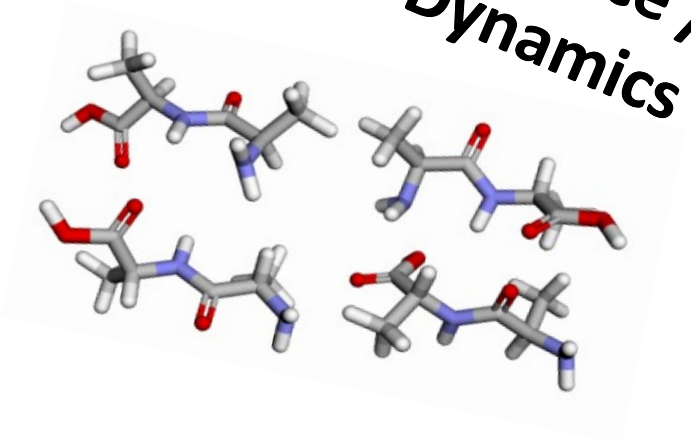
Diffusion Models, SDEs and Path Based Inference

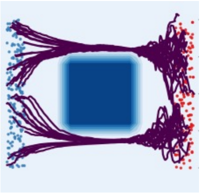
Francisco Vargas

Generative Modelling

Bayesian Inference / Molecular Dynamics



Filtering / Data Assimilation



Diffusion Models and SDEs

Lecture 1:

A very fast paced introduction to the foundations / notation.

Quick Probability Recap

Probability Space

$$\mathbb{P}(\Omega) = 1, \quad P(A) \geq 0$$

$$\mathbb{P}(\cup_{i \in \mathcal{I}} A_i) = \sum_{i \in \mathcal{I}} \mathbb{P}(A_i)$$

$$A_i \cap A_j = \emptyset, i \neq j \;, \quad \exists f : \mathcal{I} \longleftrightarrow \mathbb{N}$$

- Sample Space

e.g $\Omega = \{0, 1\}$ or $\Omega = \mathbb{R}$

- Probability Measure

$$(\Omega, \Sigma, \mathbb{P})$$

- Event Space e.g $2^{\{0,1\}}$

/ Sigma Algebra: is a algebra/system of sets that are "closed" under countable # of operations $\cup, \cap, \setminus\Omega$ and $\Omega, \emptyset \in \Sigma \subseteq 2^{\Omega}$

Quick Probability Recap

Probability Space

$$\mathbb{P}(\Omega) = 1, \quad P(A) \geq 0$$

$$\mathbb{P}(\cup_{i\in\mathcal{I}} A_i) = \sum_{i\in\mathcal{I}} \mathbb{P}(A_i)$$

$$A_i \cap A_j = \emptyset, i \neq j \;, \;\; \exists f : \mathcal{I} \longleftrightarrow \mathbb{N}$$

- Sample Space

e.g $\Omega = \{0, 1\}$ or $\Omega = \mathbb{R}$

- Probability Measure

$$(\Omega, \mathcal{B}(\Omega), \mathbb{P})$$

- Event Space e.g $2^{\{0,1\}}$

The Borel-sigma algebra is the smallest sigma algebra containing the event space (i.e. intersect all possible sigma algebra containing Omega).

Quick Probability Recap

Filtered Probability Space

- Think of a filtration as the sample space of a time series, that is a series of sample spaces:

$$\mathcal{F} = \{\mathcal{F}_t\}_{t \in [0,T]}$$

$$s \leq t \implies \mathcal{F}_s \subseteq \mathcal{F}_t$$

$$(\Omega, \mathcal{B}(\Omega), \mathcal{F}, \mathbb{P})$$

Quick Probability Recap

Stochastic Process

- Collection of Random Variables (Measurable Maps) !

$$\{X_t\}_{t\in[0,T]} \qquad X_t(\omega) : [0,T]\times\Omega \to \mathbb{R}^d$$

$$\Big(C([0,T];\mathbb{R}^d),\mathcal{B}\big(C([0,T];\mathbb{R}^d)\big),\mathcal{F},\mathbb{P}\Big)$$

Quick Probability Recap

Brownian Motion

- Brownian motion is a Gaussian Process, and one of the simplest Stochastic Processes:
  - Pinned Origin: $W_0 = 0$
  - Independent increments $s, t > 0,\ W_{t+s} - W_t \perp\!\!\!\perp W_t$
  - $W_{t+s} - W_t \sim \mathcal{N}(0, s)$
  - W_t is continuous in t (almost surely)

$$W \sim \mathcal{GP}(0, \min(s, t))$$

Quick Probability Recap

Lebesgue Integral

$$\int_A \mathrm{d}\lambda = \lambda(A)$$

$$\int_\Omega \mathbb{I}_A(x)\mathrm{d}\lambda = \lambda(A)$$

$$\int_\Omega \sum_{i=1}^{n} a_i \mathbb{I}_{A_i}(x)\mathrm{d}\lambda = \sum_{i=1}^{n} a_i \lambda(A_i)$$

$$\int_A f\mathrm{d}\lambda = \sup\left\{\int s\mathrm{d}\lambda : 0 \leq s \leq f, s = \sum_{i=1}^{n} \alpha_i \mathbb{I}_{A_i}(x)\right\}$$

Quick Probability Recap

Lebesgue-Stjelties Integral

$$\int_A f \mathrm{d}\lambda = \sup\left\{\int s \mathrm{d}\lambda : 0 \leq s \leq f, s = \sum_{i=1}^{n} \alpha_i \mathbb{I}_{A_i}(x)\right\}$$

$$\int_A f(x)\mathrm{d}\lambda(x) = \int_A f(x)\mathrm{d}x = \int_A f(x)\lambda(\mathrm{d}x)$$

We can replace lambda with a probability distribution/measure yielding the familiar expectation:

$$\int_A f(x)\mathrm{d}P(x) = \mathbb{E}_P[f(X)]$$

Quick Probability Recap

Radon Nikodym Theorem – Change of Measure

$$\mu << \lambda := \lambda(A) = 0 \implies \mu(A) = 0$$

$$\mu(A) = \int_A \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x)\mathrm{d}\lambda(x)$$

$$\int_A f(x)\mathrm{d}\mu(x) = \int_A f(x)\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x)\mathrm{d}\lambda(x)$$

Quick Probability Recap

Radon Nikodym Theorem – Probaility Density Function

$$\mathbb{P} << \lambda$$

$$\mathbb{P}(A) = \int_A \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\lambda}(x)\mathrm{d}\lambda(x)$$

Now For sake of simplicity assume Reimann Integrability

$$\mathbb{P}(A) = \int_A \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\lambda}(x)\mathrm{d}\lambda(x) = \int_A \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\lambda}(x)\mathrm{d}x$$

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\lambda}(x) = \text{Probability Density Function !}$$

Quick Probability Recap

Radon Nikodym Theorem – Importance Sampling

$$\mathbb{P} << \mathbb{Q}$$

$$\int_\Omega f(x) \mathrm{d}\mathbb{P}(x) = \int_\Omega f(x) \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x) \mathrm{d}\mathbb{Q}(x)$$

$$\mathbb{E}_\mathbb{P}[f(X)] = \mathbb{E}_\mathbb{Q}\left[f(X) \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(X)\right]$$

$$\mathbb{E}_\mathbb{P}[f(X)] = \mathbb{E}_\mathbb{Q}\left[f(X) \frac{p(X)}{q(X)}\right]$$

Quick Probability Recap

Modes of equality/convergence of r.v.s.

$$
\begin{tikzcd}
\lim\limits_{n\to\infty} \mathbb{P}\left(|X - X_n| > \epsilon\right) = 0 \arrow[dr, Rightarrow] & \mathbb{P}\left(\lim\limits_{n\to\infty} X_n = X\right) = 1 \arrow[l, Rightarrow] \arrow[d, Rightarrow] \\
\lim\limits_{n\to\infty} \mathbb{E}\left[|X - X_n|^p\right] = 0 \arrow[u, Rightarrow] \arrow[r, Rightarrow] & \mathrm{Law}\, X = \lim\limits_{n\to\infty} \mathrm{Law}\, X_n
\end{tikzcd}
$$

SDEs

Heuristic 1 – Discrete Time Markov Chain (Euler Maruyama Discretisation)

$$X_0 \sim \pi,$$

$$\epsilon_n \sim \mathcal{N}(0, \gamma I)$$

$$X_{n+1} = X_n + f(X_n, n)\delta t + \sqrt{\delta t}\epsilon_n,$$

SDEs

Heuristic 2 – Langevin Dynamics and White Noise

- Consider the ODE + Noise

$$X_0 \sim \pi,$$

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = f(X_t, t) + \gamma w(t),$$

$$w(\cdot) \sim \mathcal{GP}(0, \mathbb{I}_{s=t})$$

SDEs

Stochastic Integrals - Types

$$Y_t = \int_0^t X_s \mathrm{d}s$$

- Can think of this as a Reimann integral with convergence asserted in the $\mathscr{L}^p(\mathbb{P})$ sense

$$Z_t = \int_0^t Y_s \mathrm{d}X_s$$

- Now integrating against/wrt to random variable. Not so simple to define. Reimann conditions fail

SDEs

Stochastic Integrals – Counter Example

$$\mathbb{E}\left[\sum_{k=1}^{n} W_{t_k}(W_{t_{k+1}} - W_{t_k})\right] = 0$$

$$\mathbb{E}\left[\sum_{k=1}^{n} W_{t_{k+1}}(W_{t_{k+1}} - W_{t_k})\right] = t$$

- Where you evaluate the integrand (within the grid) changes the result, thus violating the conditions required to be Reimann integrable (remember upper and lower Darboux sums must much)

SDEs

Stochastic Integrals - Definition

- First partition the grid [0,t] $t_{k+1} - t_k = \frac{t}{N}$
- Now we make the following assumption

$$\lim_{n\to\infty} \mathbb{E}\left[\int_0^t |Y_t - Y_t^{(n)}|^2 \mathrm{d}s\right] = 0 \quad \text{s.t.} \quad Y^{(n)}(t) = \sum_{k=1}^{n} Y_{t_k} \mathbb{I}_{t\in[t_k,t_{k+1})}(t)$$

- Then the Ito Integral is defined as:

$$\int_0^t Y_s \mathrm{d}\overset{\approx \mathcal{N}(0,\mathrm{d}s)}{W_s} \overset{\mathscr{L}^2(\mathbb{P})}{=} \lim_{n\to\infty} \sum_{k=1}^{n} Y_{t_k}(W_{t_{k+1}} - W_{t_k})$$

Martingales

Conditional Expectation - Martingale

$$\mathbb{E}\left[\mathbf{X}_t \middle| \mathcal{F}_s\right] = \mathbf{X}_s$$

$$\Downarrow$$

$$\mathbb{E}\left[\mathbf{X}_t \middle| \mathbf{X}_s\right] = \mathbb{E}\left[\mathbf{X}_t \middle| \sigma\left(\mathbf{X}_s\right)\right] = \mathbf{X}_s$$

Conditional Expectation, MSE

Quick Aside (Useful Later)

The optimal predictor of X as a function of Y (Hilbert projection)

$$\underset{f-\text{is measurable}}{\arg\min} \quad \mathbb{E}\left(X - f(Y)\right)^2$$

Is given by the conditional expectation:

$$f^*(Y) = \mathbb{E}[X|Y]$$

Martingales

Martingales – Intuitive Intro

The optimal predictor of the future as a function of the past in a martingale:

$$\underset{f-\text{is measurable}}{\arg\min} \quad \mathbb{E}\left(X_{t+\delta} - f(X_t)\right)^2$$

Is given by past itself:

$$f^*(X_t) = \mathbb{E}[X_{t+\delta}|X_t] = X_t$$

Martingales

Stochastic Integrals - Martingales

$$\mathbb{E}\left[\int_0^t X_\tau \mathrm{d}W_\tau\right] = \mathbb{E}\left[\mathbb{E}\left[\int_0^t X_\tau \mathrm{d}W_\tau \middle| \mathcal{F}_0\right]\right]$$

$$= \mathbb{E}\left[\int_0^0 X_\tau \mathrm{d}W_\tau\right] = 0$$

SDEs

Formal Definition - Stochastic Piccard Lindeloff Theorem

- Assumptions (Lipchitz + Linear Growth):

$$|\mu(x,t)-\mu(y,s)|+|\sigma(x,t)-\sigma(y,s)|\leq L(|x-y|+|t-s|)$$

$$|\mu(x,t)|+|\sigma(x,t)|\leq C(1+|x|)$$

- Then we have existence and uniqueness of (in $\mathscr{L}^p(\mathbb{P})$):

$$X_0\sim\pi$$

$$X_t=X_0+\int_0^t\mu(X_s,s)\mathrm{d}s+\int_0^t\sigma(X_s,s)\mathrm{d}W_s$$

$$\mathrm{d}X_t=\mu(X_t,t)\mathrm{d}t+\sigma(X_t,t)\mathrm{d}W_t$$

Diffusion Models and SDEs

Lecture 2:

SDE Properties, Linear SDEs, Time Reversal and the h-transform

SDE Properties

Quadratic Variation of Brownian Motion

$$\lim_{n\to\infty} \mathbb{E}\left(t - \sum_{i=1}^{n}(W_{t_{i+1}} - W_{t_i})^2\right)^2 = 0$$

SDE Properties

Quadratic Variation of Brownian Motion

$$\lim_{n\to\infty} \mathbb{E}\left(t - \sum_{i=1}^{n}(W_{t_{i+1}} - W_{t_i})^2\right)^2 = 0$$

	$\mathrm{d}W_t$	$\mathrm{d}t$
$\mathrm{d}W_t$	$\mathrm{d}t$	0
$\mathrm{d}t$	0	0

SDE Properties

Ito's Lemma

Given the SDE:

$$\mathrm{d}X_t = \mu(X_t, t)\mathrm{d}t + \sigma(X_t, t)\mathrm{d}W_t$$

Consider a function $f(t, x)$ doubly differentiable in space and admitting single derivatives in time. Then the process $Y_t = f(t, X_t)$ satisfies:

$$\mathrm{d}Y_t = \left(\partial_t f + \nabla f^\top \mu(X_t, t) + \frac{1}{2}\mathrm{tr}(\sigma(X_t, t)^\top \nabla\nabla f \sigma(X_t, t))\right)\mathrm{d}t + \nabla f^\top \sigma(X_t, t)\mathrm{d}W_t$$

SDE Properties

Ito's Lemma - Exercise : Geometric Brownian Motion

Let us solve the SDE:

$$\mathrm{d}X_t = \mu X_t \mathrm{d}t + \sigma X_t \mathrm{d}W_t$$

now consider the transformation $Y_t = \ln X_t$ what are ?

$$\partial_t f = ??, \quad \partial_x f = ?? \quad \partial_x^2 f = ??$$

SDE Properties

Ito's Lemma - Exercise : Geometric Brownian Motion

Let us solve the SDE:

$$\mathrm{d}X_t = \mu X_t \mathrm{d}t + \sigma X_t \mathrm{d}W_t$$

now consider the transformation $Y_t = \ln X_t$ what are ?

$$\partial_t f = 0, \quad \partial_x f = 1/x \quad \partial_x^2 f = -1/x^2$$

$$\mathrm{d}Y_t = \left(\frac{\mu}{X_t} \cdot X_t - \frac{\sigma^2}{2X_t^2} \cdot X_t^2 \right) \mathrm{d}t - \frac{\sigma}{X_t} \cdot X_t \mathrm{d}W_t$$

SDE Properties

Ito's Lemma - Exercise : Geometric Brownian Motion

Let us solve the SDE:

$$\mathrm{d}X_t = \mu X_t \mathrm{d}t + \sigma X_t \mathrm{d}W_t$$

now consider the transformation $Y_t = \ln X_t$ what are ?

$$\partial_t f = 0, \quad \partial_x f = 1/x \quad \partial_x^2 f = -1/x^2$$

$$\mathrm{d}Y_t = \left(\mu - \frac{\sigma^2}{2}\right)\mathrm{d}t - \sigma \mathrm{d}W_t$$

SDE Properties

Ito's Lemma - Exercise : Geometric Brownian Motion

Now let us solve the SDE:

$$\mathrm{d}Y_t = \left(\mu - \frac{\sigma^2}{2}\right)\mathrm{d}t - \sigma \mathrm{d}W_t$$

$$Y_t = Y_0 + \left(\mu - \frac{\sigma^2}{2}\right)\int_0^t \mathrm{d}s - \sigma \int_0^t \mathrm{d}W_s = Y_0 + \left(\mu - \frac{\sigma^2}{2}\right) t + \sigma W_t$$

Remember $Y_t = \ln X_t$ thus:

$$X_t = e^{Y_t} = X_0 e^{\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t}$$

Fokker Plank Equation

How does the marginal density evolve (SDEs ⇔ Parabolic PDEs)

What is the probability density of the SDE solution at a given time ?

$$\mathrm{Law} X_t = p_t(x) = ???$$

There's a special PDE (think heat equation) whose solution yield the marginal density:

$$\partial_t p_t(x) = -\sum_{i=1}^{d} \partial_{x_i}[\mu_i(t, x_i) p_t(x)] + \sum_{i,j=1}^{d} \partial_{x_i, x_j}[\sigma\sigma_{ij}^{\top}(t, x) p_t(x)]$$

Fokker Plank Equation

How does the marginal density evolve (SDEs ⇔ Parabolic PDEs)

What is the probability density of the SDE solution at a given time ?

$$\mathrm{Law} X_t = p_t(x) = ???$$

There's a special PDE (think heat equation) whose solution yield the marginal density:

$$\partial_t p_t(x) = \mathcal{P}(p_t)$$

Infinitesimal Generator

Uniquely Characterises PDE and Adjoint to FPK Operator

Consider the following operator for a given SDE

$$\mathcal{A}_t[f(x)] = \lim_{t \to 0} \frac{\mathbb{E}[f(X_t)] - x}{t}$$

Can be shown to reduce to:

$$\mathcal{A}_t[f] = \partial_t f + \mu \cdot \nabla f + \frac{1}{2} \sum_{ij} [\sigma \sigma^\top]_{ij}(x,t) \partial_{x_i, x_j} f$$

$$= \partial_t f + \mathcal{P}^\dagger(f)$$

Linear SDEs

OU - Process

Mean reverting process. Reverts you back to mu.

$$X_0 \sim \pi$$

$$\mathrm{d}X_t = \alpha(\mu - X_t)\mathrm{d}t + \sqrt{2\alpha}\mathrm{d}W_t$$

Linear SDEs

OU - Process

For simplicity focus on the 0-mean case.

$$X_0 \sim \pi$$

$$\mathrm{d}X_t = -\alpha X_t \mathrm{d}t + \sqrt{2\alpha}\mathrm{d}W_t$$

Linear SDEs

OU - Process

Can be solved analytically via Integrating factor + Ito's Lemma (notice how X_t looks like the DDPM kernel):

$$X_t = X_0 e^{-\alpha t} + (1 - e^{-2\alpha t})^{1/2} W_1$$

$$X_t = X_0 e^{-\alpha t} + W_{1-e^{-2\alpha t}}$$

Linear SDEs

OU - Process

Intuitively you can see how the limit behaves:

$$\lim_{t \to \infty} X_t \overset{??}{=} W_1 \sim \mathcal{N}(0, I)$$

This is a completely informal/heuristic treatment. Calling it a heuristic is kind, but you can see where it is going.

Linear SDEs

OU - Process

More formal arguments can be made:

$$||\mathrm{Law} X_t - \mathcal{N}(0, I)||_{\mathrm{TV}} \leq C e^{-\alpha t}$$

Can be a bit tricky to show from scratch, typically involves working with the Fokker Plank Equation + Using an Eigen decomposition of its semi group. Alternatively, Martingale methods have also been used.

Convergence in KL, W_p can also be attained see Bakry, Gentil, Ledoux **Analysis and Geometry of Markov Diffusion Operators**.

Non Linear SDEs - Simply Discretise

Euler Maruyama (EM) Discretisation

To solve SDEs of the form

$$\mathrm{d}X_t = \mu(X_t, t)\mathrm{d}t + \sigma(X_t, t)\mathrm{d}W_t$$

We simply discretize them via EM

$$X_0 \sim \pi,$$
$$\epsilon_{t_k} \sim \mathcal{N}(0, \gamma I)$$
$$X_{t_{k+1}} = X_{t_k} + \mu(X_{t_k}, t_k)\delta t + \sqrt{\delta t}\sigma(X_{t_k}, t_k)\epsilon_{t_k},$$

Can prove convergence in $\mathscr{L}^p(\mathbb{P})$. Can we design better integrators ?

Time Reversal - Chain Rule

A discrete time “heuristic” sketch

Via the chain rule we can decompose the joint in either direction,

$$p_{t|t+\delta}(x|y)p_{t+\delta}(y)=p_{t+\delta|t}(y|x)p_t(x)$$

Now consider an EM approx transition density, for the forward kernel:

$$p_{t+\delta|t}(y|x) = \mathcal{N}(y|x + f^+(x)\delta, \delta\sigma^2)$$

$$p_{t|t+\delta}(x|y) = ?$$

Time Reversal - Chain Rule

A discrete time "heuristic" sketch

Via the chain rule we can decompose the joint in either direction,

$$p_{t|t+\delta}(x|y)p_{t+\delta}(y)=p_{t+\delta|t}(y|x)p_t(x)$$

Now consider an EM approx transition density, for the forward kernel:

$$p_{t+\delta|t}(y|x) = \mathcal{N}(y|x + f^+(x)\delta, \delta\sigma^2)$$

$$p_{t|t+\delta}(x|y) = p_{t+\delta|t}(y|x)\frac{p_t(x)}{p_{t+\delta}(y)}$$

Time Reversal - Chain Rule

A discrete time "heuristic" sketch

Via Taylors Theorem we can expand time t marginal around y:

$$p_{t|t+\delta}(x|y) = p_{t+\delta|t}(y|x)\frac{p_t(y)e^{(x-y)^\top \nabla_y \ln p_t(y)+\mathcal{O}(\delta^2)}}{p_{t+\delta}(y)}$$

Assuming $|\ln p_t(x) - \ln p_s(x)| = \mathcal{O}(|t-s|^2)$

$$p_{t|t+\delta}(x|y) = p_{t+\delta|t}(y|x)e^{(x-y)^\top \nabla_y \ln p_t(y)+\mathcal{O}(\delta^2)}$$

Time Reversal - Chain Rule

A discrete time “heuristic” sketch

Regrouping and completing the square:

$$p_{t|t+\delta}(x|y) = \frac{e^{-\frac{||x-(y-f^{+}(y)\delta+\sigma^2\nabla_y \ln p_t(y)\delta||^2}{\sigma^2\delta}+\mathcal{O}(\delta^2)}}{\sqrt{2\pi}\delta^{d/2}\sigma^d}$$

Which corresponds to the Euler Maruyama discretization of the following SDE (seem familiar ?):

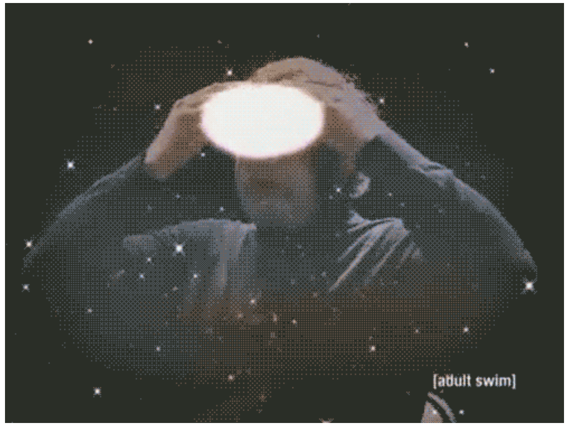
$$\mathrm{d}X_t = \left(-f^{+}(X_t, T-t) + \sigma^2\nabla_{X_t} \ln p_{T-t}(X_t)\right)\mathrm{d}t + \sigma \mathrm{d}W_t$$

Time Reversal - Chain Rule

A discrete time “heuristic” sketch

Inspecting the relationship between the drifts yields Nelsons duality formula:

$$f^{-}(x,t) + f^{+}(x,T-t) = \sigma^2 \nabla_x \ln p_{T-t}(x)$$



Time Reversal - Chain Rule

A discrete time "heuristic" sketch

Inspecting the relationship between the drifts yields Nelsons duality formula:

$$f^-(x,t) + f^+(x,T-t) = \sigma^2 \nabla_x \ln p_{T-t}(x)$$

Looks slightly different to Song et al. 2021, why ?

Time Reversal - Chain Rule

Nelsons Relation – Semantics Clarification

Looks slightly different to Song et al. 2020, why ?

Due to 2 equivalent ways of representing time reversals:

$$\mathrm{d}Y_t = f^+(Y_t, t)\mathrm{d}t + \sigma \mathrm{d}W_t$$

Forward SDE (e.g. De Bortoli 2021)

- Travels forward in time

$$\mathrm{d}\mathbf{X_t} = \mathbf{f^-}(\mathbf{X_t}, \mathbf{t})\mathrm{d}\mathbf{t} + \sigma \mathrm{d}\mathbf{W_t}$$

$$\mathbf{f^-}(\mathbf{x}, \mathbf{t}) + \mathbf{f^+}(\mathbf{x}, \mathbf{T} - \mathbf{t}) = \sigma^2 \nabla_{\mathbf{x}} \ln \mathbf{p_{T-t}}(\mathbf{x})$$

- Flips / No longer the same joint

$$\mathrm{Law}(\mathbf{x_t})_{\mathbf{t=0}}^{\mathbf{T}} = \mathrm{Law}(\mathbf{y_{T-t}})_{\mathbf{t=0}}^{\mathbf{T}}$$

Backwards SDE (e.g. Song 2021)

- Travels Backwards in time

$$\mathrm{d}\mathbf{X_t^-} = \mathbf{f^-}(\mathbf{X_t^-}, \mathbf{t})\mathrm{d}\mathbf{t} + \sigma \mathrm{d}\mathbf{W_t^-}$$

$$\mathbf{f^-}(\mathbf{x}, \mathbf{t}) - \mathbf{f^+}(\mathbf{x}, \mathbf{t}) = \sigma^2 \nabla_{\mathbf{x}} \ln \mathbf{p_t}(\mathbf{x})$$

- Encodes the same joint

$$\mathrm{Law}(\mathbf{x_t})_{\mathbf{t=0}}^{\mathbf{T}} = \mathrm{Law}(\mathbf{y_t})_{\mathbf{t=0}}^{\mathbf{T}}$$

Time Reversal – Generative Modelling

Time reversing VP-SDE / OU Process [Song 2021, De Bortoli 2021]

Consider the time homogenous VP-SDE (OU Process):

$$X_0 \sim p_{\text{data}}$$
$$\mathrm{d}X_t = -\beta X_t \mathrm{d}t + \sqrt{2\beta}\mathrm{d}W_t$$

Then its time reversal $(Y_t)_{t=0}^T \overset{d}{=} (X_{T-t})_{t=0}^T$ satisfies the score SDE [Song 2021]:

$$Y_0 \sim p_T \approx \mathcal{N}(0, I)$$
$$\mathrm{d}Y_t = (\alpha Y_t + 2\alpha \nabla_{Y_t} \ln p_{T-t}(Y_t))\,\mathrm{d}t + \sqrt{2\alpha}\mathrm{d}B_t$$

Where $Y_T \sim p_{\text{data}}$, thus we could instead sample approximately $Y_0 \sim \mathcal{N}(0, I)$ and have $\mathrm{Law} Y_T \approx p_{\text{data}}$ following the mixing rate of the OU [De Bortoli 2021]

Doobs – Transform (Quick Version)

Introduction

Given the SDE with transition density $p_{t|s}(x|y)$

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + \sigma \mathrm{d}W_t$$

We would like to find the process arising from conditioning the above SDE to hit a deterministic end point.

$$p_{t|s,T}(x_t|x_s, x_T = z) = \frac{p_{t|s,T}(x_T = z|x_t)p_{t|s}(x_t|x_s)}{p(x_T = z|x_s)}$$

Is this process itself an SDE ? Turns out it is.

Doobs – Transform (Quick Version)

Formal(ish) Statement

Given the SDE

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + \sigma \mathrm{d}W_t$$

Then its conditioning to hit a point at time T is given by

$$\mathrm{d}Z_t = (f(Z_t, t) + \sigma^2 \nabla \ln p_{T|t}(z|Z_t))\mathrm{d}t + \sigma \mathrm{d}W_t$$

Where $Z_T \sim \delta_z$ and $p^h_{t|s}(z_t|z_s) = p_{t|s,T}(z_t|z_s, z_T = z)$

relevant result for conditional generation (e.g. inpainting)

Doobs – Transform – Example Pinned Brownian

Generative Modelling / Sampling with Pinned Brownian Motion

Consider a Brownian Motion, starting from an arbitrary distribution

$$X_0 \sim p_{\text{data}}$$

$$\mathrm{d}X_t = \sigma \mathrm{d}W_t$$

Then its conditioned SDE to hit 0 at time t is given by

$$\mathrm{d}Z_t = -\frac{X_t}{T-t}\mathrm{d}t + \sigma \mathrm{d}W_t$$

Where $Z_T \sim \delta_0$ note the time reversal of Z_t maps from 0 to the data distribution, learning its score provides us with an alternative generative model to VP-SDE / OU see [Vargas et al. 2022, Ye et al 2022.].

Diffusion Models and SDEs

Lecture 3:

Girsanov Theorem, KL Divergence, Half Bridges, FK- Formula

Reminder

Conditional Expectation Property

The optimal predictor of X as a function of Y (Hilbert projection)

$$\underset{f-\text{is measurable}}{\arg\min} \quad \mathbb{E}\left(X - f(Y)\right)^2$$

Is given by the conditional expectation:

$$f^*(Y) = \mathbb{E}[X|Y]$$

Tractable Score matching loss

Last Lecture – Song Score Matching Objective

$$s^* = \underset{s-\text{is measurable}}{\arg\min}\ \mathbb{E}\left[\int_0^T \left|\left|\nabla \ln p_{t|0}(X_t|X_0) - s(t, X_t)\right|\right|^2 \mathrm{d}t\right]$$

Tractable Score matching loss

Last Lecture – Song Score Matching Objective

$$s^* = \underset{s-\text{is measurable}}{\arg\min} \ \mathbb{E}\left[\int_0^T \left|\left|\nabla \ln p_{t|0}(X_t|X_0) - s(t, X_t)\right|\right|^2 \mathrm{d}t\right]$$

$$s^*(t,x) = \mathbb{E}_{X_0|X_t}[\nabla \ln p_{t|0}(X_t|X_0)|X_t = x]$$

Tractable Score matching loss

Last Lecture – Song Score Matching Objective

$$s^* = \underset{s-\text{is measurable}}{\arg\min} \; \mathbb{E}\left[\int_0^T \left|\left|\nabla \ln p_{t|0}(X_t|X_0) - s(t, X_t)\right|\right|^2 \mathrm{d}t\right]$$

$$s^*(t,x) = \mathbb{E}_{X_0|X_t}[\nabla \ln p_{t|0}(X_t|X_0)|X_t = x]$$

$$s^*(t,x) = \int p_{0|t}(x_0|x)\nabla \ln p_{t|0}(x|x_0)\mathrm{d}x_0$$

Tractable Score matching loss

Last Lecture – Song Score Matching Objective

$$s^* = \underset{s-\text{is measurable}}{\arg\min} \; \mathbb{E}\left[\int_0^T \left|\left|\nabla \ln p_{t|0}(X_t|X_0) - s(t, X_t)\right|\right|^2 \mathrm{d}t\right]$$

$$s^*(t,x) = \mathbb{E}_{X_0|X_t}[\nabla \ln p_{t|0}(X_t|X_0)|X_t = x]$$

$$s^*(t,x) = \int p_{0|t}(x_0|x)\nabla \ln p_{t|0}(x|x_0)\mathrm{d}x_0$$

$$s^*(t,x) = \int \frac{p_{t|0}(x|x_0)p_0(x_0)}{p_t(x)}\nabla \ln p_{t|0}(x|x_0)\mathrm{d}x_0$$

Tractable Score matching loss

Last Lecture – Song Score Matching Objective

$$s^*(t,x) = \int \frac{p_{t|0}(x|x_0)p_0(x_0)}{p_t(x)} \nabla \ln p_{t|0}(x|x_0) \mathrm{d}x_0$$

Tractable Score matching loss

Last Lecture – Song Score Matching Objective

$$s^*(t,x) = \int \frac{p_{t|0}(x|x_0)p_0(x_0)}{p_t(x)} \nabla \ln p_{t|0}(x|x_0)\mathrm{d}x_0$$

$$s^*(t,x) = \frac{1}{p_t(x)} \int p_0(x_0)\nabla p_{t|0}(x|x_0)\mathrm{d}x_0$$

Tractable Score matching loss

Last Lecture – Song Score Matching Objective

$$s^*(t,x) = \int \frac{p_{t|0}(x|x_0)p_0(x_0)}{p_t(x)} \nabla \ln p_{t|0}(x|x_0)\mathrm{d}x_0$$

$$s^*(t,x) = \frac{1}{p_t(x)} \int p_0(x_0) \nabla p_{t|0}(x|x_0)\mathrm{d}x_0$$

$$s^*(t,x) = \frac{1}{p_t(x)} \nabla \int p_0(x_0) p_{t|0}(x|x_0)\mathrm{d}x_0$$

Tractable Score matching loss

Last Lecture – Song Score Matching Objective

$$s^*(t,x) = \int \frac{p_{t|0}(x|x_0)p_0(x_0)}{p_t(x)} \nabla \ln p_{t|0}(x|x_0) \mathrm{d}x_0$$

$$s^*(t,x) = \frac{1}{p_t(x)} \int p_0(x_0) \nabla p_{t|0}(x|x_0) \mathrm{d}x_0$$

$$s^*(t,x) = \frac{1}{p_t(x)} \nabla \int p_0(x_0) p_{t|0}(x|x_0) \mathrm{d}x_0$$

$$s^*(t,x) = \frac{1}{p_t(x)} \nabla p_t(x) = \nabla_x \ln p_t(x)$$

Girsanov Theorem I

General Statement

Given Novikovs condition and a Brownian motion in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ follows that:

$$B_t = W_t + \int_0^t \Theta(t)\mathrm{d}s$$

Is a Brownian motion in the probability space $(\Omega, \mathcal{F}, \mathbb{Q})$. Where

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \exp\left(-\int_0^T \Theta(t)^\top \mathrm{d}W_t - \frac{1}{2}\int_0^T ||\Theta(t)||^2 \mathrm{d}t\right)$$

Girsanovs Theorem - Corollary

General Statement

Given the SDE

$$\mathrm{d}W_t^\sigma = \sigma(W_t^\sigma, t)\mathrm{d}W_t$$

With probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then it follows that:

$$B_t = W_t - \int_0^t \mu(W_s^\sigma, s)\sigma^{-1}(W_s^\sigma, s)\mathrm{d}s$$

Is a Brownian motion in the probability space $(\Omega, \mathcal{F}, \mathbb{Q})$. Where

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \exp\left(\int_0^T \sigma^{-1}(W_t^\sigma, t)\mu(W_t^\sigma, t)^\top \mathrm{d}W_t - \frac{1}{2}\int_0^T \sigma^{-2}(W_t^\sigma, t)||\mu(W_t^\sigma, t)||^2 \mathrm{d}t\right)$$

Girsanovs Theorem - Corollary

General Statement

Given the SDE

$$\mathrm{d}W_t^\sigma = \sigma(W_t^\sigma, t)\mathrm{d}W_t$$

With probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then it follows that:

$$dB_t = dW_t - \mu(W_t^\sigma, t)\sigma^{-1}(W_t^\sigma, t)\mathrm{d}t$$

Is a Brownian motion in the probability space $(\Omega, \mathcal{F}, \mathbb{Q})$. Where

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \exp\left(\int_0^T \sigma^{-1}(W_t^\sigma, t)\mu(W_t^\sigma, t)^\top \mathrm{d}W_t - \frac{1}{2}\int_0^T \sigma^{-2}(W_t^\sigma, t)||\mu(W_t^\sigma, t)||^2 \mathrm{d}t\right)$$

Girsanovs Theorem - Corollary

General Statement

Furthermore, we have that

$$\begin{aligned} \mathrm{d}W_t^\sigma &= \sigma(W_t^\sigma, t)(\mathrm{d}B_t + \sigma^{-1}\mu(W_t^\sigma, t)\mathrm{d}t) \\ &= \mu(W_t^\sigma, t)\mathrm{d}t + \sigma(W_t^\sigma, t)\mathrm{d}B_t \end{aligned}$$

Thus, in the space $(\Omega, \mathcal{F}, \mathbb{Q})$ the process W_t^σ weakly solves the SDE

$$\mathrm{d}X_t = \mu(X_t, t)\mathrm{d}t + \sigma(X_t, t)\mathrm{d}B_t$$

With:

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \exp\left(\int_0^T \sigma^{-1}(W_t^\sigma, t)\mu(W_t^\sigma, t)^\top \mathrm{d}W_t - \frac{1}{2}\int_0^T \sigma^{-2}(W_t^\sigma, t)||\mu(W_t^\sigma, t)||^2 \mathrm{d}t\right)$$

Girsanovs Theorem – RND Corollary

Importance Sampling Again

Then we have that:

$$\mathbb{E}_{\mathbb{Q}}[f(X)] = \mathbb{E}_{\mathbb{P}}\left[\exp\left(\int_0^T \sigma_t^{-1}\mu_t^\top \mathrm{d}W_t - \frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t||^2 \mathrm{d}t\right) f(W^\sigma)\right]$$

Which is effectively the statement of the RN theorem, so it follows that

$$\frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\mathbb{P}_{W^\sigma}}(W^\sigma) = \exp\left(\int_0^T \sigma_t^{-1}\mu_t^\top \mathrm{d}W_t - \frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t||^2 \mathrm{d}t\right)$$

Girsanovs Theorem – RND Corollary

Caveat !!

This result gives us the RND when evaluated on a sample from W^σ if instead we wanted to evaluate the RND on a sample from X we would have to apply Girsanovs theorem with a sign flip starting from the SDE solving X and transforming it to the law of W^σresulting in:

$$\frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\mathbb{P}_{W^\sigma}}(X) = \exp\left(\int_0^T \sigma_t^{-1}\mu_t^\top \mathrm{d}W_t + \frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t||^2 \mathrm{d}t\right)$$

So, remember depending on what we take expectations with respect to the signs in the RND will change.

Optional bonus exercise with 1d Gaussians to be added to homework.

RNDs – General Result

Likelihood Ratio Between Diffusions

Given 2 SDEs (with the same initial condition X_0=Y_0=x):

$$\mathrm{d}X_t = \mu(X_t, t)\mathrm{d}t + \sigma(X_t, t)\mathrm{d}B_t\,, \qquad \mathrm{d}Y_t = \rho(Y_t, t)\mathrm{d}t + \sigma(Y_t, t)\mathrm{d}B_t$$

satisfying all the conditions we have discussed. It follows that:

$$\frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\mathbb{P}_Y}(X) = \exp\left(\int_0^T \sigma_t^{-1}(\mu_t - \rho_t)^\top \mathrm{d}W_t + \frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t - \rho_t||^2 \mathrm{d}t\right)$$

$$\frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\mathbb{P}_Y}(Y) = \exp\left(\int_0^T \sigma_t^{-1}(\mu_t - \rho_t)^\top \mathrm{d}W_t - \frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t - \rho_t||^2 \mathrm{d}t\right)$$

KL- Divergence

Likelihood Ratio Between Diffusions

Remember (changing notation a bit P^f refers to the SDE with drift f)

$$D_{KL}(\mathbb{P}^{\mu}||\mathbb{P}^{\rho}) = \mathbb{E}_{X\sim\mathbb{P}^{\mu}}\left[\ln\frac{\mathrm{d}\mathbb{P}^{\mu}}{\mathrm{d}\mathbb{P}^{\rho}}(X)\right]$$

Now applying Girsanov's theorem (e.g. the corollaries we derived):

$$D_{KL}(\mathbb{P}^{\mu}||\mathbb{P}^{\rho}) = \mathbb{E}_{X\sim\mathbb{P}^{\mu}}\left[\int_0^T \sigma_t^{-1}(\mu_t-\rho_t)^\top \mathrm{d}W_t + \frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t-\rho_t||^2\mathrm{d}t\right]$$

$$= \mathbb{E}_{X\sim\mathbb{P}^{\mu}}\left[\frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t-\rho_t||^2\mathrm{d}t\right]$$

KL- Divergence – Score Matching

Likelihood Ratio Between Diffusions – OU time reversal

Remember the Ito integral is a Martingale (1st Lecture) and thus has 0 expectation resulting in:

$$D_{KL}(\mathbb{P}^{\mu}||\mathbb{P}^{\rho}) = \mathbb{E}_{X\sim\mathbb{P}^{\mu}}\left[\frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t - \rho_t||^2 \mathrm{d}t\right]$$

Now consider the case where X is the time reversal of an OU process and we can parametrize P^\rho as a score network SDE, which results in:

$$D_{KL}(\mathbb{P}^{\beta x+\nabla \ln p_{T-t}(x)}||\mathbb{P}^{\beta x+s^{\rho}_{T-t}(x)}) = \mathbb{E}_{X\sim\mathbb{P}^{\mu}}\left[\frac{1}{2}\int_0^T \sigma^2_{T-t}||\nabla \ln p_{T-t} - s^{\rho}_{T-t}||^2 \mathrm{d}t\right]$$

KL- Divergence – Score Matching

Likelihood Ratio Between Diffusions – OU time reversal

$$D_{KL}(\mathbb{P}^{\mu}||\mathbb{P}^{\rho}) = \mathbb{E}_{X\sim\mathbb{P}^{\mu}}\left[\frac{1}{2}\int_0^T \sigma_{T-t}^2||\nabla \ln p_{T-t} - s_{T-t}^{\rho}||^2 \mathrm{d}t\right]$$

Now remember we can sample X_t via sampling Z_{T-t} where Z_t is the original (non reversed) noising OU process thus we have:

$$D_{KL}(\mathbb{P}^{\mu}||\mathbb{P}^{\rho}) = \mathbb{E}_{Z\sim\mathbb{P}^{\mu}}\left[\frac{1}{2}\int_0^T \sigma_t^2||\nabla \ln p_t - s_t^{\rho}||^2 \mathrm{d}t\right]$$

Same mean squared error objective as in Song et al. 2021 !

Chain Rule – Disintegration Theorem

The chain rule is a little bit more complicated for path measures

$$\mathbb{P}(A_0 \times A_{(0,T]}) = \int_{A_0} \mathbb{P}_{\cdot|0}(A_{(0,T]}|x)\mathrm{d}\mathbb{P}_0(x)$$

Which under certain regularity assumptions (which SDEs satisfies) implies

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\cdot) = \frac{\mathrm{d}\mathbb{P}_{\cdot|0}(|x)}{\mathrm{d}\mathbb{Q}_{\cdot|0}(|x)}\frac{\mathrm{d}\mathbb{P}_0}{\mathrm{d}\mathbb{Q}_0}(x)$$

Sometimes written as

$$\mathrm{d}\mathbb{P} = \mathrm{d}\mathbb{P}_{\cdot|0}(|x)\mathrm{d}\mathbb{P}_0(x)$$

Chain Rule – Disintegration Theorem

The chain rule is a little bit more complicated for path measures

$$\mathbb{P}(A_0 \times A_{(0,T]}) = \int_{A_0} \mathbb{P}_{\cdot|0}(A_{(0,T]}|x)\mathrm{d}\mathbb{P}_0(x)$$

Which under certain regularity assumptions (which SDEs satisfies) implies

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\cdot) = \frac{\mathrm{d}\mathbb{P}_{\cdot|0}(|x)}{\mathrm{d}\mathbb{Q}_{\cdot|0}(|x)}\frac{\mathrm{d}\mathbb{P}_0}{\mathrm{d}\mathbb{Q}_0}(x)$$

Sometimes written as

$$\mathrm{d}\mathbb{P} = \mathrm{d}\mathbb{P}_{\cdot|0}(|x)\mathrm{d}\mathbb{P}_0(x)$$

Half Bridges – Constrained KL minimisation

Constrained Optimisation

$$\mathbb{P}^* = \underset{\mathbb{P}\,:\,\text{s.t. }\mathbb{P}_T = \pi}{\arg\min}\; D_{KL}(\mathbb{P}||\mathbb{P}^\rho)$$

Then

$$\mathrm{d}\mathbb{P}^* = \mathrm{d}\mathbb{P}^\rho \frac{\mathrm{d}\pi}{\mathrm{d}\mathbb{P}^\rho_T}$$

Half Bridges – Constrained KL minimisation

Unconstrained Formulation – Stochastic Control

$$\mathbb{P}^* = \arg\min_{\mathbb{P}} D_{KL}(\mathbb{P}^\mu || \mathbb{P}^*)$$
$$= \arg\min_{\mathbb{P}} D_{KL}(\mathbb{P}^\mu || \mathbb{P}^\rho) - \mathbb{E}\left[\ln \frac{\mathrm{d}\pi}{\mathrm{d}\mathbb{P}^\rho_T}\right]$$

$$\mathbb{P}^\mu_0 = \mathbb{P}^*_0$$

Now applying Girsanovs Theorem (Stochastic Control Objective)

$$\arg\min_{\mu} \mathbb{E}_{X\sim\mathbb{P}^\mu}\left[\frac{1}{2}\int_0^T \sigma_t^{-2} ||\mu_t - \rho_t||^2 \mathrm{d}t\right] - \mathbb{E}\left[\ln \frac{\mathrm{d}\pi}{\mathrm{d}\mathbb{P}^\rho_T}\right]$$

Generative Modelling and Sampling/Inference

2 Sides of the same Coin

Vargas, F., Grathwohl, W. and Doucet, A., 2023. Denoising diffusion samplers. ICLR 2023

Generative Modelling	**Sampling / Inference**
- **Access to samples.**	- **Access to a density up to constant**
$$p_{\text{data}} = \frac{1}{N}\sum_n \delta_{x_i}$$	$$p_{\text{data}}(x) = \frac{e^{-U(x)}}{\mathcal{Z}}$$
- **Typically optimises Forward KL**	- **Usually optimises Reverse KL**
$$\text{argmin}_{\mathbb{P}}\text{KL}(\mathbb{Q}\|\|\mathbb{P})$$	$$\text{argmin}_{\mathbb{Q}}\text{KL}(\mathbb{Q}\|\|\mathbb{P})$$
- **e.g Score Matching, DDPM, MLE**	- **e.g DDS, PIS, DIS**

All fall under the half bridge framework !

Diffusion Models and SDEs

Lecture 4:

Schrodinger Bridges, IPF/Sinkhorn, Entropic Optimal Transport

Schrodinger Bridges – Intuition

Schrodinger 1931/32

In 1931/32, Erwin Schrodinger proposed the following Gedankenexperiment [52, 53]:

Consider the evolution of a cloud of N independent Brownian particles in R^3 . This cloud of particles has been observed having at the initial time t = 0 an empirical distribution equal to π_0 .

Schrodinger Bridges – Intuition

Schrodinger 1931/32

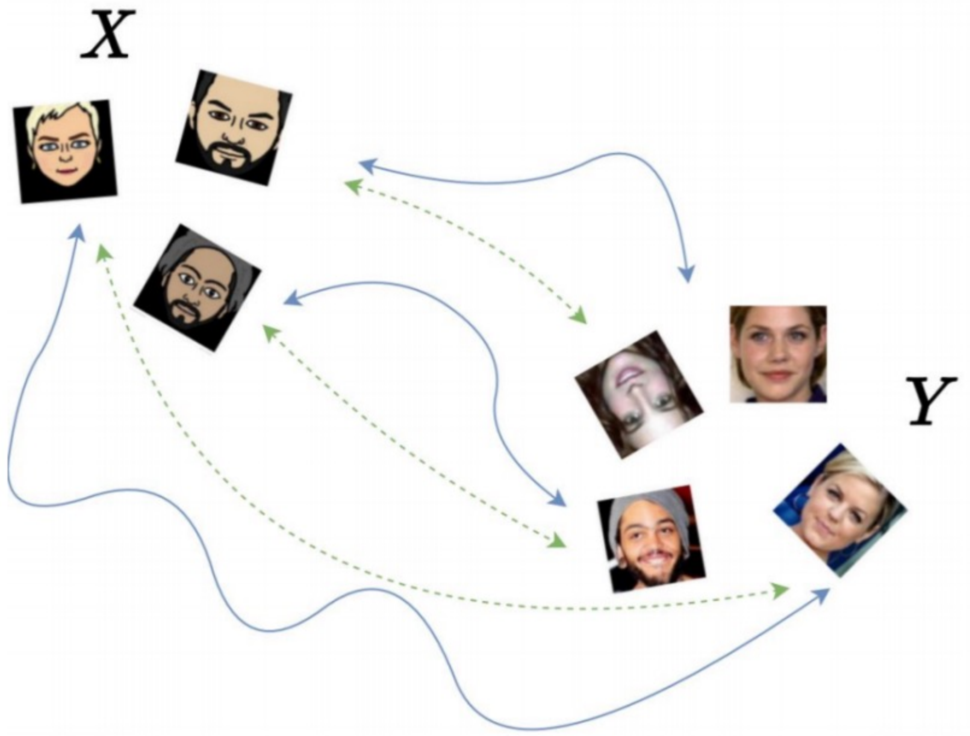
At time t = T, an empirical distribution π_1 is observed which considerably differs from what it should be according to the law of large numbers (N is large, typically of the order of Avogadro's number), namely

$$\pi_1(y) \neq \int_{\mathbb{R}^3} \mathcal{N}(y; x, T)\pi_0(x)dx$$

It seems that the particles have been transported in an unlikely way. But of the many unlikely ways in which this could have happened, which one is the most likely?

Schrodinger Bridges – Motivation

Schrodinger 1931/32



$$\arg\min_{\approx \in (\cdot,\cdot)} D_{KL}(\approx \| \equiv)$$

Schrodinger Bridges – Constrained KL minimisation

Constrained Optimisation

$$\mathbb{P}^* = \underset{\mathbb{P}\,:\,\text{s.t. }\mathbb{P}_0=\pi_0,\mathbb{P}_T=\pi_1}{\arg\min} D_{KL}(\mathbb{P}||\mathbb{P}^\rho)$$

Much harder problem than half bridges. Does not admit such a simple unconstrained formulation. Lets disintegrate:

$$\underset{\mathbb{P}\,:\,\text{s.t. }\mathbb{P}_0=\pi_0,\mathbb{P}_T=\pi_1}{\arg\min} D_{KL}(\mathbb{P}_{0,T}||\mathbb{P}^\rho_{0,T}) + \mathbb{E}_{\mathbb{P}_{0,T}} D_{KL}(\mathbb{P}_{|0,T}||\mathbb{P}^\rho_{|0,T})$$

Schrodinger Bridges – Entropic Optimal Transport

From Dynamic SBP to Static Entropic OT

$$\underset{\mathbb{P}\,:\,\text{s.t. }\mathbb{P}_0=\pi_0,\mathbb{P}_T=\pi_1}{\arg\min} \quad D_{KL}(\mathbb{P}_{0,T}||\mathbb{P}^{\rho}_{0,T}) + \mathbb{E}_{\mathbb{P}_{0,T}} \cancel{D_{KL}(\mathbb{P}_{|0,T}||\mathbb{P}^{\rho}_{|0,T})}$$

$$\underset{\mathbb{P}\,:\,\text{s.t. }\mathbb{P}_0=\pi_0,\mathbb{P}_T=\pi_1}{\arg\min} \quad D_{KL}(\mathbb{P}_{0,T}||\mathbb{P}^{\rho}_{0,T})$$

$$\underset{p(x,y):\,\text{s.t.}\,p(x)=\pi_0,p(y)=\pi_1}{\arg\min} \quad \mathbb{E}[\sigma^2 \ln p^{\rho}_{T|0}(y|x)] + \sigma^2 H(p)$$

Already looking like Entropic OT simply let p(x|y) = exp(-c(x,y)/sigma^2) and we arrive at your usual entropic OT objective.

Schrodinger Bridges – Entropic Optimal Transport

From Dynamic SBP to Static Entropic OT

$$\min_{p(x,y):\ \mathrm{s.t.} p(x)=\pi_0, p(y)=\pi_1} \mathbb{E}[\sigma^2 \ln p^{\rho}_{T|0}(y|x)] + \sigma^2 H(p)$$

Let \rho=0 then we have :

$$\min_{p(x,y):\ \mathrm{s.t.} p(x)=\pi_0, p(y)=\pi_1} \mathbb{E}[||y-x||^2] + \sigma^2 H(p) = \mathcal{W}^2_{2,\sigma^2}(\pi_0, \pi_1)$$

Aka the entropy regularized Wasserstein distance between the boundary distributions.

Schrodinger Bridges – IPF/Sinkhorn Algorithm

Solution - Alternating Subproblems (Coordinate Ascent - Sinkhorn Algorithm)

$$\mathbb{P}_0^* = \mathbb{P}^\rho$$

$$\mathbb{Q}_i^* = \underset{\mathbb{Q}\,:\,\text{s.t. }\mathbb{Q}_T=\pi_1}{\arg\min} \quad D_{KL}(\mathbb{Q}||\mathbb{P}_i^*)$$

$$\mathbb{P}_{i+1}^* = \underset{\mathbb{P}\,:\,\text{s.t. }\mathbb{P}_0=\pi_0}{\arg\min} \quad D_{KL}(\mathbb{P}||\mathbb{Q}_i^*)$$

The above IPF (Iterative Proportional Fitting) iterates also known as sinkhorn have been proved to converge to the Schrodinger bridge solution. This approach dates back to Kullback.

Schrodinger Bridges – IPF/Sinkhorn Algorithm

Solution - Alternating Subproblems (Coordinate Ascent - Sinkhorn Algorithm)

These should look familiar

$$\mathbb{Q}_i^* = \underset{\mathbb{Q}\,:\,\text{s.t. } \mathbb{Q}_T = \pi_1}{\arg\min} \; D_{KL}(\mathbb{Q}||\mathbb{P}_i^*)$$

$$\mathbb{P}_{i+1}^* = \underset{\mathbb{P}\,:\,\text{s.t. } \mathbb{P}_0 = \pi_0}{\arg\min} \; D_{KL}(\mathbb{P}||\mathbb{Q}_i^*)$$

They are half bridges, and we know how to solve via score matching or stochastic control (i.e., via minimizing forward or reverse KL iteratively).

Vargas, F., Thodoroff, P., Lamacraft, A. and Lawrence, N., 2021. Solving schrödinger bridges via maximum likelihood. *Entropy*, *23*(9), p.1134.

Schrodinger Bridges – Schrodinger System

Solution – Functional System of Potentials

Another way to formulate the solution (and construct iterations) is based on the Schrodinger system:

$$\hat{\phi}_0(x)\phi_0(x) = \pi_0(x), \quad \hat{\phi}_1(y)\phi_1(y) = \pi_1(y)$$

$$\phi_0(x) = \int p_{T|0}(x|y)\phi_1(y)dy, \quad \hat{\phi}_1(y) = \int p_{T|0}(y|x)\hat{\phi}_0(x)dx$$

Result can be arrived at via Disintegration Theorem -> Lagrange Multipliers -> Calc of Variations. (The potentials are the Lagrange multipliers).

Schrodinger Bridges – Schrodinger System

Solution – Functional System of Potentials

Then given the potentials we have that

$$X_0 \sim \pi_0$$

$$dX_t = \left(\rho + \sigma^2 \left(\nabla_{X_t} \ln \int \phi_1(z) p^{\rho}_{T|t}(z|X_t) \mathrm{d}x\right)\right) \mathrm{d}t + \sigma \mathrm{d}W_t$$

$$Y_0 \sim \pi_1$$

$$dY_t = \left(\rho - \sigma^2 \left(\nabla_{Y_t} \ln \int \hat{\phi}_0(z) p^{\rho}_{t|0}(Y_t|z) \mathrm{d}z\right)\right) \mathrm{d}t + \sigma \mathrm{d}W_t^-$$

Solve The Schrodinger Bridge when the path measures represent SDE solutions.

Schrodinger Bridges – Schrodinger System

Solution – PDE Formulation

Furthermore, the potentials

$$\phi_t(x) = \int \phi_1(z) p^{\rho}_{T|t}(z|x)\mathrm{d}x \qquad \hat{\phi}_t(y) = \int \hat{\phi}_0(z) p^{\rho}_{t|0}(y|z)\mathrm{d}z$$

Solve the Following PDEs (remember space-time regularity from Doobs transform):

$$-\partial_t \phi_t = \nabla \phi_t \cdot \rho + \sigma^2 \Delta \phi_t, \qquad \hat{\phi}_0(x)\phi_0(x) = \pi_0(x)$$

$$\partial_t \hat{\phi}_t = -\nabla \cdot (\hat{\phi}_t \rho) + \sigma^2 \Delta \hat{\phi}_t, \qquad \hat{\phi}_1(y)\phi_1(y) = \pi_1(y)$$

These are just the FPK and the backward Kolmogorov equations. With funky boundary conditions.

Schrodinger Bridges – HJB/Hopf-Cole/Flemming

Solution – PDE Formulation

Via reversing Flemings/Hopf-Cole transform that is:

$$\psi_t(x) = \exp(\phi_t(x)), \quad \hat{\psi}_t(y) = \exp(\hat{\phi}_t(y))$$

Then through some standard calculus we arrive at the following HJB-PDEs:

$$-\partial_t \psi_t = ||\sigma \nabla \psi_t||^2 + \nabla \psi_t \cdot \rho + \sigma^2 \Delta \psi_t, \qquad \hat{\psi}_0(x) + \psi_0(x) = \ln \pi_0(x)$$

$$\partial_t \hat{\psi}_t = ||\sigma \nabla \hat{\psi}_t||^2 - \nabla \hat{\psi}_t \cdot (\rho - \ln p_t) + \sigma^2 \Delta \hat{\psi}_t, \quad \hat{\psi}_1(y) + \psi_1(y) = \ln \pi_1(y)$$

And thus, connecting to stochastic control / verification results etc.

Recap and Take Aways

OU and Pinned Brownian Motion

We studied two SDEs which transform complex distributions into simple distributions:

$$X_0 \sim \pi$$

$$\mathrm{d}X_t = \alpha(\mu - X_t)\mathrm{d}t + \sqrt{2\alpha}\mathrm{d}W_t$$

$$X_0 \sim \pi$$

$$\mathrm{d}X_t = \frac{\mu - X_t}{T - t}\mathrm{d}t + \sqrt{\sigma}\mathrm{d}W_t$$

The OU process which rapidly mixes into a Gaussian, and the Pinned Brownian motion which instantaneously maps any distribution into a point mass.

$$Z_0 \sim \mathrm{law} X_T \approx \mathcal{N}(\mu, 1)$$

$$\mathrm{d}Z_t = (\alpha(Z_t - \mu) + 2\alpha \nabla \ln p_{T-t}(Z_t))\mathrm{d}t + \sqrt{2\alpha}\mathrm{d}B_t$$

$$Z_0 = \mu$$

$$\mathrm{d}Z_t = \left(\frac{Z_t - \mu}{t} + \sigma^2 \nabla \ln p_{T-t}(Z_t)\right)\mathrm{d}t + \sigma \mathrm{d}B_t$$

Their respective time reversals provide us with tractable generative models!

Recap and Take Aways

OU and Pinned Brownian Motion

In both settings we can learn the score and thus the time reversal via solving simple MSE/Regression objectives where we sample from the original noising processes to generate the "data" for the objectives.

$$Z_0 \sim \mathrm{law} X_T \approx \mathcal{N}(\mu, 1)$$

$$\mathrm{d}Z_t = (\alpha(Z_t - \mu) + 2\alpha \nabla \ln p_{T-t}(Z_t))\mathrm{d}t + \sqrt{2\alpha}\mathrm{d}B_t$$

$$Z_0 = \mu$$

$$\mathrm{d}Z_t = \left(\frac{Z_t - \mu}{t} + \sigma^2 \nabla \ln p_{T-t}(Z_t)\right)\mathrm{d}t + \sigma \mathrm{d}B_t$$

In both cases learning the score / time reversal has an equivalent variational formulation in terms of half/full bridges:

$$\underset{\mathbb{P}\,:\,\text{s.t. } \mathbb{P}_T = \pi}{\arg\min} \; D_{KL}(\mathbb{P}||\mathbb{P}^{\alpha(\mu - x)})$$

$$\underset{\mathbb{P}\,:\,\text{s.t. } \mathbb{P}_0 = \delta_0, \mathbb{P}_T = \pi}{\arg\min} \; D_{KL}(\mathbb{P}||\mathbb{P}^0)$$

Which can be applied to gen modelling, sampling, path simulation, etc.

What did we miss ??

- Feynman Kac Formula (Useful for re-expressing marginals ,deriving ELBOs)
- Trading scores with divergences via integration by parts (Allows for a Hutchinsons type estimator)
- Thorough introduction to backwards Ito integrals and divergence based conversion formula.
- Stochastic Control, HJB Equation, Equivalence between time reversal and control.
- Discrete time convergence results (De Bortoli et al 2021, De Bortoli 2022, Chen et al 2022 …)
- And much much more …

Shameless plug - Presented Wednesday

In this paper we introduce a novel unifying framework for diffusion-based models, that engulfs both sampling and generative modelling. Additionally, we also make connections to statistical mechanics (Crooks Fluctuation Theorem / Jarzynski Equality) and sequential importance sampling.

Transport, Variational Inference and Diffusions: with Applications to Annealed Flows and Schrödinger Bridges

Francisco Vargas [1] **Nikolas Nüsken** [2]

Appendix

Feynman - Kac Formula

PDE Solving via MC – Path Integral

Consider the linear Parabolic PDE

$$v_0(x) = \phi(x)$$

$$\partial_t v_t(x) = -\sum_{i=1}^{d} \mu_i(t, x_i)\partial_{x_i} v_t(x) - \sum_{i,j=1}^{d} [\sigma\sigma^\top]_{ij}(t, x)\partial_{x_i,x_j} v_t(x)] + v_t(x)V(x,t) - f(x,t)$$

Then subject to Lip conditions it follows that

$$v_t(x) = \mathbb{E}_{X\sim Q}\left[\int_t^T e^{-\int_t^s V(X_s,s)\mathrm{d}r} f(X_s, s)\mathrm{d}s + e^{-\int_t^T V(X_r,r)\mathrm{d}r}\phi(X_T)\right]$$

with

$$\mathrm{d}X_t = \mu(X_t, t)\mathrm{d}t + \sigma(X_t, t)\mathrm{d}W_t$$

Doobs – Transform (Quick Version)

Proof Sketch – Part I : Transition Density

First condition and apply Bayes Theorem

$$p_{t+\delta|t,T}(z_{t+\delta}|z_t, z_T = z) = \frac{p_{T|t,t+\delta}(z_T = z|z_t, z_{t+\delta})p_{t+\delta|t}(z_{t+\delta}|z_t)}{p_{T|t}(z_T = z|z_t)}$$

Now the Markov property

$$p_{t+\delta|t,T}(z_{t+\delta}|z_t, z_T = z) = \frac{p_{T|t+\delta}(z_T = z|z_{t+\delta})p_{t+\delta|t}(z_{t+\delta}|z_t)}{p_{T|t}(z_T = z|z_t)}$$

Now we need to find an SDE with this transition density.

Doobs – Transform (Quick Version)

Proof Sketch – Part 2 : Space Time Regular

The h-transform satisfies (since it satisfies backward Kolmogorov):

$$\mathcal{A}_t(p_{T|t+\delta}(z_T = z|z_{t+\delta})) = 0$$

Doobs – Transform (Quick Version)

Proof Sketch – Part 3 : Finding the drift

Take time derivatives see what happens

$$\partial_t p_{t|s,T}(z_t|z_s, z_T = z) = \frac{1}{p_{T|s}(z_T = z|z_s)} \partial_t p_{T|t}(z_T = z|z_t) p_{t|s}(z_t|z_s)$$

$$= \frac{1}{p_{T|s}(z_T = z|z_s)} (p_{t|s}(z_t|z_s)\partial_t p_{T|t}(z_T = z|z_t) + p_{T|t}(z_T = z|z_t)\partial_t p_{t|s}(z_t|z_s))$$

$$= \frac{1}{h(z_s, s)} (-p_{t|s}(z_t|z_s)\mathcal{P}^\dagger h(z_t, t) + h(z_t, t)\mathcal{P} p_{t|s}(z_t|z_s))$$

$$= \frac{1}{h(z_s, s)} \left(\mathcal{P} h(z_t, t) p_{t|s}(z_t|z_s) + \nabla p_{t|s}(z_t|z_s) \cdot \nabla h(z_t, t) + p_{t|s}(z_t|z_s)\Delta h(z_t, t)\right)$$

$$= \frac{1}{h(z_s, s)} \left(\mathcal{P} h(z_t, t) p_{t|s}(z_t|z_s) + \nabla \cdot (h(z_t, t) p_{t|s}(z_t|z_s) \nabla \ln h(z_t, t))\right)$$