $\partial_k \log p(y) = -\mathbb{E}[\delta(u_k)|X_T = y]$.
- Final form:
$$\nabla_y \log p(y) = -\gamma_{X_T}^{-1}(y - Y_T\mathbb{E}[X_0|X_T = y])$$
- $\nabla \log p \in L^2(\mathbb{R}^m)$ if $p(y)$ is sufficiently smooth (e.g., $p \in H^1(\mathbb{R}^m)$) and $\mathbb{E}[X_0|X_T = y]$ is well-defined..

Malliavin-Bismut: Regression Insight

- Score reduces to estimating $\mathbb{E}[X_0|X_T = y]$.
- Transforms score matching into regression problem.
- Simplifies computation via neural networks.
- The regression is well-posed in $L^2(\Omega)$, assuming X_0 and X_T are jointly integrable.

Covariance in Malliavin and Fokker-Planck

- **Linear SDE**: $dX_t = b(t)X_t\,dt + \sigma(t)\,dW_t$, with initial condition $X_0 = x_0$.
- **Fokker-Planck Approach**:
  - Solves the PDE for the density $p_t(x) = \mathcal{N}(\mu_t, \Sigma_t)$, where:

$$\mu_t = Y_t x_0, \quad \Sigma_t = Y_t \left(\int_0^t Y_s^{-1} \sigma(s)\sigma(s)^\top (Y_s^{-1})^\top \, ds \right) Y_t^\top$$

  - $Y_t = \exp\left(\int_0^t b(s)\,ds\right)$ is the fundamental matrix.
- **Malliavin Approach**:
  - Malliavin matrix: $\gamma_{X_T} = \int_0^T D_r X_T (D_r X_T)^\top \, dr$, where $D_r X_T = Y_T Y_r^{-1} \sigma(r)$ is the Malliavin derivative.
  - Compute:

$$\gamma_{X_T} = Y_T \left(\int_0^T Y_r^{-1} \sigma(r)\sigma(r)^\top (Y_r^{-1})^\top \, dr \right) Y_T^\top$$

- **Result**: $\gamma_{X_T} = \Sigma_T$, showing that the Malliavin matrix (stochastic sensitivity) equals the Fokker-Planck covariance (statistical variance).

Score Function in Malliavin and Fokker-Planck

- **Fokker-Planck Score**: For $p_t(x) = \mathcal{N}(Y_t x_0, \Sigma_t)$ with deterministic $X_0 = x_0$:

$$\nabla_x \log p_t(x) = -\Sigma_t^{-1}(x - Y_t x_0)$$

- **Malliavin-Bismut Score**: General form for X_T at time T:

$$\nabla_y \log p(y) = -\gamma_{X_T}^{-1}(y - Y_T \mathbb{E}[X_0 | X_T = y])$$

- **Equivalence (Deterministic X_0)**:
  - If $X_0 = x_0$ is fixed, then $\mathbb{E}[X_0 | X_T = y] = x_0$.
  - Thus: $\nabla_y \log p(y) = -\gamma_{X_T}^{-1}(y - Y_T x_0)$.
  - Since $\gamma_{X_T} = \Sigma_T$, this equals $-\Sigma_T^{-1}(y - Y_T x_0)$, matching the Fokker-Planck score.

Malliavin-Bismut: Algorithm Overview

- **Algorithm**: Malliavin Diffusion Framework.
- Steps:
  1. Simulate forward SDE: X_t.
  2. Compute $\gamma_{X_t}^{-1}$ and Y_t.
  3. Train NN for $\mathbb{E}[X_0|X_t, t]$.
  4. Sample reverse SDE with score.

Malliavin-Bismut: Practical Considerations

- NN predicts $\mathbb{E}[X_0|X_t, t]$ (e.g., U-Net).
- Cost: Matrix inversion of γ_{X_t} per time step.
- Scales with dimension m and time steps N.

VP SDE, sub-VP SDE Definition

- VP SDE: $dX_t = -\frac{1}{2}\beta(t)X_t\,dt + \sqrt{\beta(t)}\,dB_t$.
- $\beta(t) = \beta_{\min} + (\beta_{\max} - \beta_{\min})\frac{t}{T}$, $\beta \in C([0,T])$.
- sub-VP SDE: $dX_t = -\frac{1}{2}\beta(t)X_t\,dt + \sqrt{\beta(t)(1 - e^{-2\int_0^t \beta(s)\,ds})}\,dB_t$.
- Variance-preserving: Maintains signal variance.

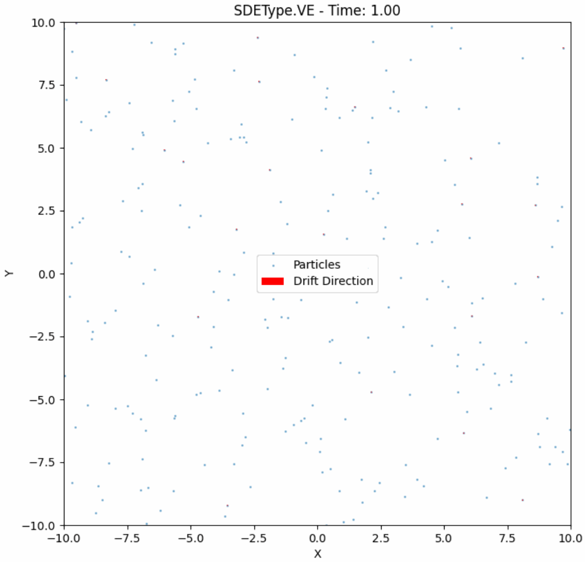
Observation: Singularity in VP SDE

- Result: $\gamma^{-1}(t) = O\left(\frac{1}{t}\right)$ as $t \to 0$.
- From: $\gamma(t) \approx \beta_{\min} t I$.
- Causes numerical issues in score near $t = 0$.
- $\gamma(t)$'s eigenvalues scale as t, with $\|\gamma^{-1}(t)\| \to \infty$ in $L^{\infty}([0, \epsilon])$, violating uniform ellipticity.

Observation: Singularity in Sub-VP SDE

- Result: $\gamma^{-1}(t) = O\left(\frac{1}{t^2}\right)$ as $t \to 0$.
- From: $\gamma(t) \approx \beta_{\min}^4 t^2 I$.
- Stronger singularity than VP.

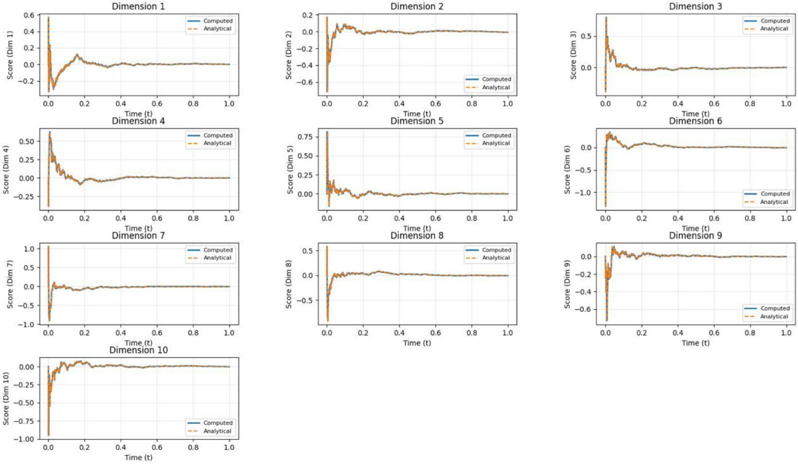
An Experiment: Checkerboard



Implications and Mitigations

- **Implications**: Instability near $t = 0$ affects sampling.
- **Mitigations**:
  - Regularise $\sigma(t)$: Linear growth near 0.
  - Adjust drift: Add damping term.
  - Tikhonov regularisation: Perturb $\gamma(t)$.
- Regularisation ensures $\gamma(t)$ is invertible in $L^2([0, T])$, akin to Tikhonov's method for ill-posed operators.

Malliavin Score vs Analytical Score



Discussion: Summary of Contributions

- Developed Malliavin-Bismut framework for linear diffusion generative models.
- A promising framework for nonlinear diffusion generative models (next work)
- Reduced score matching to regression problem.
- Analysed singularities: VP $O(1/t)$, Sub-VP $O(1/t^2)$.

Conclusions: Future Work

- Extend to nonlinear SDEs (Part II).
- Implement in generative tasks (e.g., image synthesis).
- Explore regularisation for stability.

Conclusions: Broader Impact

- Enhances robustness of diffusion models.
- Enables nonlinear modelling with stochastic tools.
- Encourages Malliavin calculus in ML research.

Thank You