Operator Learning for the Fokker-Planck Equation with appications in Data Assimilation

Adam James Sheppard, Ruediger Schack

May 30, 2024

Contents

I Diffusion Processes 3
0.1 Ordinary Differential Equations 4
0.1.0.1 Solution Operators for Differential Equations in $\mathbb{R}^n$ 4
0.1.0.2 Boundedness of Differential Equations 5
0.2 Stochastic Differential Equations 5
0.2.0.1 The Wiener Process 6
0.2.1 Itô Calculus 10
0.2.1.1 Construction of the Itô Integral 10
0.2.2 Discussion 16
0.3 Further mathematical background 17
0.3.1 The Itô Formula 17
0.3.2 Discussion 18
0.3.3 Multidimensional Itô Formula 18
0.3.4 The Markov Model 19
0.3.5 Hidden Markov Models 20
0.3.6 The Fokker-Planck Equation 20
0.3.6.1 Infinitesimal Generators 20
0.3.6.2 Deriving the Fokker-Planck Equation 23
0.3.6.3 Generators or Fokker-Planck Equation? 29
0.3.6.4 One Dimensional Example of Fokker-Planck and Generator 29
0.3.6.5 Numerical Example: Stochastic Lorenz System 30

II Bayesian Data Assimilation 37
0.4 Introduction and Reason for Data Assimilation 38
0.5 Mathematical Formulation of Bayesian Data Assimilation 40
0.5.1 More on Hidden Markov Models 40
0.5.2 Bayesian Data Assimilation 41
0.5.3 Filtering - Brief History 41
0.5.4 Nonlinear Filtering for Diffusion Processes 42
0.5.5 Forecasting 42
0.5.6 Updating the distribution 43
0.5.7 Utilizing the Fokker-Planck Equation for Filtering 43
0.6 Stochastic Data Assimilation 44
0.7 Numerical Example: Stochastic Data Assimilation for Lorenz 1963 45

III Supervised Learning of Operators 49
0.8 Operator Learning 50
0.8.1 Mathematical Formulation of Operator Learning 50

A Further Mathematical Definitions **53**
A.1 Analysis and Measures 53

Part I

Diffusion Processes

Here we give a brief introduction to diffusion processes, which are a class of differential equations which allow for random variables within the formulation and solution. An object is diffusing if it experiences some erratic and disordered motion through the n-dimensional real-numbers. Diffusion processes play a central role in the theory of Data Assimilation.

0.1 Ordinary Differential Equations

Many applications of dynamical systems, ordinary and partial differential equations are used to model phenomena and describe the evolution of an object. Complete information about the how a system evolves over time is provided by a variable $\mathbf{X}(t)$ called the state variable - essentially describing the condition the system is in at any given time t. Given some state variable, which is a vector-function parameterised by time in the n-dimensional real-numbers, we can express the derivative of the state as a function $\mathbf{f}$ of $\mathbf{X}$ and some parameter set θ

$$\frac{d\mathbf{X}}{dt} = \mathbf{f}\left(\mathbf{X}, t; \theta\right)$$

which is given in matrix form as

$$\frac{d}{dt}\begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{pmatrix} = \begin{pmatrix} f_1(\mathbf{X}, t; \theta) \\ f_2(\mathbf{X}, t; \theta) \\ \vdots \\ f_n(\mathbf{X}, t; \theta) \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{pmatrix}$$

where $\mathbf{f}$ is a function of time t and position $\mathbf{X}$ and θ which is additional information about $\mathbf{X}$ which is directly measurable or estimated. If the initial conditions for this differential equation are provided - this differential equation has a solution.

However, in practice in order to estimate θ based on some observations of the state $\mathbf{X}$ other factors should be taken into account. In many cases, the values of these parameters are not directly measurable but need to be inferred from observed data of the system's state $\mathbf{X}$. This process is known as parameter estimation or system identification.

0.1.0.1 Solution Operators for Differential Equations in $\mathbb{R}^n$

Generally we consider the differential equation with continuous first order derivatives in the n-dimensional real numbers

$$\frac{d\mathbf{x}}{dt}(t) = f(\mathbf{x}(t)), \quad \mathbf{x}_0 = u$$

and assuming a solution exists for all u and $t > 0$; for a given u. An equilibrium point for the differential equation is a point $\mathbf{x}^*$ such that $f(\mathbf{x}^*) = \mathbf{0}$. Initialising the equation at $u = \mathbf{x}^*$ results in a solution $u(t) = \mathbf{x}^*$ for all $t \geq 0$.

If the differential equation has a solution for every equilibrium point $u \in \mathbb{R}^n$ and every $t \in \mathbb{R}^+$ then we have the following properties for an operator Ψ defined as the solution of the differential equation with a parameter t being time.

1. $u(t) = \Psi(\mathbf{x}^*; t), \quad t \geq 0$
2. $\Psi(\Psi(\mathbf{x}^*; s); t) = (\Psi \circ \Psi)(\mathbf{x}^*;\ s, t) = \Psi(\mathbf{x}^*;\ t + s)$
3. $\Psi(\Psi(\mathbf{x}^*; t); -t) = \Psi(\Psi(\mathbf{x}^*; -t); t) = \Psi(\mathbf{x}^*; 0) = \mathbf{x}^*$
4. $\Psi\left(\Psi(\Psi(\mathbf{x}^*; s); t); r\right) = \Psi\left(\Psi(\Psi(\mathbf{x}^*; s); r); t\right) = \Psi(\mathbf{x}^*;\ t + s + r)$

In general, properties (1)-(4) form a (semi)group (appendix A) of nonlinear (or linear) operators of a single parameter. The map $\Psi(\cdot,\cdot)$ is said to be a solution operator for the ordinary differential equation if there exists a function $u(t)$ which satisfies (1)-(4).

To give some intuition about why groups come into the picture we give an analogy. A ball is rolling on a landscape, the landscape represents a function $f(x)$ which is a differential equation. The steeper the slope the faster the ball rolls in that direction (think positive or negative being the direction in 1D). The balls position at some time t represents the solution $u(t)$ of the differential equation. The starting position of the ball $u(0)$ is denoted by x^*. This is an equilibrium point, where the ball wouldn't move because the slope is zero and hence $f(u(0) = f(x^*) = 0$

Now we define a mathematical object Ψ which is a operator, the property of the operator is that it takes some point equilibrium point x^* and outputs a function $u(t)$ to give the position of the ball at any time after the point x^*. Furthermore, if we start away from an equilibrium point at a point, say, v then the operator Ψ still returns a $u(t)$ after the point v and describes the dynamical behaviour thereafter. In essence, without an equilibrium point as the starting point, the solution operator becomes a tool to track the ball's motion

0.1.0.2 Boundedness of Differential Equations

Suppose for the differential equation above that there exists $\alpha, \beta \geq 0$ such that

$$\begin{aligned} \langle f(\mathbf{x}), \mathbf{x} \rangle &\leq \alpha + \beta |\mathbf{x}|^2 \\ &= \alpha + \beta \left(\sqrt{\sum_{k=0}^{n} x_k(t)} \right)^2 \end{aligned}$$

where $|\cdot|$ is the Euclidean norm on $\mathbb{R}^n$ (in a vector space or for matrices then Frobenius norm or inner product may be used) then,

$$\begin{aligned} \frac{1}{2}\frac{d|\mathbf{x}|^2}{dt} &= \left\langle \mathbf{x}, \frac{d\mathbf{x}}{dt} \right\rangle \\ &= \langle \mathbf{x}, f(\mathbf{x}) \rangle \\ &\leq \alpha + \beta |\mathbf{x}|^2 \end{aligned}$$

This expression puts a bound on the growth of a differential equation over time. This means it can not grow rapidly By bounding this growth rate, we prevent unbounded or explosive behaviour in the system. Without such bounds, solutions to the differential equation might exhibit erratic behavior, making it difficult to analyze or control the system. The inequality provides insights into the long-term behavior of the system by bounding the growth of the systems solutions. Data assimilation algorithms can use this information to make more accurate long-term forecasts while ensuring that the predictions remain consistent with observed data and stable over time. As we shall see further on this bound ensures stability in the evolution of probability densities described by the Fokker-Planck equation. Stability is crucial in stochastic processes to prevent probability densities from diverging or becoming unbounded over time.

0.2 Stochastic Differential Equations

Consider a simple of example of population growth given by the differential equation

$$\frac{dN}{dt} = \alpha(t)N(t), \quad N(0) = N_0$$

where N is the size of the population at some time t and α is the rate of growth of N at the same time t. Now assuming we do not know all information about α then we can introduce a parameter set θ

$$\frac{dN}{dt} = \alpha(t;\theta)N(t)$$

however, the parameter set may be an infinite set which is uncountable, there are many factors that can affect population growth - especially in todays societies. It may be tempting, for simplification to add some random 'noise' the model to account for all uncertainty with respect to our incomplete parameter set θ. As such we remodel α into

$$\alpha(t) = v(t) + \eta$$

where η is distributed according to some probability distribution and represents the uncertainty in the growth rate α that can not be fully explained by the deterministic model; we can model some aspects of randomness, but there are limitations to our knowledge about the underlying processes, we do not know the exact behaviour of the noise term η. $v(t)$ is assumed to be deterministic - nonrandom. Therefore, our population model becomes

$$\begin{aligned}\frac{dN}{dt} &= \alpha(t)N(t)\\ &= \left(v(t) + \eta\right)N(t)\end{aligned}$$

the equation we have obtained is a 'stochastic differential equation' - one in which randomness in the coefficients is allowed. Since the input of the equation involves some variable distributed according to a probability distribution - then it is reasonable to expect the solution to $N(t)$ to now also be distributed by some probability distribution. We can only then express qualitatively - something about the solution as a probability distribution of the solutions. Instead of obtaining a single trajectory for $N(t)$, we express its behaviour using probability distributions to account for the range of possible outcomes and the likelihood of them.

0.2.0.1 The Wiener Process

For our general differential equation

$$\frac{dX}{dt} = f\left(X, t; \theta\right) \tag{0.2.1}$$

supposing we add some random variable (or noise) term to the input and attempt to account for unexplained fluctuations in the output, for such a construction we would obtain

$$\frac{dX}{dt} = f(X, t; \theta) + g(X, t; \theta)W(t) \tag{0.2.2}$$

where g is a deterministic function and $W(t)$ is some stochastic processes - a collection of random variables which are indexed by time. However, for this equation to have a solution the random term W must satisfy some constraints. The term W must be a **Wiener Process** which is charaterised by:

1. $\mathbb{P}\{W(0) = 0\} = 1$, i.e. the process must start at the origin - there is no noise present at $t = 0$ in the sense that the g term does cause any initial displacement in $\frac{dX}{dt}$ at $t = 0$.

2. The increments $W\left(t_1\right) - W\left(t_0\right), \ldots, W\left(t_n\right) - W\left(t_{n-1}\right)$ must be mutually independent for some time interval $0 \leq t_0 < \cdots < t_n$ ensuring the behaviour at a future time interval is only influenced by current behaviour - i.e. independent of its past history. This is an example of the Markovian Property

(a) This is useful for complex systems modeling where instances may arise where some disturbances occur at different times which are not correlated which each other - allowing for modeling of systems that have multiple sources of uncertainty that evolve independently over time.

3. The difference operator Δ applied to W is normally distributed such that

$$\begin{aligned}\Delta W &= W(t+h) - W(t)\\ \mathbb{E}\left[\Delta W\right] &= 0\\ \mathbb{V}\left[\Delta w\right] &= \sigma^2 h\end{aligned}$$

(a) Where $h > 0$ is some small number and σ^2 is the variance of W in the increment. We can define $t + h = s$ and such we have

$$W(t+h) - W(t) = W(s) - W(t)$$

(b) The expected value being zero indicates that there is no bias in the increments

$$\begin{aligned}\lim_{h\to 0}\left\{\mathbb{E}\left[W(t+h) - W(t)\right]\right\} &= \lim_{h\to 0}\left\{\mathbb{E}\left[W(t) - W(t)\right]\right\}\\ &= \lim_{h\to 0}\left[0\right]\\ &= 0\end{aligned}$$

(c) The variance, which is proportional to the time increment h (and thus the size of the interval considered), reflects the erratic random fluctuations over time and that these fluctuations are continuous. Since $s > t$ we can re-write the variance

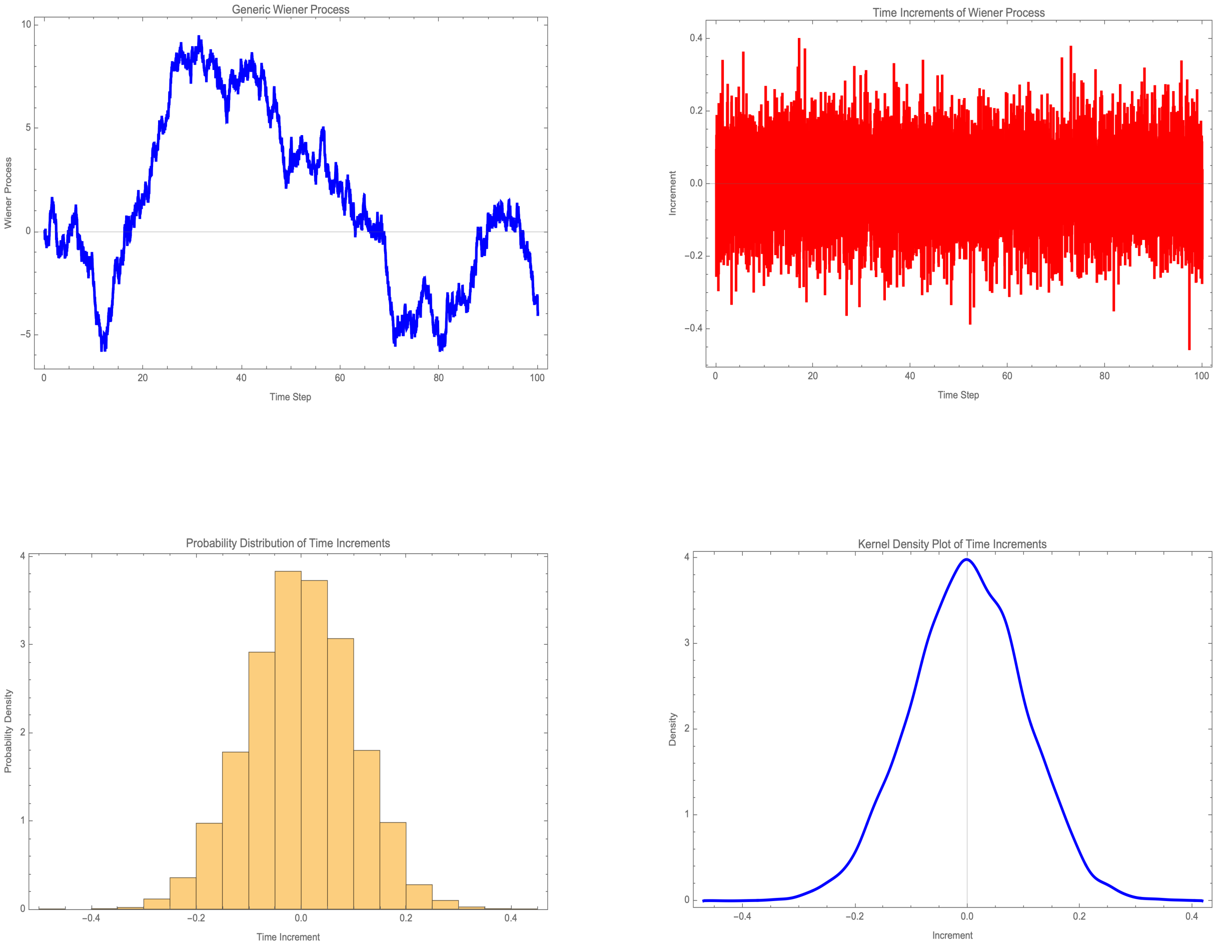
$$\begin{aligned}\mathbb{V}\left[W(s) - W(t)\right] &= \sigma^2|s-t|\\ &= \mathbb{E}\left[W(s)W(t)\right]\\ &= \mathbb{E}\left[\left[(W(s))^2 + W(s)\right]\left[W(t) - W(s)\right]\right]\\ &= \mathbb{E}\left[(W(s))^2\right] + 0\\ &= \mathbb{V}\left[W(s)\right](s-t)\,, \quad t, s > 0\\ &= \sigma^2\left|s-t\right|\\ &= \sigma^2\left|t+h-t\right|\\ &= \sigma^2 h\end{aligned}$$

therefore, it is essential that the provided term W is indeed a Wiener Process to ensure consistency and continuity as the system evolves, this allows for further mathematical formulation of the solutions to (2). Providing a realisations of a Wiener Process we can see that the Wiener Process has a Gaussian distribution

Remark 0.1. For property (1) of the Wiener Process that the definition of the starting point $W(0) = 0$ may be extended to allow for a more general starting point, say, $W(0) = w$ - however considerations to the offset in the initial condition of the differential equation must also be accounted for in this scenario.

We shall now focus on the distributions of a Wiener Process W. Suppose we are given a Wiener process $W(u) = x$ where u represents the time and is positive or zero and x is a real-valued variable. Conditional on this is $W(t)$ which is normally distributed with $\mathcal{N}(x, t-u)$ for all $t \geq u$ - this is essentially considering the same Wiener Process at two different times, t and u where t is at a later time than u - we can develop a distribution function F for $W(t)$.

Figure 0.2.1: Wiener Process and Gaussian Increments



Given $W(u) = x$ and for all $u \geq 0$ and $u \in \mathbb{R}$ then the conditional cumulative distribution function is given by

$$F(t, y|u, x) = \mathbb{P}\left(W(t) \leq y|W(u) = x\right)$$

which has density function

$$\frac{\partial F}{\partial y}$$

this represents the rate of change of the probability of an event occurring with respect to some variable y.

We know that $W(t) - W(t+h)$ is distributed normally and follows a normal distribution with mean 0 and variance h from property $3(a)$-Gaussian increments, therefore, $W(t) - x$ is also normally distributed with mean x and variance $(t-u)$ (see remark 1) via the shifting property for normal distributions. So we have a normal Probability Density Function (PDF) as

$$\rho(t) = \frac{\partial F}{\partial y} = \frac{1}{\sqrt{2\pi(t-u)}} \exp\left\{-\frac{(y-x)^2}{2(t-u)}\right\}$$

Example 0.1. If $W(t)$ is a Wiener process in the real numbers $\mathbb{R}$, then

$$\begin{aligned}
\mathbb{E}\left|\Delta W\right|^p &= \mathbb{E}\left|W(t) - W(s)\right|^p \\
&= \frac{1}{\sqrt{2\pi|t-s|}} \int_{\mathbb{R}} |x|^p \exp\left\{-\frac{|x|^2}{2|t-s|}\right\} dx \\
&= \frac{1}{\sqrt{2\pi|t-s|}} \int_{-\infty}^{\infty} |x|^p \exp\left\{-\frac{|x|^2}{2|t-s|}\right\} dx
\end{aligned}$$

where $p \geq 0$, by using the substitution

$$\frac{x}{\sqrt{|t-s|}} = y \text{ and } dx = \sqrt{|t-s|}dy$$

then

$$\begin{aligned}
\mathbb{E}|\Delta W|^p &= \frac{\left(\sqrt{|t-s|}\right)^p}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |x|^p \exp\left(-\frac{|x|^2}{2}\right) dx \\
&= |t-s|^{p/2} \int_{-\infty}^{\infty} |x|^p \exp\left(-\frac{|x|^2}{2}\right) dx \\
&= C|t-s|^{p/2} \\
&= C|\Delta t|^{p/2} \qquad (0.2.3)
\end{aligned}$$

where

$$C = \int_{-\infty}^{\infty} |x|^p \exp\left(-\frac{|x|^2}{2}\right) dx \qquad (0.2.4)$$

which still possesses a Gaussian distribution.

0.2.1 Itô Calculus

If we re-write equation (2) as a difference equation, we can approximate the behaviour of the system by using a finite difference approximation - we employ the Euler-Maruyama method to obtain

$$\begin{aligned}\frac{d\mathbf{X}}{dt} &\approx \frac{\mathbf{X}(t_{n+1})-\mathbf{X}(t_n)}{\Delta t} && (0.2.5)\\ &= \mathbf{f}(\mathbf{X}(t_n),t_n;\theta)+g(\mathbf{X}(t_n),t_n;\theta)W(t_n) && (0.2.6)\\ \mathbf{X}(t_{n+1})-\mathbf{X}(t_n) &= \mathbf{f}(\mathbf{X}(t_n),t_n;\theta)\Delta t+g(\mathbf{X}(t_n),t_n;\theta)W(t_n)\Delta t && (0.2.7)\\ &= \mathbf{f}(\mathbf{X}(t_n),t_n;\theta)\Delta t+g(\mathbf{X}(t_n),t_n;\theta)\Delta W\,(t_n) && (0.2.8)\end{aligned}$$

The general form for the difference equation representing components of $\mathbf{X}$ at time $n+1$ in terms of the components at time n for time step $\Delta t_j = t_{j+1}-t_j$ for $j \geq 0$ and increments of the Wiener process $\Delta W(t_j) = W(t_{j+1}) - W(t_j)$ and $\Delta W(t_0) = 0$ we obtain

$$\begin{aligned}\mathbf{X}_0 &= \mathbf{f}_0(\mathbf{X}_0,0;\theta)\Delta t_0\\ \mathbf{X}_1(t_1) &= \mathbf{X}_0+\mathbf{f}_1(\mathbf{X}(t_1),t_n;\theta)\Delta t_1+g_1(\mathbf{X}(t_1),t_1;\theta)\Delta W\,(t_1)\\ \mathbf{X}_2(t_2) &= \mathbf{X}_0+\mathbf{f}_2(\mathbf{X}(t_2),t_2;\theta)\Delta t_2+g_2(\mathbf{X}(t_2),t_2;\theta)\Delta W\,(t_2)\\ &\vdots\\ \mathbf{X}_n(t_n) &= \mathbf{X}_0+\mathbf{f}_{n-1}(\mathbf{X}(t_{n-1}),t_{n-1};\theta)\Delta t+g_{n-1}(\mathbf{X}(t_{n-1}),t_{n-1};\theta)\Delta W\,(t_{n-1}) && (0.2.9)\end{aligned}$$

where we can express this as a sum over $n-1$ elements with the additional final term

$$\mathbf{X}(t_n) = \mathbf{X}_0+\sum_{j=0}^{n-1}\mathbf{f}_j(\mathbf{X}(t_j),t_j;\theta)\Delta t_j+\sum_{j=0}^{n-1}g_j(\mathbf{X}(t_j),t_j;\theta)\Delta W\,(t_j) \qquad (0.2.10)$$

since we know that each ΔW_j is independent and normally distributed with mean 0 and and is continuous, then for all $t_1 < t_2 < \cdots < t_k = t$ we have the continuous analogue of the discrete-time solution for $\mathbf{X}(t)$ as $\Delta t_j \to 0$

$$X(t) = X_0+\int_0^t f(X(s),s;\theta)ds+\int_0^t g(X(s);\theta)dW(s) \qquad (0.2.11)$$

The existence of the integral

$$\int_0^t g(X(s);\theta)dW(s) \qquad (0.2.12)$$

will be assumed without proof, for the proof see [1, 2]. So we shall adopt that the solution to the stochastic differential equation (2) is a stochastic process which satisfies (3).

0.2.1.1 Construction of the Itô Integral

To construct such an integral it is reasonable to start by defining the integral as a Riemann-Stieltjes sum

$$\begin{aligned}\int_0^t g(X(s);\theta)dW(s) &= \lim_{n\to\infty}\sum_{j=0}^{n-1} g\left(t_j\right)\Delta W\left(t_j\right) && (0.2.13)\\ &= \lim_{n\to\infty}\sum_{j=0}^{n-1} g\left(t_j\right)\left[W\left(t_{j+1}\right)-W\left(t_j\right)\right] && (0.2.14)\end{aligned}$$

Recall that the Wiener process has a variance which is proportional to the interval it is defined on. This means that the fluctuation or variation in the value of $W(t)$ over the interval becomes larger, making it increasingly difficult to approximate the integral using a Riemann sum.

As the length of the interval over which you're integrating increases, the increments of the Wiener process also increase in size due to its property of having a variance proportional to the length of the interval. These larger increments result in larger variations in the values of the Wiener process.

The Riemann sum relies on subdividing the interval into smaller sub-intervals, but the erratic behavior of the Wiener process makes it difficult to accurately capture its behavior over these sub-intervals. So as the sub-intervals are divided into discrete subsets the total amount of erratic behaviour inside the sub-intervals increases, therefore, a basic approximation scheme based upon rectangles or trapezoids will not capture the complex geometric shapes which are present in the intervals as they get larger.

Definition 0.1. A continuous function $f : [0,1] \to \mathbb{R}$ is a function of bounded variation if

$$V_f^{(1)}(t) = \sup \left\{ \sum_{j=1}^{k} |f(t_j) - f(t_{j-1})| \right\} < \infty \tag{0.2.15}$$

i.e it finds the largest possible total jump (up and down) the function makes as you move between the specified points. Which is the maximum total vertical displacement in each time interval.

Therefore, the concept of bounded variation ensures that the total "jumps" (up and down) of the function throughout the interval have a finite (limited) sum.

This gives a method of measuring how “bumpy” or “irregular” a function is when it fluctuates on the vertical axis within some specified points.

Theorem 0.1. *The variation of a Wiener Process path does not converge, with probability one*

$$\lim_{n\to\infty} \sum_{j=0}^{n-1} [W(t_{j+1}) - W(t_j)] \to \infty \tag{0.2.16}$$

Proof. Suppose that the sequence of partitions $0 \le t_0 < \cdots < t_n \le \tau$ is a nested partition, that is, at each step, say some interval between $[0,\tau]$ given by the times $t_k$and t_{k+1}

$$0 < \cdots < t_k < t_{k+1} < \cdots < \tau$$

inside the interval $[t_k, t_{k+1}]$ we partition l times $t_k \le u_1 < \cdots < u_l \le t_{k+1}$ and so on for every sub-interval in $[0,\tau]$ and intervals generated therein, more formally we have:

Consider the sequence of nested partition of the interval $[0,\tau]$ denoted by $\{t_k\}_{k=0}^{n}$ where each partition $[t_k, t_{k+1}]$ is partitioned into l_k sub-intervals then as the number of sub-intervals $l_k \to \infty$ then the length of sub-intervals $t_{k+1} - t_k \to 0$ denoted as:

$$\lim_{k\to\infty} \sup_{1\le j\le k} \{t_{j+1} - t_j\} \to 0 \tag{0.2.17}$$

This means that the partitions become finer and finer as k increases, resulting in smaller and smaller intervals between each t_j and t_{j+1} , then By the properties of the Wiener process, we know that the increment $W(t_{j+1}) - W(t_j)$ is normally distributed with mean 0 and variance $t_{j+1} - t_j$ therefore, As the length of each sub-interval tends to zero, the variance of each increment also tends to zero, the expected value of $[W(t_{j+1}) - W(t_j)]^2$ is $t_{j+1} - t_j$

$$\lim_{k\to\infty}\left\{\sum_{j=0}^{k}\mathbb{E}\left[W\left(t_{j+1}\right)-W\left(t_{j}\right)\right]^{2}\right\}=\lim_{k\to\infty}\left\{\sum_{j=0}^{k}\left(t_{j+1}-t_{j}\right)\right\}\to\tau \tag{0.2.18}$$

which is just the length of the original interval, but now

$$\limsup_{\tau\to\infty}\left\{\sum_{j=0}^{k}\left[W\left(t_{j+1}\right)-W\left(t_{j}\right)\right]^{2}\right\}\to\infty \tag{0.2.19}$$

so as $\tau\to\infty$ and the number of nested partitions increase so will the limit of the sum. Therefore, the Wiener process has unbounded variation. While the variance of each increment tends to zero, the number of increments within each partition (i.e., the number of terms in the sum) increases. This is because as τ approaches infinity, the number of partitions (represented by k) also increases.

Therefore, even though each increment becomes smaller (due to the decreasing variance), the number of increments increases, and this leads to the sum of squares of the increments diverging to infinity as τ approaches infinity because the increasing number of partitions leads to a net increase in the overall variance contribution from all the increments. □

As it can be seen from Theorem 1 the term

$$\int_0^t g(X(s;\theta)dW(s)$$

can not be interpreted using the well-known Riemann-Stieltjes sum the theorem and proof demonstrate that the total variation of a Wiener process path across its entire domain diverges to infinity as the number of partitions and interval length increase. This means that the sum of absolute value changes across the entire path becomes infinitely large, not zero.

We aim to define the stochastic integral. This integral should have the properties that for

$$g(X(1);\theta)=1\implies\int_0^t dW(s)=W(t)-W(0)$$

In this way one can integrate a constant process with respect to a Wiener Process. Sums of integrals over different integrals should also be possible. This means that the stochastic integral is defined for process which are constant on a finite number of intervals. Then by the limit the integral can then be defined for more general processes.

If we take a non-random process $\mathcal{X}(t)$ which is still a function of time but does not depend on W then we can partition the integral $[0,t]$ such that

$$\mathcal{X}(t)=\sum_{j=0}^{n}c_j\mathbb{I}_{(t_i,t_{j+1}]}(t) \tag{0.2.20}$$

where $\mathbb{I}$ is an indicator function which tells us whether t is in the sub-interval (a fine partition of $[0,t]$) and is given by

$$\mathbb{I}_{(t_i,t_{j+1}]}=\begin{cases}1 & t\in(t_j,t_{j+1}]\\ 0 & t\notin(t_j,t_{j+1}]\end{cases},\forall j\in\mathbb{N}\cup\{0\} \tag{0.2.21}$$

it is essentially another way to define (or approximate) any function in terms of whether a functions output is contained in a range of interest, using this idea the Itô integral can be defined as

$$\int_0^t \mathcal{X}(s)dW(s) = \int_0^t \sum_{j=0}^{n} c_j \mathbb{I}_{(t_i,t_{j+1}]}(s)dW(s) \tag{0.2.22}$$

$$= \sum_{j=0}^{n} c_j \left(\int_{t_j}^{t_{j+1}} dW(s) \right) \tag{0.2.23}$$

$$= \sum_{j=0}^{n} c_j \left(W(t_{j+1}) - W(t_j) \right) \tag{0.2.24}$$

$$= \sum_{j=0}^{n} c_j \Delta W(t_j) \tag{0.2.25}$$

for some constants c_j. Furthermore, since we know that the difference operator applied to the Wiener process is a Gaussian Random Variable with mean zero and variance

$$\mathbb{V}\left[\int_0^t \mathcal{X}(s)dW(s)\right] = \mathbb{V}\left[\sum_{j=0}^{n} c_j \left(W(t_{j+1}) - W(t_j) \right)\right] \tag{0.2.26}$$

$$= \sum_{j=0}^{n} \mathbb{V}\left[c_j \left(W(t_{j+1}) - W(t_j) \right)\right] \tag{0.2.27}$$

$$= \sum_{j=0}^{n} c_j^2 \left(t_{j+1} - t_j \right) \tag{0.2.28}$$

$$= \sum_{j=0}^{n} c_j^2 \Delta t_j \tag{0.2.29}$$

If we now replace constant functions c_j with non-constant but deterministic function g to obtain

$$\sum_{j=0}^{n} g(s,t_j) \left(W(t_{j+1}) - W(t_j) \right) \to \int_0^t g(X(s);\theta)dW(s) \tag{0.2.30}$$

where the function $g(s,t_j)$ must satisfy both

1. Lipschitz Continuity - a property of functions that captures how much of the functions output changes with respect to some change in it's input. Say for x and y some input then it should be less than some number L which scales the input.

   $$|f(y,t) - f(x,t)| + |g(y,t) - g(x,t)| \leq L|x-y|$$

2. Bounded Growth Condition, which refers to the property of functions where values do not grow beyond a certain point, for any input x. If there exists some upper bound M such that for all x in the domain of g then

   $$|g(x,t)| \leq M$$

So noting that g must be a bounded continuous function then one can define a set over all time t such that the history of the Wiener process up to time t is captured inside it. Such a set would take the form of

$$\{s : B_{t_1}(s) \in \mathbb{R}^n, \ldots, B_{t_k}(s) \in \mathbb{R}^n, t_k \leq t\}$$

This says that a reasonable definition of an integral in the Itô sense will have a successful approximation to g provided that the functions $g(s, t_j)$ only depend on the behaviour of the Wiener process up to t_j. Therefore, an integral involving g cannot have expectations involving future values for the stochastic process.

This all together forms a definition of the type of function that we can integrate with respect to a Wiener process, that is

Definition 0.2. Let $\mathcal{G}(s, t)$ be the class of functions

$$g(s, t) : [0, \infty) \times \mathbb{R} \to \mathbb{R}$$

such that

1. g is bounded and continuous
2. $g(s, t-1)$ only depends on the time up to time $t-1$
3. The expected value of any integrable interval is finite

$$\mathbb{E}\left\{\int_a^b g(s,t)dt\right\} < \infty$$

These assumptions on g can be nicely charaterised to give existence and uniqueness of solutions of an equation. We enforce that there exists an $\alpha, \beta \in \mathbb{R}$ such that

$$\langle g(X,t), X\rangle \leq \alpha + \beta X^2$$

where $\alpha, \beta > 0$ later showing that the test models prescribed in the following sections do follow this constraint.

Example 0.2. We can show the following integral exists

$$\int_0^t W(s)dW(s) = \frac{1}{2}W(t)^2 - \frac{1}{2}t$$

Take $g(X,t) = W(t)$ then by (2.30) we can express the integral as a sum in the following sense

$$\int_0^t W(s)dW(s) = \lim_{n\to\infty}\left\{\sum_{j=0}^{n} W(t_j)\left(W(t_{j+1}) - W(t_j)\right)\right\}$$

however, we now need to set up an approximation procedure. We need to partition the interval $[0, t]$ into a n finite sub-intervals of equal length. Define

$$W\left(\frac{jt}{n}\right) \leq W(t) \leq W\left(\frac{(j+1)t}{n}\right)$$

for $\frac{jt}{n} \leq t < \frac{(j+1)t}{n}$ and $n \in \mathbb{N}$. The partitioned interval is half-open to avoid double counting points at each t_j this is done to explicitly emphasize how such an interval may be partitioned. So now

$$\begin{aligned}\int_0^t W(s)dW(s) &= \lim_{n\to\infty}\left\{\sum_{j=0}^{n} W(t_j)\left(W(t_{j+1})-W(t_j)\right)\right\}\\ &= \lim_{n\to\infty}\left\{\sum_{j=0}^{n} W\left(\frac{jt}{n}\right)\left[W\left(\frac{(j+1)t}{n}\right)-W\left(\frac{jt}{n}\right)\right]\right\}\end{aligned}$$

By Theorem 1 we know that the limiting sum

$$\lim_{n\to\infty}\left\{\sum_{j=0}^{n}\left[W\left(\frac{(j+1)t}{n}\right)-W\left(\frac{jt}{n}\right)\right]^2\right\}=t$$

therefore, if we can express the integral in terms of this summation then we can use the variation result from theorem 1 to conclude the result of the integrand. Noting that here we have normally distributed incremenets with

$$\begin{aligned}W\left(\frac{(j+1)t}{n}\right)-W\left(\frac{jt}{n}\right) &\sim \mathcal{N}\left(0,\frac{(j+1)t}{n}-\frac{jt}{n}\right)\\ &= \mathcal{N}\left(0,\frac{t}{n}\right)\end{aligned}$$

and so

$$\mathbb{V}\left[W\left(\frac{(j+1)t}{n}\right)-W\left(\frac{jt}{n}\right)\right]=\frac{t}{n}$$

which is the length of the time intervals and for $n\to\infty$ then $\mathbb{V}[\Delta W]\to 0$ and as $t\to\infty$ we have unbounded variance and also variation. Thus restricting ourselves to a finite time interval. This is true for all partitions and thus true for the limit too. Manipulating the term $W(t_j)\left(W(t_{j+1})-W(t_j)\right)$ into the sort seen in theorem 1 can be done via noticing that

$$\begin{aligned}W(t_j)\left(W(t_{j+1})-W(t_j)\right) &= \frac{1}{2}\left(W^2(t_{j+1})-W^2(t_j)\right)-\underbrace{\frac{1}{2}\left(W(t_{j+1})-W(t_j)\right)^2}_{\text{Term we want}}\\ &= \frac{1}{2}\left(W^2\left(\frac{(j+1)t}{n}\right)-W^2\left(\frac{jt}{n}\right)\right)-\frac{1}{2}\left(W\left(\frac{(j+1)t}{n}\right)-W\left(\frac{jt}{n}\right)\right)^2\end{aligned}$$

$$\frac{1}{2}\left(W\left(\frac{(j+1)t}{n}\right)-W\left(\frac{jt}{n}\right)\right)^2=\frac{1}{2}\left(W^2\left(\frac{(j+1)t}{n}\right)-W^2\left(\frac{jt}{n}\right)\right)-W\left(\frac{jt}{n}\right)\left(W\left(\frac{(j+1)t}{n}\right)-W\left(\frac{jt}{n}\right)\right)$$

now that the expression has been manipulated into a familiar form we can use theorem 1 to conclude a result for the integral, first simplifying the right hand side into a better shape

$$\begin{aligned}&\frac{1}{2}\left(W^2\left(\frac{(j+1)t}{n}\right)-W^2\left(\frac{jt}{n}\right)\right)-W\left(\frac{jt}{n}\right)\left(W\left(\frac{(j+1)t}{n}\right)-W\left(\frac{jt}{n}\right)\right)\\ &\quad=\frac{1}{2}W^2\left(\frac{(j+1)t}{n}\right)-W\left(\frac{(j+1)t}{n}\right)W\left(\frac{jt}{n}\right)+\frac{1}{2}W^2\left(\frac{jt}{n}\right)\end{aligned}$$

which is a quadratic form. Moreover noting that when $j=0$ then $W(0)=W_0$

$$\begin{aligned}\frac{1}{2}\sum_{j=0}^{n}\left(W\left(\frac{(j+1)t}{n}\right)-W\left(\frac{jt}{n}\right)\right)^2 &= \overbrace{\frac{1}{2}\sum_{j=0}^{n}W^2\left(\frac{(j+1)t}{n}\right)}^{\text{Increment by one}}-\sum_{j=0}^{n}W\left(\frac{(j+1)t}{n}\right)W\left(\frac{jt}{n}\right)+\underbrace{\frac{1}{2}\sum_{j=0}^{n}W^2\left(\frac{jt}{n}\right)}_{\text{join with first term}}\\ &= \frac{1}{2}W^2\left(t+\frac{t}{n}\right)+\frac{1}{2}\sum_{j=0}^{n}W^2\left(\frac{jt}{n}\right)\\ &\quad -\sum_{j=0}^{n}W\left(\frac{(j+1)t}{n}\right)W\left(\frac{jt}{n}\right)+\frac{1}{2}\sum_{j=0}^{n}W^2\left(\frac{jt}{n}\right)\\ &= \frac{1}{2}W^2\left(t+\frac{t}{n}\right)+\frac{1}{2}\sum_{j=0}^{n}W^2\left(\frac{jt}{n}\right)-\sum_{j=0}^{n}W\left(\frac{(j+1)t}{n}\right)W\left(\frac{jt}{n}\right)\\ &= \frac{1}{2}W^2\left(t+\frac{t}{n}\right)+\sum_{j=0}^{n}W\left(\frac{jt}{n}\right)\left(W\left(\frac{jt}{n}\right)-W\left(\frac{(j+1)t}{n}\right)\right)\end{aligned}$$

So by rearranging terms on the RHS with the term on the LHS we arrive at the desired expression for the integral

$$\sum_{j=0}^{n}W\left(\frac{jt}{n}\right)\left(W\left(\frac{jt}{n}\right)-W\left(\frac{(j+1)t}{n}\right)\right)=\frac{1}{2}W^2\left(t+\frac{t}{n}\right)-\frac{1}{2}\sum_{j=0}^{n}\left(W\left(\frac{(j+1)t}{n}\right)-W\left(\frac{jt}{n}\right)\right)^2$$

by taking the limit and invoking Theorem 1 we have the existence of the integral:

$$\begin{aligned}\int_0^t W(s)dW(s) &= \lim_{n\to\infty}\left\{\sum_{j=0}^{n}W\left(\frac{jt}{n}\right)\left[W\left(\frac{(j+1)t}{n}\right)-W\left(\frac{jt}{n}\right)\right]\right\}\\ &= \lim_{n\to\infty}\left\{\frac{1}{2}W^2\left(t+\frac{t}{n}\right)\right\}-\frac{1}{2}\lim_{n\to\infty}\left\{\sum_{j=0}^{n}\left(W\left(\frac{(j+1)t}{n}\right)-W\left(\frac{jt}{n}\right)\right)^2\right\}\\ &= \frac{1}{2}W^2(t)-\frac{1}{2}t\end{aligned}$$

as desired.

In this section the stochastic differential equation and the Itô integral was defined. However, during the following computations involving these types of integrals, the stochastic differential equations will only be interpreted through the Itô sense of the integral. Other interpretations do exist, however they are not addressed here.

0.2.2 Discussion

Diffusion processes are fundamental to data assimilation, it provides a rigorous framework to account for unknown information and error and include it in analysis. They facilitate the gradual dispersion and integration of observational data and model predictions, enabling the refinement of system state estimates and the reduction of uncertainties inherent in dynamic systems. Within state filtering and information fusion, data assimilation serves as a theoretical framework for fusing observation data and the case where one possesses a dynamical model which describes the process of interest to aid accurate future state prediction. It encompasses the iterative synthesis of diverse data modalities, optimal integration of observational evidence and model-derived predictions to yield refined estimates of system states.

State estimation for Linear Gaussian Processes, the Kalman Filter, stands as a cornerstone of state estimation. However, its efficacy hinges upon an understanding of diffusion processes and stochastic differential equations. These conceptual underpinnings govern the propagation of uncertainty and the refinement of state estimates. Furthermore, the propagation of uncertainty and time evolution of a probability density function for a stochastic differential equation is of upmost interest in this research.

0.3 Further mathematical background

In this section some useful definitions and brief summaries from topics will be introduced.

0.3.1 The Itô Formula

Following on from the definition of the Itô Integral the Fundamental Theorem of Calculus and the chain rule from the Riemannian perspective are preserved in this formulation too, in fact only on the basis of the fundamental theorem of calculus and the chain rule are explicit (nonnumerical) calculations possible in this framework involving stochastic integrals. Before looking at the formulas for stochastic integrals some further properties of the Wiener process will be stated but not proved here, see [2] for full derivations involving telescoping series.

Remark 0.2. For $\Delta W = W(t) - W(s)$ for $t > s$ and Δt we know is a Gaussian random variable.

The following useful identities are shown without proof

$$\mathbb{E}[\Delta W] = 0 \tag{0.3.1}$$

$$\mathbb{E}\left[(\Delta W)^2\right] = \Delta t \tag{0.3.2}$$

$$\mathbb{V}\left[\Delta W\right] = \Delta t \tag{0.3.3}$$

$$\mathbb{V}\left[(\Delta W)^2\right] = 2(\Delta t)^2 \tag{0.3.4}$$

Most notably the variance of the squared increment of the Wiener process is a function of just time. Thus $\Delta t \to 0$ then so does the variance, in fact, since $\Delta t < 1$ then $(\Delta t)^2 \to 0$ at a fast rate than $(\Delta W)^2 \to 0$. This means that the variability introduced by the stochastic process diminishes more rapidly than the process itself evolves. Δt becomes infinitesimally small, the variance diminishes quadratically with respect to Δt. This rapid decrease in variance relative to the process's behavior is a key aspect that leads to the behavior becoming more deterministic.

The benefit of this, when constructing an integral with respect to some stochastic process then one can observe the following relationships in the continuous case via generalisation

$$(dW(t))^2 = dt \tag{0.3.5}$$

$$dt \cdot dW(t) = 0 \tag{0.3.6}$$

Using equation (2) and multiplying through by dt we get the one dimensional *Itô process* given by

$$dX(t) = fdt + gdW(t)$$

where the parameters of f and g have been omitted. The *Itô formula* can be constructed with a second order Taylor Approximation of $dX(t)$ with the assumption that f and g are are contained in the class $\mathcal{G}$.

Let ϕ be in the class of functions with continuous first and seconder order partial derivatives over the real number line, denoted as $C^2(\mathbb{R})$ and define an Itô process

$$Y(t) = \phi(X(t), t)$$

then the Taylor expansion of Y with respect to x and t is given

$$\begin{aligned} dY(t) &= \sum_{n=0}^{\mathcal{D}} \left\{ \frac{1}{n!} \sum_{k=0}^{n} \begin{pmatrix} n \\ k \end{pmatrix} \frac{\partial^n \phi}{\partial t^{n-k} \partial x^k} dt^{n-k} dX^k \right\} && (0.3.7) \\ &= \sum_{n=0}^{2} \left\{ \frac{1}{n!} \sum_{k=0}^{n} \begin{pmatrix} n \\ k \end{pmatrix} \frac{\partial^n \phi}{\partial t^{n-k} \partial x^k} dt^{n-k} dX^k \right\} && (0.3.8) \\ &= \frac{\partial \phi}{\partial t} dt + \frac{\partial \phi}{\partial x} dX + \frac{1}{2} \frac{\partial^2 \phi}{\partial x^2} (dX)^2 + \frac{1}{2} \frac{\partial^2 \phi}{\partial t^2} (dt)^2 + \frac{\partial^2 \phi}{\partial x \partial t} dt \cdot dX && (0.3.9) \\ &= \frac{\partial \phi}{\partial t} dt + \frac{\partial \phi}{\partial x} dX + \frac{1}{2} \frac{\partial^2 \phi}{\partial x^2} (dX)^2 && (0.3.10) \end{aligned}$$

where $\mathcal{D}$ is the highest order of derivative that we are working with, so with respect to the random variable X taking values $x \in \mathbb{R}$ and time $t \geq 0$ where $t \in \mathbb{R}$ then $\mathcal{D} = 2$.

so now plugging in our expression for $dX(t)$ we have

$$dY(t) = \left(\frac{\partial \phi}{\partial t} + f \frac{\partial \phi}{\partial x} + \frac{g^2}{2} \frac{\partial^2 \phi}{\partial x^2} \right) dt + g \frac{\partial \phi}{\partial x} dW$$

0.3.2 Discussion

The above expression is relevant because it is much easier to work with than the stochastic integrals considered before i.e. a nonlinear combination of stochastic differentials. Overall, the Taylor expansion approach simplifies the representation of stochastic processes and makes them more amenable to analysis using traditional calculus techniques, which can often be more intuitive and computationally tractable than working directly with stochastic integrals. Please refer to example two for comparison and note the measure theoretic properties considered. We do note that due to the presence of stochastic differentials in the Taylor expansion that direct application of traditional calculus techniques are not applicable however the simplification through this method does enable one to bridge analysis and stochastic analysis. For direct application of traditional calculus refer to Stratonovich stochastic calculus [4] however either case comes with advantage and disadvantage [4] particular to note is that Stratonovich calculus is best suited when working on manifolds and when symmetry properties of stochastic processes hold true [4].

0.3.3 Multidimensional Itô Formula

Providing a generalisation of the Itô formula to handle multiple variables. These are heavily important in the following sections.

Definition 0.3. Let $\mathbf{W} = (W_1, \ldots, W_d)$ be a d-dimensional Wiener process and let $\mathbf{g}(\mathbf{w}, t) \in \mathbb{M}_{n \times m}(\mathbb{R})$ then

$$\int_a^b \mathbf{g}(\mathbf{w}, t) d\mathbf{W}(t) = \sum_{j=0}^{d} \int_a^b g_{ij}(w_{ij}, t) d\mathbf{W}_j(t) \tag{0.3.11}$$

where $\mathbf{g} : \mathbb{R}^n \times [0, \infty) \to \mathbb{R}^{n \times m}$ and is the extension of the class $\mathcal{G}$ to d-dimensional functions denoted as $\mathcal{G}^{m \times n}$. From the construction of the stochastic integration above the conditions above are still sufficient to carry out construction as before. Each g_{ij} satisfies conditions (1)-(3) laid out in the definition of $\mathcal{G}$.

Definition 0.4. Let f and g be Lipschitz then the multidimensional $m \times n$ Itô process is

$$\begin{pmatrix} dX_1 \\ \vdots \\ dX_m \end{pmatrix} = \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix} dt + \begin{pmatrix} g_{11} & \cdots & g_{1n} \\ \vdots & \ddots & \vdots \\ g_{m1} & \cdots & g_{mn} \end{pmatrix} \begin{pmatrix} dW_1 \\ \vdots \\ dW_n \end{pmatrix} \tag{0.3.12}$$

$$d\mathbf{X}(t) = \mathbf{f}(w,t)dt + \mathbf{g}(w,t)d\mathbf{W}(t) \tag{0.3.13}$$

with differential components corresponding to the multidimensional Itô formula.

$$dY_k = \frac{\partial \phi_k}{\partial t}dt + \sum_{i=0}^{n} \frac{\partial \phi_k}{\partial x_i} dX_i + \frac{1}{2}\sum_{i=0}^{n}\sum_{j=0}^{n} \frac{\partial^2 \phi_k}{\partial x_i \partial x_j} dX_i dX_j \tag{0.3.14}$$

where $dX_i dt = 0 = dt dX_j$.

0.3.4 The Markov Model

The Markov Model is a fundamental concept in the study of how things change over time. It provides a structured way to understand how systems evolve, whether it's a hopping rabbit in a garden or the fluctuations of stock prices.

A stochastic process $X(t)$ for $t \in T$ is called a Markov process if, for any parameter set in an internal $t_i < t_j$ for $i \neq j$ and for all $\lambda \in \mathbb{R}$ then

$$\begin{aligned} \mathbb{P}\left(X(w,t_j) \leq \lambda | X(w,t_1), X(w,t_2), \ldots, X(w,t_i)\right) &= \mathbb{P}\left(X(w,t_j) \leq \lambda | X(w,\tau); \tau \leq T\right) && (0.3.15) \\ &= \mathbb{P}\left(X(w,t_j) \leq \lambda | X(w,t_i)\right) && (0.3.16) \end{aligned}$$

which is the Markov Property, which simply means that the future of a system depends only on its present state, not on its entire history. This property makes the model incredibly powerful for making predictions and understanding complex processes. Mathematically, the states are contained in a set S which takes values from $1, \ldots, n$ where $n \in \mathbb{N} \cup \{0\}$.

Transition Probabilities
In a Markov chain, the dynamics of the system are charaterised by transition probabilities. These probabilities describe how the system changes from one state to another when evolving over time. These probabilities are summarised within a matrix P where P_{ij} represents the probability of transitioning from state i to state j. Each row of the matrix must sum to 1 which defines the matrix P as a stochastic matrix.

These probabilities delineate the likelihood of transitioning between states, akin to a roadmap guiding the evolution of the system over time. Through this lens, the Markov Model facilitates the quantitative analysis and prediction of dynamic processes, offering invaluable insights into their underlying mechanisms.

Example 0.3. The stochastic differential equations considered here are Markov processes,

$$\mathbf{X}(t) = \mathbf{X}_0 + \int_0^t \mathbf{f}(\mathbf{X}(s), s; \theta) ds + \int_0^t g(\mathbf{X}(s);\theta) d\mathbf{W}(s) \tag{0.3.17}$$

where it is assumed that $\mathbf{f}, \mathbf{g} \in \mathcal{G}^{n \times m}$. Since $\mathbf{X}(t)$ is a function of $t > 0$ then the Wiener process $\mathbf{W}$ is independent of $\mathbf{X}$ for $t < 0$. As $t > 0$ increases and $\mathbf{X}$ evolves it is independent of $t < 0$ given that we know $X(0)$. For a proof of this claim, see [3]. Returning to the bound on g given by the

$$\langle g(\mathbf{X}, t), \mathbf{X} \rangle \leq \alpha + \beta |\mathbf{X}|^2$$

with Frobenius inner product (Appendix A). This can lead to bounded solutions, where $\mathbf{X}$ remains within a certain region of the state space over time. This implies that the system's behavior is confined and doesn't exhibit unbounded growth or runaway effects (or explosions).

0.3.5 Hidden Markov Models

Hidden Markov Models are an extension of Markov Models. HMMs encapsulate an interplay between observable outputs and unobservable states, thus facilitating nuanced analysis and prediction of temporal dynamics.

Let $X(t)$ denote the hidden state at time t, and $Y(t)$ represent the output of the state $X(t)$. Suppose one can observe $Y(t)$ then Y can be represented as a function of X such that $Y = H(X(t))$ where H is some unknown mapping we wish to infer some information about.

In dynamical systems modeling we may wish to express a process in the framework of HMM's.

Given a dynamical model $f : \mathbb{R}^m \to \mathbb{R}^m$, and a measurement model $H : \mathbb{R}^m \to \mathbb{R}^p$ as follows

$$\mathbf{X}(t+1) = f(\mathbf{X}(t)) + \mathbf{g}(t) \tag{0.3.18}$$

$$\mathbf{Y}(t) = H(\mathbf{X}(t)) + \mathbf{r}(t) \tag{0.3.19}$$

where $\mathbf{g}, \mathbf{r}$ are random variables representing stochasticity in the model. This set up constitutes a hidden Markov model (HMM): a sequence of (possibly) hidden states linked together by a dynamical system f, such as $\mathbf{X}(t+1)$, which are only observed through an observation operator, such as H. The HMM is a useful framework for building inference techniques for dynamical systems. In many data assimilation techniques, the function f is only available as a binary executable, where we do not know the internal structure of the dynamical equation which produces the outputs that we observe. Moreover, this invites the notion of treating the dynamical model f as a black box - furthermore the dynamics of a forecast equation (or law) are typically a crude discretisations (3rd or 4th order Runge-Kutta approximations or Finite Element Methods discussed later) of an underlying time-continuous physical system.

0.3.6 The Fokker-Planck Equation

The Fokker-Planck equation, also known by the Kolmogorov Forward equation describes the evolution of a probability distribution for a stochastic process. If we take the bounded function g we know it is bounded above by

$$\langle g(\mathbf{X}, t), \mathbf{X} \rangle \leq \alpha + \beta |\mathbf{X}|^2$$

with Frobenius inner product, consider the case when $\beta > 0$ then a positive β term suggests a dissipative effect. The product $g \cdot \mathbf{X}$ (or the inner product in higher dimensions) is always less than or equal to a term that includes $\beta|\mathbf{X}|^2$ (or βx^2 in 1D). This implies that some of the system's energy (represented by $|\mathbf{X}|^2$) is being dissipated or lost over time. This could be due to factors like friction, damping, or other mechanisms that counteract the influence of g.

0.3.6.1 Infinitesimal Generators

Remark 0.3. The solution operator Ψ is a (semi)group of linear or nonlinear operators. This tells us that the solution operator Ψ behaves like a collection of linear or linear-like operators that can be composed together.

Definition 0.5. Let $\Psi(\cdot;t)$ be a solution operator that is a (semi)group and continuous in t over a normed vector (Appendix A) space in $\mathbb{R}^n$, then for a function $\psi : \mathbb{R}^n \to \mathbb{R}$ then the infinitesimal for the function $\psi(\mathbf{X})$ is

$$\lim_{h\to 0}\left\{\frac{\Psi(\psi;h)-\Psi(\psi;0)}{h}\right\} = \mathcal{L}\psi$$

To apply this definition to our process to find a generator for the Ito process we have, consider the process $dY(t) = \psi(\mathbf{X},t)$ and using the Taylor expression expand as follows

Since a stochastic differential equation in the Itô sense defines a Markov process, we define a *diffusivity matrix* $\Gamma \in M_{d\times d}(\mathbb{R})$ by

$$\Gamma = g(\mathbf{X},t)g^{\mathsf{T}}(\mathbf{X},t)$$

which encodes how the dissipation β effects how the function is controlled due to random fluctuations in the dynamics of the stochastic component g. Essentially this characterises the spread of a probability density over time. This notation shall be used when explicit statements of the stochastic part of the Ito process is not needed during calculations.

$$\begin{aligned}
d\mathbf{Y}(t) &= \psi(\mathbf{X}(t),t) - \psi(\mathbf{X}(0),0)\\
&=\sum_{i=0}^{n}\frac{\partial\psi}{\partial x_i}dX_i + \frac{1}{2}\sum_{i=0}^{n}\sum_{j=0}^{n}\frac{\partial^2\psi}{\partial x_i\partial x_j}dX_i dX_j\\
&=\sum_{i=0}^{n}\frac{\partial\psi}{\partial x_i}(f_i(X_i,t)dt + g_i(X_i,t)dW_i)\\
&\quad+\frac{1}{2}\sum_{i=0}^{n}\sum_{j=0}^{n}\frac{\partial^2\psi}{\partial x_i\partial x_j}\left(f_i(X_i,t)dt + g_i(X_i,t)dW_i\right)\left(f_j(X_j,t)dt + g_j(X_j,t)dW_j\right)\\
&=\sum_{i=0}^{n}\frac{\partial\psi}{\partial x_i}f_i dt + \sum_{i=0}^{n}\frac{\partial\psi}{\partial x_i}g_i dW_i + \frac{1}{2}\sum_{i=0}^{n}\sum_{j=0}^{n}\frac{\partial^2\psi}{\partial x_i\partial x_j}f_i f_j (dt)^2\\
&\quad+\frac{1}{2}\sum_{i=0}^{n}\sum_{j=0}^{n}\frac{\partial^2\psi}{\partial x_i\partial x_j}(f_i g_j + f_j g_i)dt\,(dW_i + dW_j) + \frac{1}{2}\sum_{i=0}^{n}\sum_{j=0}^{n}\frac{\partial^2\psi}{\partial x_i\partial x_j}g_i g_j dW_i dW_j\\
&=\sum_{i=0}^{n}\frac{\partial\psi}{\partial x_i}f_i dt + \frac{1}{2}\sum_{i=0}^{n}\sum_{j=0}^{n}\frac{\partial^2\psi}{\partial x_i\partial x_j}g_i g_j dt + \sum_{i=0}^{n}\frac{\partial\psi}{\partial x_i}g_i dW_i
\end{aligned}$$

Now considering $\Psi(\psi(\mathbf{X},t)) - \Psi(\psi(0,t))$

$$
\begin{aligned}
\Psi(\psi(\mathbf{X},t)) - \Psi(\psi(0,t)) =& \int_t^{t+h} \left\{ \sum_{i=0}^{n} \frac{\partial \psi}{\partial x_i} f_i + \frac{1}{2} \sum_{i=0}^{n} \sum_{j=0}^{n} \frac{\partial^2 \psi}{\partial x_i \partial x_j} g_i g_j \right\} ds + \int_t^{t+h} \sum_{i=0}^{n} \frac{\partial \psi}{\partial x_i} g_i dW_i ds \\
=& \int_t^{t+h} \left\{ \sum_{i=0}^{n} \frac{\partial \psi}{\partial x_i} f_i + \frac{1}{2} \sum_{i=0}^{n} \sum_{j=0}^{n} \frac{\partial^2 \psi}{\partial x_i \partial x_j} g_i g_j \right\} ds + 0 \\
\frac{\Psi(\psi(\mathbf{X},t)) - \Psi(\psi(0,t))}{h} =& \frac{1}{h} \int_t^{t+h} \left\{ \sum_{i=0}^{n} \frac{\partial \psi}{\partial x_i} f_i + \frac{1}{2} \sum_{i=0}^{n} \sum_{j=0}^{n} \frac{\partial^2 \psi}{\partial x_i \partial x_j} \Gamma_{ij} \right\} ds \\
\lim_{h \to 0} \left\{ \frac{\Psi(\psi(\mathbf{X},t)) - \Psi(\psi(0,t))}{h} \right\} =& \lim_{h \to 0} \left(\frac{1}{h} \int_t^{t+h} \left\{ \sum_{i=0}^{n} \frac{\partial \psi}{\partial x_i} f_i + \frac{1}{2} \sum_{i=0}^{n} \sum_{j=0}^{n} \frac{\partial^2 \psi}{\partial x_i \partial x_j} \Gamma_{ij} \right\} ds \right) \\
& \text{by the Fundemental Theorem of Calculus} \\
=& \lim_{h \to 0} \left\{ \frac{1}{h} \left[\psi(X_i(t+h), t+h) - \psi(X_i(h), h) \right] \right\} \\
=& \frac{\partial \psi}{\partial t} \\
=& \sum_{i=0}^{n} \frac{\partial \psi}{\partial x_i} f_i + \frac{1}{2} \sum_{i=0}^{n} \sum_{j=0}^{n} \frac{\partial^2 \psi}{\partial x_i \partial x_j} \Gamma_{ij} \\
=& \mathbf{f} \cdot \nabla \psi + \frac{1}{2} \mathbf{g}\mathbf{g}^{\mathsf{T}} \cdot \nabla^2 \psi \\
=& \mathbf{f}(\mathbf{X},t) \cdot \nabla \psi + \frac{1}{2} \Gamma \cdot \nabla \nabla \psi \\
=& \mathcal{L} \psi
\end{aligned}
$$

We are dealing with a solution operator Ψ acting on a function ψ. Each function ψ is a function of $\mathbf{X}$ which evolves over time t. We assert that Ψ is continuous over t within the vector space $\mathbb{R}^n$ it operates in. If one imagines a rocky and hilly landscape being generated in time which represents the behaviour of an Ito process. The height $\mathbf{X}$ and time t determine how likely the process is to move in one direction or another. We have a function ψ which describe some property of this path, like its height or average curvature. The solution operator Ψ acts on this function. As before it will take some initial value of the function (does not have to be zero) and tells you how that property ψ will evolve over some small time interval h and then in the limiting case, at a fixed point t. The solution operator helps us understand how properties of the process evolve, but the infinitesimal generator $\mathcal{L}$ gives us a rule to calculate the rate of change of the evolution directly, most importantly without knowledge of the solution operator itself.

Definition 0.6. The Differential Operator (Infinitesimal Generator for the Ito Markov Process) $\mathcal{L}$ is given as

$$
\begin{aligned}
\mathcal{L} &= \mathbf{f} \cdot \nabla + \frac{1}{2} \mathbf{g}\mathbf{g}^{\mathsf{T}} \cdot \nabla^2 \\
&= \mathbf{f} \cdot \nabla + \frac{1}{2} \Gamma \cdot \nabla^2
\end{aligned}
\tag{0.3.20}
$$

We get this using Itô's lemma, which was originally obtained from Taylor expansion. The diffusion term, characterized by the diffusivity matrix $\Gamma = gg^{\mathsf{T}}$, encodes the spread of the probability density due to random fluctuations in the process. These fluctuations are driven by the stochastic component

represented by the function g. Higher values in the diffusivity matrix indicate a larger spread due to stronger random effects.
For an input ψ then,

- $\mathbf{f} \cdot \nabla\psi$ captures the change in $\mathbf{f}$ which describes an average tendency to move in one direction or another.
- The diffusion term $\frac{1}{2}\mathbf{g}\mathbf{g}^{\mathsf{T}} \cdot \nabla^2\psi$ captures the change due to random fluctuations with the second order derivative representing the spreading effect on the probability density function.

$\mathcal{L}$ serves as a linearisation of the underlying Markov transition mapping (defined by a matrix), which governs state transitions. This linearisation becomes particularly advantageous when coupled with a Taylor series expansion around a specific time point, typically t $=$ 0. The first-order term of this expansion, a linear operator, approximates the instantaneous rate of change in the probability distribution of the process at a specific state. This effectively captures the dominant effect on the process's evolution over small time intervals. The elegance of $\mathcal{L}$ lies in its ability to transform the non-linear transition probabilities coming from $\mathbf{g}\mathbf{g}^{\mathsf{T}}d\mathbf{W}$ into a tractable linear framework, facilitating the analysis of the process's short-term dynamics. While the linearisation provides an excellent initial approximation, it's crucial to acknowledge its limitations. For longer time horizons or processes exhibiting more intricate behavior, incorporating higher-order terms from the Taylor series expansion might be necessary for a more accurate representation, it is also worth noting that the operator $\mathcal{L}$ does not include the Wiener process either.

While the solution operator Ψ directly provides the solution at different time points, the infinitesimal generator doesn't. However, understanding the instantaneous rate of change (captured by $\mathcal{L}$) allows one to piece together the function's evolution over time. By repeatedly applying the concept of the instantaneous rate of change using the infinitesimal generator, you can build an approximation of the solution operator's behavior for finite time intervals without knowing Ψ.

The generator $\mathcal{L}$ plays a crucial role in computing the rate of change of functions ψ of solutions to a stochastic differential equation via the Itô formula for stochastic integration.

0.3.6.2 Deriving the Fokker-Planck Equation

Problem Setting

Suppose one drops a particle into a liquid at some position, say $\mathbf{x}(0)$ at time $t = 0$. The subsequent evolution of the particles position is noisy due to the constant bombardment of particles from the liquid hitting the particle making the trajectory jagged and resemble that of a Wiener process - or commonly known as a Brownian motion. There is no way to be sure about where the particle will be at some other time. Then the best way to talk about it is in terms of probabilities.

We would like to give a measure of our uncertainty. What probability distribution tells us the most likely position the particle will be - what is the most likely state of it over time and how does it change with respect to time?

We don't care about any particular path the particle takes at some time t. Instead we ask the question of the probability distribution that the particle at $\mathbf{x}(t)$ for some t disregarding the previous history of positions $\mathbf{x}\left(t_{j-1}\right), \mathbf{x}\left(t_{j-2}\right), \ldots$ to answer this question we turn towards the Fokker-Planck equation or the Kolmogorov Forward equation.

The problem setting can be generalised to that of the stochastic population model. Let's clarify what is meant by this, recall that the stochastic population dynamics model that was presented earlier is

$$\frac{dN}{dt} = (v(t) + \eta)\, N(t)$$

where η is distributed with some probability distribution function $\rho(t)$. Not only are these phenomena observed in physics but also situations where we can express unexplained variations in the dynamics of a system over time. This makes this framework general and mathematically rich.

Modelling the Evolution of the Probability Density

Since a stochastic differential equations in the form that we consider are Markovian we have a state transition function which has continuous probability density ρ.

Let $\mathbf{X}$ be a stochastic process and $\psi : \mathbb{R}^n \times [0, t] \to \mathbb{R}$ be a function in $C^{1,2}(\mathbb{R}^n \times [0, t], \mathbb{R})$ which denotes the space of functions that are once continuously differentiable with respect to the spatial variables $\mathbf{X}$ and twice continuously differentiable with respect to the temporal variable t. Then the process $\psi(\mathbf{X}, t)$ where $0 \le t$ with transition probability density function $\rho : \mathbb{R}^n \times [0, t] \to \mathbb{R}$ has the expected value given by

$$\mathbb{E}\left[\psi(\mathbf{X}, t)\right] = \int_{\mathbb{R}} \psi(\mathbf{x}(s), s)\rho(\mathbf{x}(s), s) ds \tag{0.3.21}$$

In terms of the problem setting the function $\psi(\mathbf{X}, t)$ represents some observable or quantity of interest associated with the stochastic process $\mathbf{X}$ at time t. This could be, for example, the position of the particle at time t, or some function of its position. The integral calculates this expected value by integrating the product of ψ, which is the value of the quantity of interest at a particular position $\mathbf{x}$, and ρ, which is the probability density function describing the likelihood of finding the particle at position $\mathbf{x}$ at time t, over all possible positions $x_0, x_1, ..., x_n$. In essence, the expected value provides a way to summarize the behavior of the stochastic process in terms of its average or typical behavior, which is essential for understanding the overall dynamics of the system and predicting its future behaviour.

In summary the expected value, calculated by integrating the product of the position function ψ and the probability density ρ, gives us an average picture of where the particle is likely to be at a particular time. However the Brownian motion described in terms of the Wiener process is a dynamic process - it's a dynamical system and with dynamical systems the position or state of an object as time increases, changes due to random fluctuations. We are not just interested in the average position of a single time, but as alluded to previously, how the average position itself evolves over time. In this case, by taking the derivative of the expected value (which depends on time t), we're essentially asking: at what rate is the average position of the particle changing with respect to time? Therefore we give

$$\frac{d\mathbb{E}}{dt} = \int_{\mathbb{R}} \psi(\mathbf{x}(s), s) \frac{\partial \rho}{\partial s} ds \tag{0.3.22}$$

However by the Itô Formula (0.2.41), we express ψ in terms of the drift $\mathbf{f}$ and diffusion $\mathbf{g}$ coefficients, leading to a differential expression involving these terms.

$$\begin{aligned}
d\psi &= \frac{\partial\psi}{\partial t}dt + \frac{\partial\psi}{\partial\mathbf{x}}dX + \frac{1}{2}\frac{\partial^2\psi}{\partial\mathbf{x}^2}(dX)^2 \\
&= \frac{\partial\psi}{\partial t}dt + \frac{\partial\psi}{\partial\mathbf{x}}(\mathbf{f}(\mathbf{X},t)dt + \mathbf{g}(\mathbf{X},t)d\mathbf{W}(t)) + \frac{1}{2}\left(\frac{\partial^2\psi}{\partial\mathbf{x}^2}(\mathbf{f}(\mathbf{X},t)dt + \mathbf{g}(\mathbf{X},t)d\mathbf{W}(t))^2\right) \\
&= \frac{\partial\psi}{\partial t}dt + \frac{\partial\psi}{\partial\mathbf{x}}\mathbf{f}(\mathbf{X},t)dt + \frac{\partial\psi}{\partial\mathbf{x}}\mathbf{g}(\mathbf{X},t)d\mathbf{W}(t) \\
&+ \frac{1}{2}\left(\frac{\partial^2\psi}{\partial\mathbf{x}^2}(\mathbf{f}(\mathbf{X},t))^2dt^2 + 2\mathbf{f}(\mathbf{X},t)\mathbf{g}(\mathbf{X},t)\frac{\partial^2\psi}{\partial\mathbf{x}^2}d\mathbf{W}(t)dt + \frac{\partial^2\psi}{\partial\mathbf{x}^2}(\mathbf{g}(\mathbf{X},t))^2d\mathbf{W}(t)^2\right) \\
&= \frac{\partial\psi}{\partial t}dt + \frac{\partial\psi}{\partial\mathbf{x}}\mathbf{f}(\mathbf{X},t)dt + \frac{\partial\psi}{\partial\mathbf{x}}\mathbf{g}(\mathbf{X},t)d\mathbf{W}(t) + \frac{\partial^2\psi}{\partial\mathbf{x}^2}(\mathbf{g}(\mathbf{X},t))^2d\mathbf{W}(t)^2 \\
&= \frac{\partial\psi}{\partial t}dt + \frac{\partial\psi}{\partial\mathbf{x}}\mathbf{f}(\mathbf{X},t)dt + \frac{\partial\psi}{\partial\mathbf{x}}\mathbf{g}(\mathbf{X},t)d\mathbf{W}(t) + \frac{\partial^2\psi}{\partial\mathbf{x}^2}(\mathbf{g}(\mathbf{X},t))^2dt \\
&= \left(\frac{\partial\psi}{\partial t} + \frac{\partial\psi}{\partial\mathbf{x}}\mathbf{f}(\mathbf{X},t) + \frac{1}{2}(\mathbf{g}(\mathbf{X},t))^2\frac{\partial^2\psi}{\partial\mathbf{x}^2}\right) + \frac{\partial\psi}{\partial\mathbf{x}}\mathbf{g}(\mathbf{X},t)d\mathbf{W}(t) \qquad \text{by (0.2.34-36)} \\
&= \frac{\partial\psi}{\partial t} + \mathbf{f}\cdot\nabla\psi + \frac{1}{2}\mathbf{g}\mathbf{g}^{\mathsf{T}}\cdot\nabla^2\psi + \frac{\partial\psi}{\partial\mathbf{x}}\mathbf{g}(\mathbf{X},t)d\mathbf{W}(t) \\
&= \frac{\partial\psi}{\partial t} + \mathbf{f}\cdot\nabla\psi + \frac{1}{2}\Gamma\cdot\nabla^2\psi + \nabla\psi\cdot\mathbf{g}d\mathbf{W} \\
&= \frac{\partial\psi}{\partial t} + \mathcal{L}\psi + \nabla\psi\cdot\mathbf{g}d\mathbf{W}
\end{aligned}$$

Notice that we have an extra term in the derivative of ψ and this is because we have assumed that ψ is smooth however $\mathbf{W}$, if one does notice in the derivation of $\mathcal{L}$ that $\nabla\psi\cdot\mathbf{g}d\mathbf{W}$ did appear but in the formulation of the differential operator of Ψ the term $\nabla\psi\cdot\mathbf{g}d\mathbf{W}$ was integrated out. As can be seen when the expected value is taken with respect to t, the expected value of ψ has the same form as $\partial_t\psi$ and this is no coincidence since we are averaging.

$$\begin{aligned}
\psi(\mathbf{X},t) &= \psi(\mathbf{X}(0),0) + \int_0^t\left(\frac{\partial\psi}{\partial s} + \mathbf{f}\cdot\nabla\psi + \frac{1}{2}\Gamma\cdot\nabla^2\psi\right)ds + \int_0^t\nabla\psi\cdot\mathbf{g}d\mathbf{W}(s) \\
&= \psi(0) + \int_0^t\frac{\partial\psi}{\partial s} + \mathcal{L}\psi(\mathbf{X}(s),s)ds + \int_0^t\nabla\psi\cdot\mathbf{g}d\mathbf{W}(s)
\end{aligned}$$

assuming for $x_i \to \pm\infty$ then $\psi, \rho, \partial_{\mathbf{x}}\psi, \partial_t\rho \to 0$ then by Dynkin's Formula (Appendix A probability theory)

$$\begin{aligned}
\mathbb{E}[\psi(\mathbf{X},t)] &= \mathbb{E}\left[\psi(\mathbf{X}(0),0)\right] + \int_{\mathbb{R}^n}\int_0^t\left(\frac{\partial\psi}{\partial s} + \mathcal{L}\psi(\mathbf{X}(s),s)\right)\rho(\mathbf{x},s)dsd\mathbf{x} + \int_{\mathbb{R}^n}\int_0^t(\nabla\psi\cdot\mathbf{g})\frac{\partial\rho}{\partial s}d\mathbf{W}(s)d\mathbf{x} \\
&= \mathbb{E}\left[\psi(\mathbf{X}(0),0)\right] + \int_{\mathbb{R}^n}\int_0^t\left(\frac{\partial\psi}{\partial s} + \mathcal{L}\psi(\mathbf{X}(s),s)\right)\rho(\mathbf{x},s)dsd\mathbf{x} + 0
\end{aligned}$$

where that $\mathbb{E}[d\mathbf{W}(t)] = 0$ $(0.2.34-36)$ and by noting that

$$\begin{aligned}
\mathbb{E}[\psi(\mathbf{X},t)] &= \int_{\mathbb{R}^n}\int_0^t\left(\frac{\partial\psi}{\partial s} + \mathbf{f}\cdot\nabla\psi + \frac{1}{2}\mathbf{g}\mathbf{g}^{\mathsf{T}}\cdot\nabla^2\psi\right)\rho(\mathbf{x},s)dsd\mathbf{x} \\
&= \int_{\mathbb{R}^n}\int_0^t\frac{\partial\psi}{\partial s}\rho(\mathbf{x},t)dsd\mathbf{x} + \int_{\mathbb{R}^n}\int_0^t\left(\mathbf{f}\cdot\nabla\psi\right)\rho(\mathbf{x},t)dsd\mathbf{x} + \int_{\mathbb{R}^n}\int_0^t\frac{1}{2}\left(\mathbf{g}\mathbf{g}^{\mathsf{T}}\cdot\nabla^2\psi\right)\rho(\mathbf{x},t)dsd\mathbf{x} \\
&= \int_{\mathbb{R}^n}\left(\int_0^t\frac{\partial\psi}{\partial s}\rho(\mathbf{x},t)ds + \int_0^t\left(\mathbf{f}\cdot\nabla\psi\right)\rho(\mathbf{x},t)ds + \int_0^t\frac{1}{2}\left(\mathbf{g}\mathbf{g}^{\mathsf{T}}\cdot\nabla^2\psi\right)\rho(\mathbf{x},t)ds\right)d\mathbf{x}
\end{aligned}$$

We can deal with these integrals separately. Starting from left to right we get the following expressions for each. We can use integration by parts on the integral

$$\int_0^t \frac{\partial \psi}{\partial s} \rho(\mathbf{x}, t) ds$$

by letting

$$\begin{array}{cc} u = \rho(\mathbf{x}, t) & dv = \frac{\partial \psi}{\partial t} \\ du = \frac{\partial \rho}{\partial t} & v = \psi(\mathbf{x}, t) \end{array}$$

then

$$\begin{aligned} \int_0^t \frac{\partial \psi}{\partial s} \rho(\mathbf{x}, t) ds &= [\rho(\mathbf{x}, s)\psi(\mathbf{X}, s)]_0^t - \int_0^t \psi(\mathbf{X}, s) \frac{\partial \rho}{\partial t} ds \\ &= - \int_0^t \psi(\mathbf{X}, s) \frac{\partial \rho}{\partial t} ds \\ &= - \frac{d\mathbb{E}}{dt} \end{aligned}$$

and so we are left with

$$\begin{aligned} \int_{\mathbb{R}^n} \int_0^t \frac{\partial \psi}{\partial s} \rho(\mathbf{x}, t) ds d\mathbf{x} &= - \int_{\mathbb{R}^n} \int_0^t \psi(\mathbf{X}, s) \frac{\partial \rho}{\partial t} ds d\mathbf{x} \\ &= -\mathbb{E}\left[\psi(\mathbf{X}, t)\right] \end{aligned}$$

In this case it is assumed that the terms go to zero on the boundary of the domain due to the function ψ [5] these are known as far-field boundary conditions and these are applied to systems being analysed on unbounded domain like this one, namely the whole of the real numbers, namely that $\rho \to 0$ as $||\mathbf{X}(t)|| \to \infty$ for all $t \in [0, \infty)$. The drift $\mathbf{f}$ and diffusion term $\mathbf{g}$ cannot become discontinuous since they were assumed to be smooth functions in $\mathcal{G}^{m \times n}$ so we have $\psi = 0$ at $t = 0$ and $\psi = 0$ at t. It is only possible for them to become discontinuous across the boundary of the domain and in this case is unbounded so is always continuous. Now for the second integral

$$\int_{\mathbb{R}^n} \int_0^t (\mathbf{f} \cdot \nabla \psi)\, \rho(\mathbf{x}, t) ds d\mathbf{x}$$

By the Fubini-Tonneli theorem (Appendix A)

$$\begin{aligned} \int_{\mathbb{R}^n} \int_0^t (\mathbf{f} \cdot \nabla \psi)\, \rho(\mathbf{x}, t) ds d\mathbf{x} &= \int_0^t \int_{\mathbb{R}^n} (\mathbf{f} \cdot \nabla \psi)\, \rho(\mathbf{x}, t) d\mathbf{x} ds \\ &= \int_0^t \int_{\mathbb{R}^n} (\mathbf{f}\rho \cdot \nabla \psi)\, d\mathbf{x} ds \end{aligned}$$

Using integration by parts with

$$\begin{array}{cc} u = \rho\mathbf{f} & dv = \nabla\psi \\ du = \nabla(\rho\mathbf{f}) & v = \psi \end{array}$$

so we have

$$\begin{aligned}\int_{\mathbb{R}^n} (\mathbf{f}\rho \cdot \nabla\psi)\, d\mathbf{x} &= [\rho\mathbf{f}\psi]_{\mathbb{R}} - \int_{\mathbb{R}^n} \nabla(\rho\mathbf{f})\psi d\mathbf{x} \\ &= -\int_{\mathbb{R}^n} \nabla(\rho\mathbf{f})\psi d\mathbf{x}\end{aligned}$$

due to far field boundary conditions. So we are left with

$$\int_{\mathbb{R}^n}\int_0^t (\mathbf{f}\cdot\nabla\psi)\,\rho(\mathbf{x},t)dsd\mathbf{x} = -\int_{\mathbb{R}^n}\int_0^t \nabla(\rho\mathbf{f})\psi dsd\mathbf{x}$$

Finally for the integral

$$\int_{\mathbb{R}^n}\int_0^t \frac{1}{2}\left(\mathbf{g}\mathbf{g}^\intercal\cdot\nabla^2\psi\right)\rho(\mathbf{x},t)dsd\mathbf{x}$$

This integral shall require a double application of integration by parts since the term $\nabla^2\psi$ needs to become ψ only. By Fubini-Tonelli theorem we have

$$\int_{\mathbb{R}^n}\int_0^t \frac{1}{2}\left(\mathbf{g}\mathbf{g}^\intercal\cdot\nabla^2\psi\right)\rho(\mathbf{x},t)dsd\mathbf{x} = \frac{1}{2}\int_0^t\int_{\mathbb{R}^n}\left(\mathbf{g}\mathbf{g}^\intercal\cdot\nabla^2\psi\right)\rho(\mathbf{x},t)d\mathbf{x}ds$$

and deal with the integral

$$\int_{\mathbb{R}^n}\left(\mathbf{g}\mathbf{g}^\intercal\cdot\nabla^2\psi\right)\rho(\mathbf{x},t)d\mathbf{x} = \int_{\mathbb{R}^n}\left(\rho\mathbf{g}\mathbf{g}^\intercal\cdot\nabla^2\psi\right)d\mathbf{x}$$

The first application of integration by parts for this integral will require the substitutions

$$\begin{aligned} u_1 &= \rho\mathbf{g}\mathbf{g}^\intercal & dv_1 &= \nabla^2\psi \\ du_1 &= \nabla\left(\rho\mathbf{g}\mathbf{g}^\intercal\right) & v_1 &= \nabla(\psi)\end{aligned}$$

and so we have the following integrals to apply the second round of the by parts formula to

$$\begin{aligned}\int_{\mathbb{R}^n}\left(\mathbf{g}\mathbf{g}^\intercal\cdot\nabla^2\psi\right)\rho(\mathbf{x},t)d\mathbf{x} &= \left[\rho\mathbf{g}\mathbf{g}^\intercal\cdot\nabla\psi\right]_{\mathbb{R}} - \int_{\mathbb{R}^n}\nabla^2\cdot\left(\rho\mathbf{g}\mathbf{g}^\intercal\right)\cdot\nabla\psi d\mathbf{x} \\ &= -\int_{\mathbb{R}^n}\nabla^2\cdot\left(\rho\mathbf{g}\mathbf{g}^\intercal\right)\nabla\psi d\mathbf{x} \\ &= -\left[\nabla\cdot\left(\rho\mathbf{g}\mathbf{g}^\intercal\right)\psi\right]_{\mathbb{R}} + \int_{\mathbb{R}^n}\nabla^2\cdot\left(\rho\mathbf{g}\mathbf{g}^\intercal\right)\psi d\mathbf{x} \\ &= \int_{\mathbb{R}^n}\nabla^2\cdot\left(\rho\mathbf{g}\mathbf{g}^\intercal\right)\psi d\mathbf{x}\end{aligned}$$

with the substitutions

$$\begin{aligned} u_2 &= \nabla\cdot\left(\rho\mathbf{g}\mathbf{g}^\intercal\right) & dv_2 &= \nabla\psi \\ du_2 &= \nabla^2\cdot\left(\rho\mathbf{g}\mathbf{g}^\intercal\right) & v_2 &= \psi\end{aligned}$$

and so the final integral can be expressed as

$$\frac{1}{2}\int_{\mathbb{R}^n}\int_0^t\left(\mathbf{g}\mathbf{g}^\intercal\cdot\nabla^2\psi\right)\rho(\mathbf{x},t)dsd\mathbf{x} = \int_0^t\int_{\mathbb{R}^n}\nabla^2\cdot\left(\rho\mathbf{g}\mathbf{g}^\intercal\right)\psi d\mathbf{x}ds$$

Now bringing this all together we have

$$\begin{aligned}
\mathbb{E}[\psi(\mathbf{X},t)] &= \int_{\mathbb{R}^n}\int_0^t \left(\frac{\partial\psi}{\partial s} + \mathbf{f}\cdot\nabla\psi + \frac{1}{2}\mathbf{g}\mathbf{g}^{\mathsf{T}}\cdot\nabla^2\psi\right)\rho(\mathbf{x},s)dsd\mathbf{x} \\
&= \int_{\mathbb{R}^n}\int_0^t \frac{\partial\psi}{\partial s}\rho(\mathbf{x},t)dsd\mathbf{x} + \int_{\mathbb{R}^n}\int_0^t \left(\mathbf{f}\cdot\nabla\psi\right)\rho(\mathbf{x},t)dsd\mathbf{x} + \int_{\mathbb{R}^n}\int_0^t \frac{1}{2}\left(\mathbf{g}\mathbf{g}^{\mathsf{T}}\cdot\nabla^2\psi\right)\rho(\mathbf{x},t)dsd\mathbf{x} \\
&= -\int_{\mathbb{R}^n}\int_0^t \psi(\mathbf{X},s)\frac{\partial\rho}{\partial s}dsd\mathbf{x} - \int_{\mathbb{R}^n}\int_0^t \nabla\cdot(\rho\mathbf{f})\psi dsd\mathbf{x} + \frac{1}{2}\int_0^t\int_{\mathbb{R}^n} \nabla^2\cdot\left(\rho\mathbf{g}\mathbf{g}^{\mathsf{T}}\right)\psi dsd\mathbf{x} \\
&= \int_{\mathbb{R}^n}\int_0^t \left(\frac{1}{2}\nabla^2\cdot\left(\rho\mathbf{g}\mathbf{g}^{\mathsf{T}}\right)\psi(\mathbf{X},s) - \nabla\cdot(\rho\mathbf{f})\psi(\mathbf{X},s) - \psi(\mathbf{X},s)\frac{\partial\rho}{\partial s}\right)dsd\mathbf{x} \\
&= \int_{\mathbb{R}^n}\int_0^t \left(\frac{1}{2}\nabla^2\cdot\left(\rho\mathbf{g}\mathbf{g}^{\mathsf{T}}\right)\psi(\mathbf{X},s) - \nabla\cdot(\rho\mathbf{f})\psi(\mathbf{X},s) - \psi(\mathbf{X},s)\frac{\partial\rho}{\partial s}\right)dsd\mathbf{x} \\
&= \int_{\mathbb{R}^n}\int_0^t \psi(\mathbf{X},s)\left(\frac{1}{2}\nabla^2\cdot\left(\rho\mathbf{g}\mathbf{g}^{\mathsf{T}}\right) - \nabla\cdot(\rho\mathbf{f}) - \frac{\partial\rho}{\partial s}\right)dsd\mathbf{x} \\
&= \int_{\mathbb{R}^n}\int_0^t \psi(\mathbf{X},s)\left(\frac{\partial\rho}{\partial s} + \nabla\cdot(\rho\mathbf{f}) - \frac{1}{2}\nabla^2\cdot(\mathbf{g}\mathbf{g}^{\mathsf{T}}\rho)\right)dsd\mathbf{x}
\end{aligned}$$

Due to the far field boundary conditions then this integral should be zero since $\psi \to 0$ and so does $\mathbb{E}[\psi] \to 0$ for $t \notin (0,\infty)$

$$\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{f}) - \frac{1}{2}\nabla^2\cdot(\mathbf{g}\mathbf{g}^{\mathsf{T}}\rho) = 0$$

where we have now arrived at the Fokker-Planck partial differential equation:

$$\begin{aligned}
\frac{\partial\rho}{\partial t} &= -\nabla\cdot(\rho\mathbf{f}) + \frac{1}{2}\nabla^2\cdot(\mathbf{g}\mathbf{g}^{\mathsf{T}}\rho) \\
&= -\nabla\cdot(\rho\mathbf{f}) + \frac{1}{2}\nabla\cdot(\mathbf{g}\mathbf{g}^{\mathsf{T}}\cdot\nabla\rho)
\end{aligned} \tag{0.3.23}$$

in a more explicit formulation we have

$$\frac{\partial\rho}{\partial t}(\mathbf{x},t) = -\sum_{k=0}^{n}\frac{\partial(\rho(\mathbf{x},t)f_k(\mathbf{x},t))}{\partial x_k} + \frac{1}{2}\sum_{k=0}^{n}\sum_{l=0}^{n}\frac{\partial^2}{\partial x_k\partial x_l}\left\{\left(\sum_{j=0}^{n} g_{kj}(\mathbf{x},t)g_{lj}(\mathbf{x},t)\right)\rho(\mathbf{x},t)\right\}$$

Relating back to the particles movement, we know that this particles movement is described by a stochastic differential equation. One is interested in some property of this particles path, this property is represented by ψ (or in particular with the Fokker Planck equation ρ) which in this case describes the probability distribution of being in a particle state at time t - like it's position at t. The state of the particle is anywhere in the vector space $\mathbb{R}^n$. There is a probability density function ρ, the expected value gives an average estimate of where the particle may be at t and gives us information on the overall dynamics of the random movement. The issue lies in the crucial point highlighted above, though, we are not interested in the specific time t, but rather, all time t - we can still make use of the expected value we calculated - instead of focusing on the single expected value we can consider how fast the average value is changing. Ito's formula helps express the probability density function ρ in terms of the SDE's drift $\mathbf{f}$ and diffusion $\mathbf{g}$ coefficients. By taking the derivative of the expected value and using Ito's formula, we arrive at a differential equation. There's a hidden term in this equation that arises due to the randomness of the particle's movement $\nabla\psi\cdot\mathbf{g}d\mathbf{W}(s)$, the term disappears when we take the expected value, but it plays a crucial role in the full dynamics. By applying Dynkin's formula (Appendix A), we can eliminate this term and arrive at a new equation: the Fokker-Planck equation. This equation relates the average behavior to the underlying probability density function, however, by averaging out the random fluctuations, it loses information about the individual random paths the particle might take. In the next chapter when the Kushner-Stratonovich and Zakai equations are discussed we shall see cases where the randomness is explicitly included when calculating the probability density evolution.

The Fokker-Planck equation provides a comprehensive framework for understanding the evolution of stochastic processes, capturing the behavior of the variable $\mathbf{X}$ once its solution is obtained. The Fokker-Planck equation is an Itô diffusion process since it describes the evolution of transition functions probability density of the Markov process generated by the Itô equation, this is why the infinitesimal generator has been considered. However, solving this equation becomes challenging in high-dimensional scenarios. Subsequently, it will be illustrated that Fourier Neural Operators can serve as surrogate models to learn the Fokker-Planck equation, aiding in data assimilation. Rather than relying on Markov Chain Monte Carlo methods for sampling posterior distributions, the Fokker-Planck equation can be learned as a functional mapping through a neural operator, enabling a direct inference of the posterior distribution.

0.3.6.3 Generators or Fokker-Planck Equation?

It is worth noting why we went through all of those calculations when we could have simply used the definition of the infinitesimal generator in definition 5 to obtain the rate of change of the probability distribution with respect to time. Consider the Fokker-Planck equation

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho \mathbf{f}) + \frac{1}{2}\nabla^2 \cdot (\Gamma \rho)$$

and if we compare this to

$$\mathcal{L}\rho = \mathbf{f} \cdot \nabla \rho + \frac{1}{2}\Gamma \cdot \nabla^2 \rho$$

If we first focus on the Fokker-Planck equation. Suppose we have an initial condition at time u for the position $\mathbf{X}(u)$ this equation predicts how the probability of finding the object in different positions changes over time, that is, for any later time $t \geq u$. We need not in general have $u = 0$. Therefore, as we vary $\mathbf{X}$ in this sense, given initial condition $\rho(\mathbf{X}(u), u)$, then we can evolve the probability density for $t > u$.

In the case where the function put into the generator $\mathcal{L}$ is the probability distribution ρ , then we get the expression for $\mathcal{L}\rho$. In this scenario, we have the position at time t and we ask the question of how it got there. That is we try to reconstruct the trajectory based off of the information of the object at t for all previous time $u \leq t$ representing the change in probability distribution for the position $\mathbf{X}$ with respect to time - giving the most likely state for $\mathbf{X}$, $\mathcal{L}$ operates on the state variable $\mathbf{X}$ at t which enables iteration over t. For the general case here we would be working from the boundary values on the interval $[u, t]$. The boundary t gives the final observation of $\mathbf{X}$ that we know of.

We can schematically represent this with arrows, the arrow $\nearrow$ will represent the forward or increase in time and $\swarrow$ representing decrease in time respectively.

$$\begin{aligned} \text{Fokker-Planck (Forward) Equation:} \quad & u \nearrow t, \quad u \leq t, \quad t \in [0, \infty) - \text{Prediction} \\ \text{Generator (Backward) Equation:} \quad & t \swarrow u, \quad u \leq t, \quad t \in [0, \infty) - \text{Reconstruction} \end{aligned}$$

0.3.6.4 One Dimensional Example of Fokker-Planck and Generator

The generator, Fokker-Planck equation and their action on the probability distribution become more clear in one dimension. We can analyse the behaviour of the probability distribution in this way to learn what these equations can tell us. Given that we have a process not in $\mathbb{R}^n$ but rather $\mathbb{R}$ and then the equations are given as

$$\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial x}\left[f(X(t),t)\rho(X(t),t)\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[g^2(X(t),t)\rho(X(t),t)\right]$$

For the initial condition $\rho(X(u),u)$ where $t \geq u$. The quantity $\partial_x(f\rho)$ represents the rate of change of the expectation of the deterministic part of Ito process it is acting upon. The term $\partial_x^2[g^2\rho]$ gives the change in the spread of the probability distribution (variance) over time, both with respect to the spatial variable X. Focusing on the minus sign, this is analogous to the direction of movement of the probability distribution if the drift f is positive, indicating that the process tends to move towards higher values of X, then the probability density tends to decrease in those regions, hence the negative sign. Essentially The negative sign in front of $\partial_x(f\rho)$ ensures that the decrease in density due to the drift is balanced by the increase in density in opposing regions (positive direction vs negative direction). The negative sign ensures that regions where the drift pushes probability density are effectively "depleted" while regions opposing the drift are "filled up," maintaining the conservation of probability mass, when integrated, should be 1.

0.3.6.5 Numerical Example: Stochastic Lorenz System

We provide a numerical example of the Fokker-Planck equation with the Stochastic Lorenz 1963 Model (SLM). It is a simple model - but it will highlight a very important point about what happens to the probability density function as the system evolves over time. The SLM [6] is given by three coupled differential equations:

$$\begin{aligned}
\frac{dx_1}{dt} &= \sigma(x_2 - x_1) + d_1\frac{dW_1(t)}{dt}\\
\frac{dx_2}{dt} &= x_1(\rho - x_3) - x_2 + d_2\frac{dW_2(t)}{dt}\\
\frac{dx_3}{dt} &= x_1x_2 - \beta x_3 + d_3\frac{dW_3(t)}{dt}
\end{aligned}$$

where $d_1, d_2, d_3 \in \mathbb{R}$ and are a scaling factor to describe the intensity of the noise. How much the noise will influence the dynamics of the system. $\mathbf{W} = (W_1, W_2, W_3)$ which is a Wiener Process. To apply the Fokker-Planck equation to the Stochastic Lorenz System we define two functions $\mathbf{f}, \mathbf{g} \in \mathbb{R}^3$ and $\mathbf{X} = (x_1(t), x_2(t), x_3(t)) = (x_1, x_2, x_3) \in \mathbb{R}^3$ which are vector valued functions

Note that this system describes an Ito process in $\mathbb{R}^3$ which is expressed in the standard form:

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

for ease of notation let

$$\begin{pmatrix}\xi_1\\ \xi_2\\ \xi_3\end{pmatrix} = \frac{d\mathbf{W}}{dt}$$

so that we have

$$\mathbf{f}(\mathbf{X},t) = \begin{pmatrix}\sigma(x_2 - x_1)\\ x_1(\rho - x_3) - x_2\\ x_1x_2 - \beta x_3\end{pmatrix}, \quad \mathbf{g}(\mathbf{X},t) = \begin{pmatrix}\xi_1\\ \xi_2\\ \xi_3\end{pmatrix}, \quad \mathbf{d} = \begin{pmatrix}d_1\\ d_2\\ d_3\end{pmatrix}$$

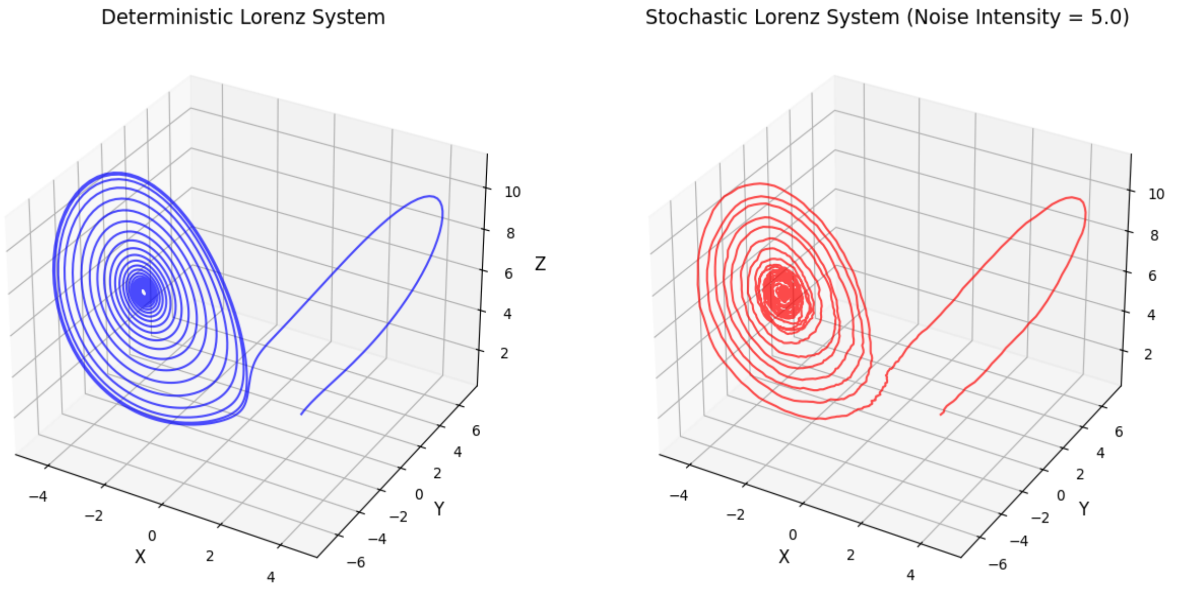
and define the diffusivity matrix $\Gamma = \mathbf{g}\mathbf{g}^\intercal$ such that

$$\Gamma = (\xi_1, \xi_2, \xi_3) \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix}$$
$$= \begin{pmatrix} \xi_1^2 & \xi_1\xi_2 & \xi_1\xi_3 \\ \xi_2\xi_1 & \xi_2^2 & \xi_2\xi_3 \\ \xi_3\xi_1 & \xi_3\xi_2 & \xi_3^2 \end{pmatrix}$$

a numerical example of the Lorenz system with stochastic forcing can be seen in the figure below. In comparison to the Deterministic version where $\mathbf{d} = \mathbf{0}$ and $\frac{d\mathbf{W}}{dt} = \mathbf{0}$ the dynamics are much simpler. If the Fokker-Planck equation was described for the deterministic version one would arrive at the Advection-Diffusion equation where $\frac{1}{2}\nabla^2(\Gamma \cdot \rho) = 0$ leaving the partial differential equation

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho \mathbf{f})$$

which can be solved analytically with appropriate boundary conditions and suitable initial conditions.


Deterministic Lorenz System

Stochastic Lorenz System (Noise Intensity = 5.0)

Now applying the Fokker-Planck equation in $\mathbb{R}^3$ with far field boundary conditions since the Lorenz system still forms a strange attractor even in the presence of noise. Therefore, the region of attraction in the Lorenz system will be an inherent constraint on the density function thus explicit boundary conditions for the Fokker Planck equation in the simulation are not needed and can extend to infinity this is for a few reasons.

1. Non-vanishing at the Edges: Due to the constant diffusion, the PDF won't be zero at the edges of the attractor. It will have very low values but not strictly zero. This reflects the possibility of the system reaching those regions through diffusion, although the probability might be tiny

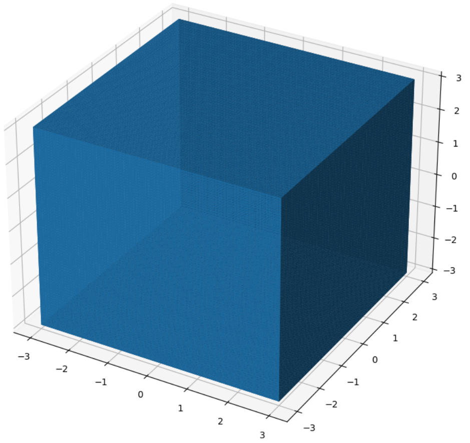
2. Normalization: The integral of the PDF over the entire space (technically, negative to positive infinity for x, y, and z) should be equal to 1. This ensures the total probability of finding the system somewhere is 100%

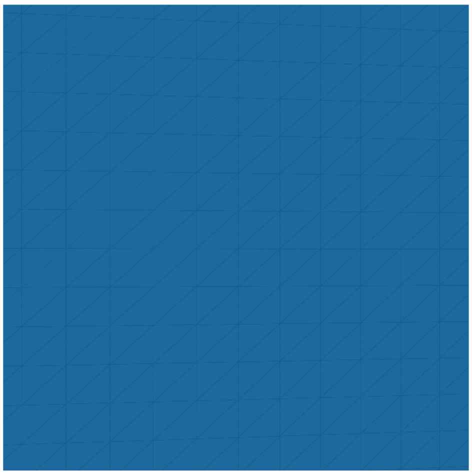
3. If a finite region of interest is needed (i.e. zooming into the attractor), one can consider reflecting or absorbing boundary conditions at the edges of that region. These would modify the FPE to account for the boundaries being artificial.

We will show that Lorenz system is continuous in $\mathbb{R}^3$. The initial assumption on ρ is that it is normally distributed with mean μ and variance α, that is at $t = 0$ as an initial starting point then $\rho(\mathbf{X}(0), 0) \sim \mathcal{N}(\mu, \alpha)$ then and we want to see how the probability density function evolves over time.

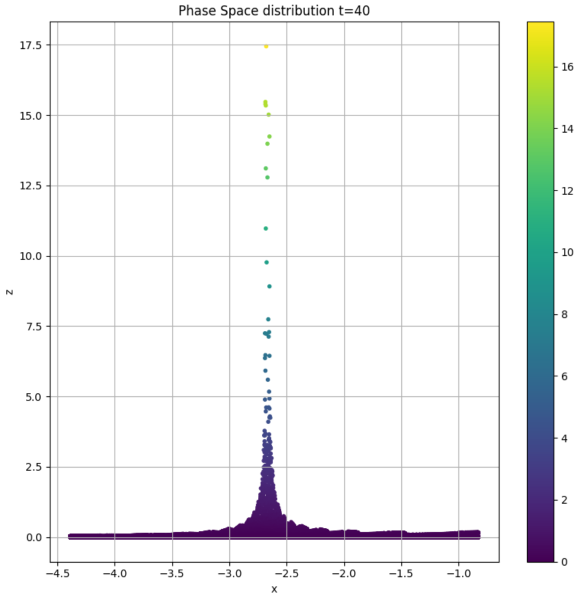
$$\begin{aligned}\frac{\partial \rho}{\partial t} &= -\nabla \cdot (\rho \mathbf{f}) + \frac{1}{2}\nabla^2 \cdot (\Gamma \rho) \\ &= -\nabla \cdot \left(\rho \left(\begin{array}{c} \sigma(x_2 - x_1) \\ x_1(\rho - x_3) - x_2 \\ x_1 x_2 - \beta x_3 \end{array} \right) \right) + \frac{1}{2} \nabla \cdot \nabla \cdot \left(\begin{pmatrix} \xi_1^2 & \xi_1\xi_2 & \xi_1\xi_3 \\ \xi_2\xi_1 & \xi_2^2 & \xi_2\xi_3 \\ \xi_3\xi_1 & \xi_3\xi_2 & \xi_3^2 \end{pmatrix} \rho \right) \\ &= -\sum_{k=0}^{3} \frac{\partial(\rho f_k)}{\partial x_k} + \frac{1}{2} \sum_{k=0}^{3}\sum_{l=0}^{3} \frac{\partial^2}{\partial x_k \partial x_l} \left\{ \left(\sum_{j=0}^{3} g_{kj} g_{lj} \right) \rho \right\} \qquad (0.3.24)\end{aligned}$$

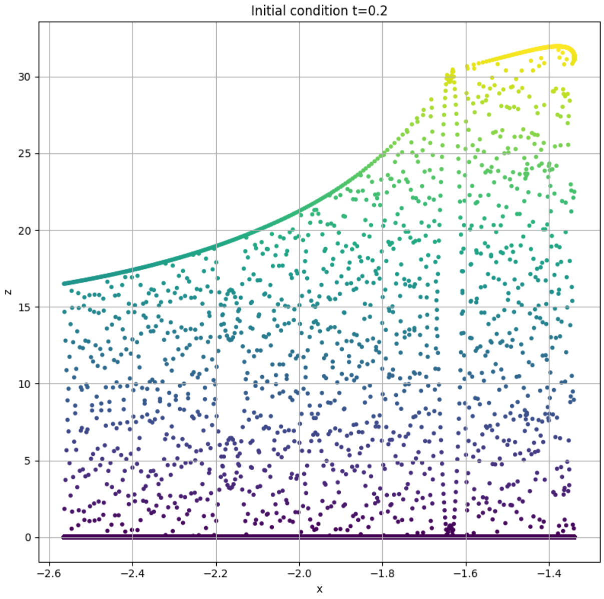
The Fokker-Planck equation is best solved numerically in this scenario since no closed form solution exists to this partial differential equation. We provide a simulation of the evolution of the probability distribution for the lorenz system with a control matrix Γ_c for reproducibility, this matrix can readily be generalized. The simulation is code, written in Python, uses the FEniCS library [7, 8 ,9, 10] with a C++ back-end, DOLFIN [9], on four Intel i5 processors, with python bindings for the Message Passing Interface (MPI). MPI distributes the computational domain Ω across each processor, each computing different parts of Ω to aid compuation and it then recombined. Ω for this problem is a cubic spatial grid in Cartesian coordinates. The box $\Omega \subset \mathbb{R}^3$ is defined such that each edge is on the interval $[-3, 3]$ these boundaries are chosen specifically such that the region of attraction is captured in the solution of the FPE. Ω is divided into 80 smaller cubes called cells of length 0.0375 and then split in half to form 160 tetrahedral elements with length 1.875×10^{-2} units each Tetrahedral element is where computation in done. Ω is called the *mesh* or *computational domain*. The set Ω produced in this set up is depicted below:

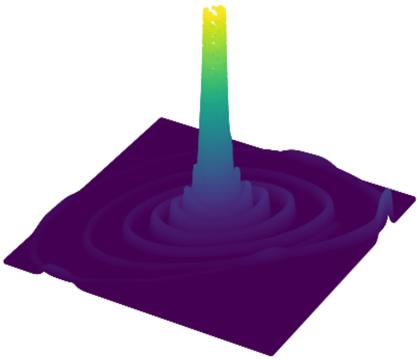


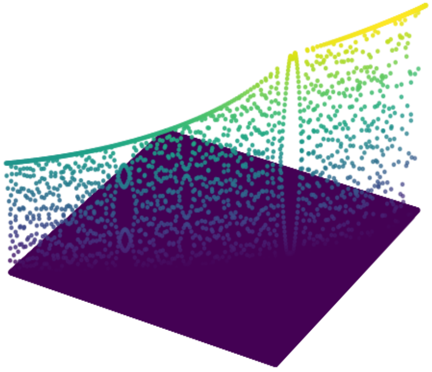


The Lorenz system parameters are given as $\sigma = 3.765432$, $\rho = 17.654323456$ and $\beta = 8/3$ which scales the attractor down to a feasible computational domain. The initial conditions are given by $\dot{x}_0 = -3.345$, $\dot{y}_0 = -3.2$ and $\dot{z}_0 = 1$. The stochastic Lorenz system is given with noise sampled from standard Gaussian distribution $\xi_n \sim \mathcal{N}(0,1)$ and scaled with $d_n = 5$ for $n \in \{1,2,3\}$. The system is then evolved for 40 time steps $[0,20] \subset \mathbb{R}$ and we notice there a single attracting set for this attractor at $(x,y,z) = (-2.6,-2.6,z)$ for all $z \in \mathbb{R}$. The initial condition for the FPE is given by the phase space distribution Φ of points of the Lorenz attractor, that is, the distribution of points in the phase space. For example, the distribution will peak around the attracting set as $t \to \infty$. The initial condition is given for $t = 0.2$ which we denote Φ_0 which is to represent a short observation period of the dynamics.









Over the entire time domain $t \in T$ where $T = [0, 20]$ of the system the Fokker-Planck equation may converge to a *stationary distribution*, that it the probability density does not change after a time $t' > \max\{T\}$ called a *steady state* distribution of the system. This is illustrated for the stochastic Lorenz example - these distribution can also be expensive to solve for and often analytical solutions are highly complex or not possible to describe with current mathematical techniques. The steady state distribution for the Lorenz system with our parameters is illustrated below - by advancing past iteration $\max\{T\}$ up to a final time step $t' > 20$. Systems with attractors, fixed points, and limit cycles, tend to exhibit stationary distributions around these points. The probability distribution concentrates around the attractor's region/basin of attraction, leading to a steady-state behavior. The FPE captures how $-\nabla \cdot (\rho\mathbf{f})$ and $\frac{1}{2}\nabla^2 \cdot (\Gamma\rho)$ interact. Initially, the probability density might be far from equilibrium. The drift term will try to pull it towards a specific state, while the diffusion term will cause it to spread out. The diffusion term tends to counteract the drift, gradually filling in any gaps created by the drift. Eventually, an equilibrium is reached where the net change in the probability density becomes zero. This is the steady-state distribution. This phenomena occurs when we have a bounded domain with far-field boundary conditions. There are Fokker-Planck equations that violate these conditions. For example, consider a system with a constant positive drift $-\nabla \cdot (\rho\mathbf{f}) = c \geq 0$ towards positive infinity. The probability density will keep spreading without ever reaching an equilibrium. This simplified example uses constant diffusivity matrix $\Gamma = \Gamma_c = \mathbf{I}_3 \in \mathbb{M}_{3\times3}(\mathbb{N})$ such that

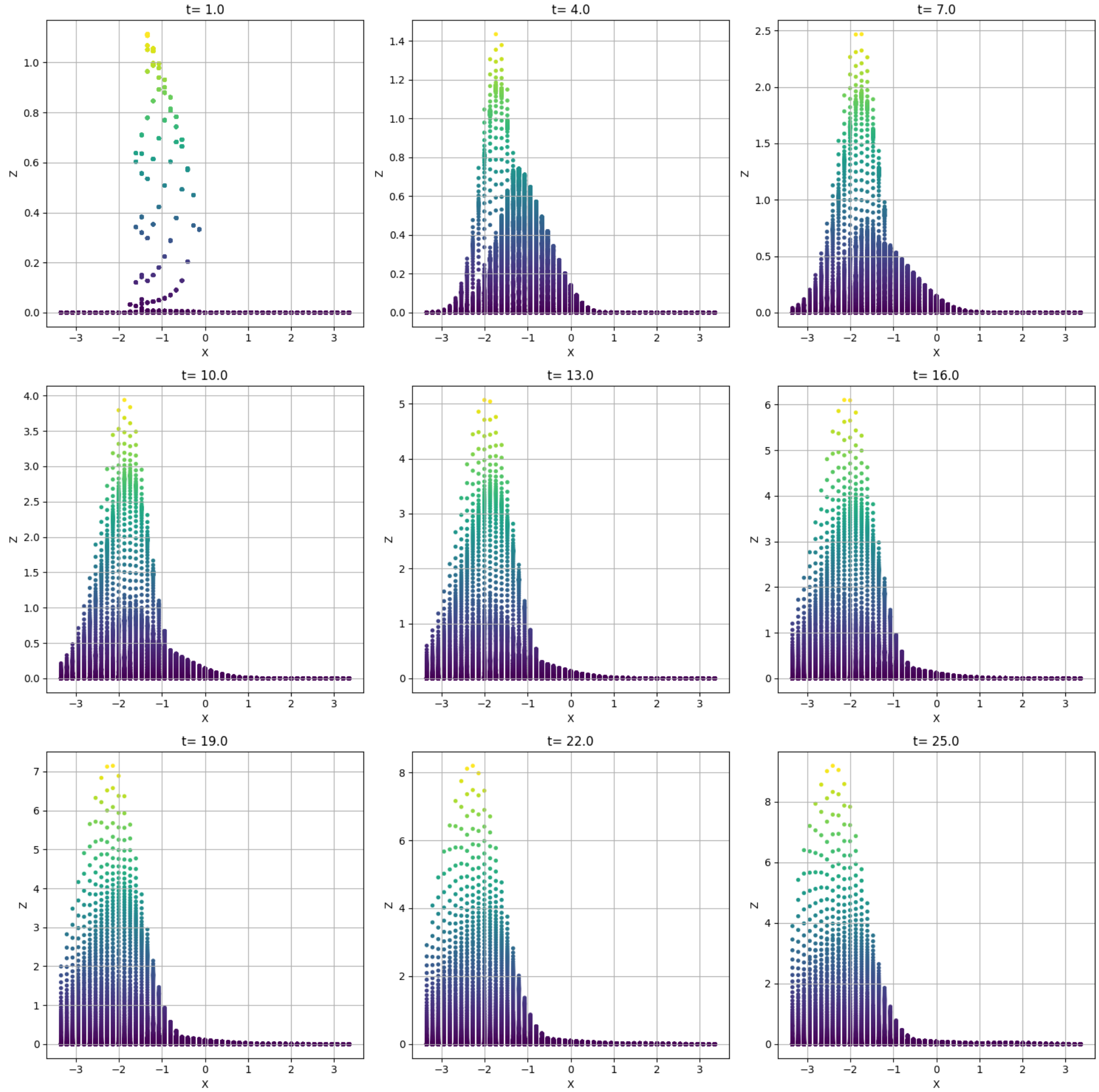
$$\begin{aligned}
\frac{\partial\rho}{\partial t} &= -\nabla \cdot \left(\rho \begin{pmatrix} \sigma(x_2 - x_1) \\ x_1(\rho - x_3) - x_2 \\ x_1x_2 - \beta x_3 \end{pmatrix}\right) + \frac{1}{2}\nabla \cdot \nabla \cdot \left(\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\rho\right), \quad \rho_0 = \Phi_0 \\
\mathbf{f}(\mathbf{X}, 0) &= \mathbf{f}((x_0, y_0, z_0), 0) \\
\rho \notin \Omega &= 0 \\
\Omega &= \{(x, y, z) : x \leq x_0, y \leq y_0; x, y, z \in \mathbb{R}\}
\end{aligned}$$

However, many Fokker-Planck equations used in physics, chemistry, and engineering do satisfy these conditions and have well-defined steady-state distributions. These are particularly useful for studying systems that eventually reach an equilibrium state. Huang, Ji, Liu and Yi [11] studied the steady state behaviour of Fokker-Planck Equations and the existence of steady state distribution - Mathematically the steady state distributions occur when there exists a solution $\rho(\mathbf{X}(t), t)$ such that

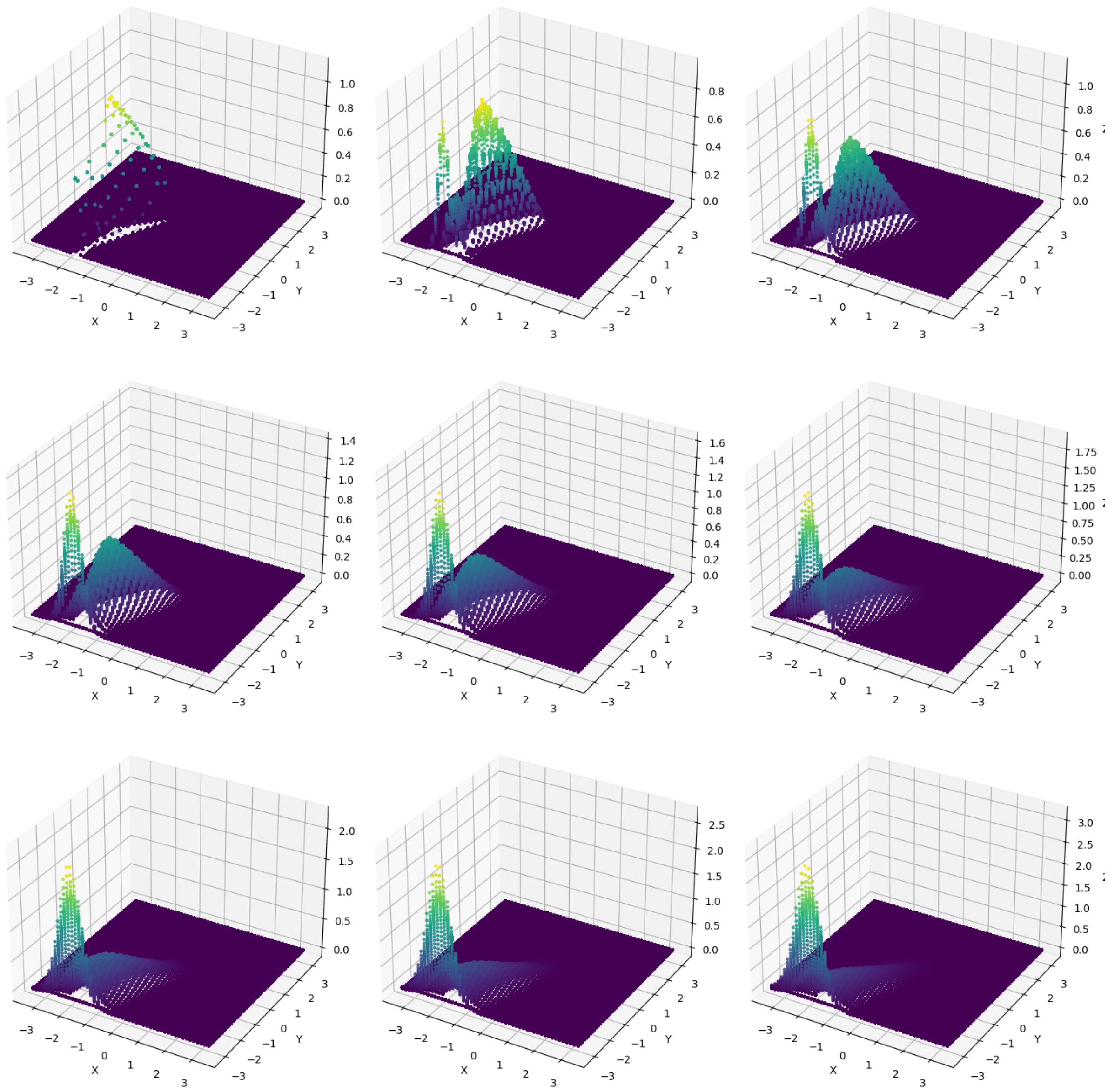
$$-\nabla \cdot (\rho \mathbf{f}) + \frac{1}{2}\nabla^2 \cdot (\Gamma \rho) = 0, \quad \rho \geq 0$$

$$\mathbb{P} = \int_0^t \rho(\mathbf{X}(t), t) d\mathbf{x} = 1$$

where $\mathbb{P}$ is the probability measure given by the cumulative distribution - of observing the particle in the steady state at t. For $\mathbf{f}, \mathbf{g} \in \mathcal{G}^{m \times n}$. Below the stationary distribution for the Stochastic Lorenz System can be seen.



where one can see that the peak of the distribution $\rho \to \Phi$. Below is a 3D representation on the interval $t = 0, ..., 9$ to show a more fine grained transition from the initial condition.


X
Y
−3
−2
−1
0
1
2
3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
0.5
1.5
2.0
2.5
3.0

Part II

Bayesian Data Assimilation

Term	Focus	Technique	Examples
Data Fusion	Combining Data	Raw Data	Sensor Fusion Financial Analysis
Data Assimilation	Type of data fusion	Raw Data Dynamical Model	Weather forecasting Oceanography Neurosciences Agent based modelling Crowd/Population dynamics
State Estimation	Current State of Systems	Data Fusion Data Assimilation Statistical	Navigation Health Monitoring
Information Fusion	Extracting information	Data Fusion Data Assimilation Machine Learning Data Analysis Logic	Social Network Anlaysis Military Intelligence Medical Diagnostics

0.4 Introduction and Reason for Data Assimilation

Data assimilation is the process of integrating model forecasts with observational data. It encompasses a set of techniques tailored for sequential, statistical inference on dynamical model variables such as position or speed, as well as static variables governing a system's evolution over time. This process is built upon the broader framework of State Estimation in control theory.

The overarching term for such mathematical frameworks is Data Fusion, which involves amalgamating multiple information sources to offer a more precise and comprehensive understanding of a given situation. Data fusion operates across various data types, formats (such as images, videos, social network data, and phone data), and datasets exhibiting different modalities (distinct statistical properties within a dataset). Its applications are diverse, ranging from combining radar and camera data for autonomous vehicles to merging financial data from disparate sources for risk assessment.

State Estimation, meanwhile, refers to leveraging available information to ascertain the current condition or characteristics of a system. It may draw upon data from diverse origins or blend data with models to derive estimates. Examples include determining the position and velocity of an aircraft using radar data or monitoring a machine's health through sensor readings. Notably, the steering of the Apollo Spacecraft utilized state estimation via the Kalman Filter algorithm, estimating variables like location, velocity, and acceleration based on accelerometer, gyroscope, and sextant measurements.

Data assimilation, in contrast, represents a specific subtype of data fusion, predominantly employed in scientific and engineering realms for modeling and forecasting purposes. It integrates observational data with mathematical models to refine the depiction of a system's state.

These concepts—Data Fusion, State Estimation, and Data Assimilation—are encapsulated within Information Fusion, the ultimate stage focusing on information extraction. This phase extends beyond raw data, involving reasoning, analysis, interpretation, and potentially employing Machine Learning techniques. Examples encompass military intelligence operations, where data from various sources converge to evaluate enemy capabilities, as well as medical diagnosis procedures, which consider patient history alongside lab tests and imaging results. Table 1 provides a concise summary of this overview.
In contrast to simpler examples of data assimilation above, such as those found above are unlike hy-

drology, meteorology, quantum mechanics, and neuroscience, more complex applications often involve physical models with numerous parameters, sometimes in the billions. These models frequently exhibit nonlinear dynamics. For instance, in hydrological applications like History Matching for oil production planning, there are a few known variables such as pressure, saturation, and humidity, but many unknown parameters in the model θ is then a large set, resulting in a large parameter space. These models often involve equations resembling Darcy Flow, which is a nonlinear partial differential equation describing thermodynamic behavior.

Similar complexities arise in meteorology when dealing with weather forecasting, where models are based on equations like Navier-Stokes equations, thermodynamic principles, and ideal gas laws. Observations in meteorology come from various sources and formats and modalities the ECMWF and Met Office models require super computers for data processing and fusion - the actual data assimilation takes much shorter time. Other applications of data assimilation are atmospheric chemistry, air quality analysis, forest fire prediction, over crowding and crushing prediction in closed indoor spaces, and climatology.

Returning to the example of population dynamics, particularly in epidemiological analysis like COVID-19 models, state variables represent the number of susceptible, infected, and recovered individuals, affecting the population size at a given time. These are often coupled nonlinear differential equations, which may incorporate spatial distribution through methods like agent-based modeling or diffusion processes. Observations in epidemiology can originate from diverse sources such as governments, health organizations, Google Flu Trends, social media analytics, data mining applications, local phone tower transmissions, and GPS analytics.

In these complex models, equations may be either linear or nonlinear. Moreover, the nature of the dynamical system can extend to abstract equation types, including neural networks, graph-based models, or other machine learning-based models, which are often nonlinear dynamical systems such as the case with neural network based surrogate models involving functions such as $\tanh(x)$ to specifically introduce nonlinearity in the learning processes.

As we focus on data assimilation here we are combining observation data with scientific information. Data assimilation serves as a powerful technique for bridging the divide between theoretical models and real-world observations in various scientific disciplines. Central to this approach is the recognition that mathematical models, while offering valuable insights into system behavior, inherently possess limitations. These limitations can arise from simplifications made during model construction or from uncertainties associated with initial conditions. Data assimilation addresses these shortcomings by incorporating observational data into the modeling framework.

The Bayesian statistical framework provides a robust foundation for data assimilation. This framework allows us to explicitly account for uncertainties inherent in both the observations (e.g., measurement errors) and the model (e.g., limitations in representing complex processes). Through Bayesian inference, data assimilation leverages observations to refine the model state estimate, resulting in a more accurate representation of the true system state.

What is this process doing exactly? The data assimilation process typically follows a cyclical approach. Initially, a mathematical model is employed, incorporating prior knowledge about the system. This model state serves as a preliminary estimate. Subsequently, real-world observations are introduced. By comparing the model predictions with these observations, discrepancies can be quantified via the Bayesian Framework. Data assimilation techniques then utilize this information to adjust the model state estimate, effectively merging theoretical understanding with empirical data. This refined state estimate offers a more realistic representation of the system's current state. Notably, this improved state estimate can then be employed for various purposes, including generating more accurate forecasts for future system behavior.

The recursive nature of data assimilation is a key strength. As new observations become available, the entire process can be repeated. With each iteration, the model incorporates the latest data, leading to a progressively more accurate understanding of the system dynamics. This continuous improvement process makes data assimilation particularly valuable in fields like weather forecasting and climate modeling, where having the most up-to-date picture of the system is paramount.

It is crucial to acknowledge that models employed in data assimilation are not perfect representations of reality. A fundamental assumption is that the core principles embedded within the model, often rooted in established physical laws, accurately describe the system's governing processes. Discrepancies typically arise from imperfections in the initial conditions provided to the model. Therefore, a central challenge in data assimilation lies in identifying the most accurate initial conditions that, when incorporated into the model, enable it to accurately represent the system dynamics up to the latest observation point and generate reliable forecasts for the future.

It's important to acknowledge that models employed in data assimilation are not flawless representations of reality. Their limitations can stem from inherent complexities of the system being modeled or simplifications made for computational efficiency. However, even with imperfect models, data assimilation offers valuable insights. By comparing model predictions with observations, systematic errors within the model can be identified. Furthermore, data assimilation can still improve forecasts by incorporating real-world data, even if the underlying model isn't perfect.

0.5 Mathematical Formulation of Bayesian Data Assimilation

In this section we formulate data assimilation in continuous time with a bayesian formulation. The bayesian formulation of data assimilation provides a clearly defined mathematical problem whose solution provide the best-in-class set of solutions. While these solutions are are the ideal solutions to such problems, in practice many simplifications must be made due to numerical stability and computational complexity. The underlying theoretical dynamical systems in continuous time behave like a stochastic differential equations as do the observations of the dynamical systems.

0.5.1 More on Hidden Markov Models

As in [3], suppose the state and observations $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^l$ respectively and generated over time t by a dynamical system $\mathbf{f} \in \mathcal{G}^{d\times d}$ and some observation operator $H : \mathbb{R}^d \to \mathbb{R}^d$ as follows

$$\mathbf{x}(t+1) = \mathbf{f}(\mathbf{x}(t), t) + \xi_t, \quad t = 0, 1, ... \tag{0.5.1}$$

$$\mathbf{y}(t) = H\left(\mathbf{x}(t), t\right) + \eta_t, \; t = 1, 2, ... \tag{0.5.2}$$

where ξ_t and η_t are Wiener processes representing measurement noise and some initial condition is specified by the phase space distribution

$$\mathbf{x}_0 \sim \rho_0 \tag{0.5.3}$$

where $\rho_0 : \mathbb{R}^d \to \mathbb{R}$. The model parameters $\mathbf{f}, H$ and Φ_0 are all assumed known. If $\rho_0 = \mathcal{N}(\mu,\mathbf{Q})$ where $\mathbf{Q} \in \mathbb{M}_{d\times d}(\mathbb{R})$ then μ and $\mathbf{Q}$ are also to be assumed, known.

The process $\{(\mathbf{x}(t), \mathbf{y}(t)); t \in \mathbb{N}\}$ constitutes a Hidden Markov Model (HMM), a sequence of hidden and non-hidden states linked together by $\mathbf{f}$, which is only observed through H. We denote the conditioned probability density function $\rho(\mathbf{y}|\mathbf{x})$. The basic properties of HMM's which can be derived from $6.1 - 6.3$ are independence relations

$$\rho(\mathbf{x}(t+1) : \mathbf{x}_{0:T}) = \rho(\mathbf{x}(t+1)|\mathbf{x}(t)) \tag{0.5.4}$$

$$\rho(\mathbf{y}(t+1) : \mathbf{x}_{0:T}) = \rho(y(t+1)|\mathbf{x}(t)) \tag{0.5.5}$$

for all time indices $t \leq T$, where the colon indicates a sequence

$$\mathbf{x}_{0:T} = \{\mathbf{x}_t; t = 0, ..., T\}$$

0.5.2 Bayesian Data Assimilation

Data assimilation is now formulated as a statistical inference problem on a Hidden Markov Model. The idea is to compute the density function $\rho(\mathbf{x}|\mathbf{y})$, given by Bayes formula

$$\rho(\mathbf{x}|\mathbf{y}) \propto \rho(\mathbf{y}|\mathbf{x})\rho(\mathbf{x})$$

where $\mathbf{y}$ is any data available and $\mathbf{x}$ is unknown - we wish to estimate $\mathbf{x}$. The density $\rho(\mathbf{x}|\mathbf{y})$ is known as the *posterior* distribution resulting from multiplication of the *likelihood* $\rho(\mathbf{y}|\mathbf{x})$ which is our knowledge from the observations $\mathbf{y}$ and the *prior* $\rho(\mathbf{x})$. The objective of data assimilation is to compute $\rho(\mathbf{x}_{0:t}|\mathbf{y}_{1:T})$. Splitting into three cases we have

1. Smoothing if $t < T$
2. Filtering if $t = T$
3. Forecasting if $t > T$

The joint posterior distribution is given by the relation

$$\begin{aligned}\rho(\mathbf{x}_{0:T}|\mathbf{y}_{1:T}) &\propto \rho(\mathbf{x}_0)\prod_{t=1}^{T}\rho(\mathbf{x}(t)|\mathbf{x}(t-1))\rho(\mathbf{y}(t)|\mathbf{x}(t)) \\ &= \rho(\mathbf{y}(t)|\mathbf{x}(t))\rho(\mathbf{x}(t)|\mathbf{x}(t-1))\rho(\mathbf{x}_{0:T}|\mathbf{y}_{1:t-1})\end{aligned}$$

which allows for inference by recursively appending $\rho(\mathbf{y}(t)|\mathbf{x}(t))\rho(\mathbf{x}(t)|\mathbf{x}(t-1))$ to the previous posterior.

0.5.3 Filtering - Brief History

This section is intended as a brief outline of nonlinear filtering theory. The study has it's origin in the work of R.L. Stratonovich. The earlier work on nonlinear filtering in continuous time was largely problem-driven and experimental.

However the history of the discrete-time filtering problem traces back to the pioneering work of Andrey Kolmogorov and Mark Krein in the late 1930s and early 1940s. Kolmogorov's contributions in 1939 and 1941, as referenced in [12], laid some foundational concepts for filtering theory. Krein further advanced this field in 1945, as documented in [13].

Simultaneously, Norbert Wiener made significant strides in the continuous filtering problem, as noted in [14]. Wiener's work, which delved into optimal estimation in dynamical systems midst noise, originated in 1942, initially for applications in defense which was classified throughout World War II, and later declassified and published as a book in 1949.

However, it was not until the 1960s that Rudolf Kalman developed the Kalman filter, as mentioned in [15] and [16]. The Kalman filter revolutionized estimation theory by providing a systematic approach to estimate the true values of measurements. It accomplishes this by predicting a value based on the system's dynamics, estimating the uncertainty associated with this prediction, and then computing a weighted average of this prediction and the measured value. Importantly, the Kalman filter assigns more weight to the value with the least uncertainty, thus optimizing the estimation process. This is what was used in the Apollo missions - which does assume Gaussian noise. The continuous time version - known as the Kalman-Bucy filter assumed a Brownian motion or Wiener processes.

Classical filtering theory involves the manipulation of signals through linear operations such as convolution, Fourier transforms, and linear differential equations. Such as telecommunications, control systems, image processing, and audio processing.

The theory of nonlinear filtering was developed in the late 1960s and early 1970s. We shall focus on *$reference$ probability* methods. This method relies on the transformation of the filtering problem under continuous changes of the underlying probability measure. [see appendix A]. The reference probability approach was developed in work first by G. Kallianpur and C. Striebel [17].

Nonlinear systems exhibit behaviors that cannot be represented by linear models. Such as systems governed by nonlinear differential equations. Such as target tracking, robotics, navigation systems, financial modeling, and biological systems modeling.

0.5.4 Nonlinear Filtering for Diffusion Processes

We are interested in multivariate Itô Processes over $\mathbb{R}^n$. The signal is taken in differential form

$$d\mathbf{X}(t) = \mathbf{f}(\mathbf{X}(t), t)dt + \mathbf{g}(\mathbf{X}(t), t)d\mathbf{W}(t), \quad X_i \sim \rho$$

where $i \leq n$, $\mathbf{f}, \mathbf{g} \in \mathcal{G}^{n \times m}$ with $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ and $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^{n \times m}$ further assume that $\mathbf{f}, \mathbf{g}$ are $C^{1,2}$ functions with bounded first and second order derivatives. $\mathbf{W}$ is an m-dimensional Wiener Process and the expected values of X_i for any order $p \leq n$ is finite, that is $\mathbb{E}[|\mathbf{X}|^p] < \infty$ which is an L_p space (Appendix A). This setup defines a Markov Diffusion Process which has generator $\mathcal{L}$, which is a differential operator that characterizes the dynamics of the process and is closely related to the state transition functions semigroup of solution operators associated with the processes time evolution, it is used to describe the time evolution of the state variable $\mathbf{X}$'s probability distribution, as defined in section 2.6 and is given as

$$\mathcal{L} = \mathbf{f} \cdot \nabla + \frac{1}{2}\mathbf{g}\mathbf{g}^\mathsf{T}\nabla^2$$

given an initial distribution $\rho_0 = \rho(0)$ then $\mathcal{L}\rho_0 : \rho_0 \to \rho(t)$ where again ρ is an $C^{1,2}$ function and is Lipschitz. The conditional densities can only be evaluated and updated when observations become available - however since the model continuous and $\mathbf{y}$ is discrete - then the transition probability of the model has to be known. However, we know that the Fokker - Planck equation describes the evolution of the transition probabilities.

0.5.5 Forecasting

For the conditional density $t_k \leq t \leq t_{k+1}$ given by $\rho(\mathbf{x}(t)|\mathbf{y}(t))$ with known initial distribution ρ_0 satisfies the Fokker-Planck equation

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho \mathbf{f}) + \frac{1}{2}\nabla^2 \cdot (\Gamma \rho)$$

when working with a data stream, a convenient approximation of the Wiener Process is to take the expected value of the noise up to time t where $\mathbb{E}[d\mathbf{W}(t)d\mathbf{W}^\mathsf{T}(t)] = \mathbf{R}(t)dt$ is a matrix $\mathbf{R} \in \mathbb{M}_{d\times d}(\mathbb{R})$ so that the resulting FPE is given by

$$\frac{\partial \rho}{\partial t} = -\sum_{k=0}^{d} \frac{\partial (\rho f_k)}{\partial x_k} + \frac{1}{2}\sum_{k=0}^{d}\sum_{l=0}^{d} \frac{\partial^2}{\partial x_k \partial x_l} \left\{ \left[\mathbf{g}\mathbf{R}\mathbf{g}^{\mathsf{T}}\right]_{kl} \rho \right\}$$

0.5.6 Updating the distribution

Assuming that past data is available, that is $\rho(\mathbf{x}(t-1)|\mathbf{y}_{1:t-1})$ is known. Then by the Markov Property the updated density function is given by

$$\rho(\mathbf{x}(t)|\mathbf{y}_{1:t-1}) = \int_{\mathbb{R}^d} \rho(\mathbf{x}(t)|\mathbf{x}(t-1))\rho(\mathbf{x}(t-1)|\mathbf{y}_{1:t-1})d\mathbf{x}(t-1)$$

with posterior distribution

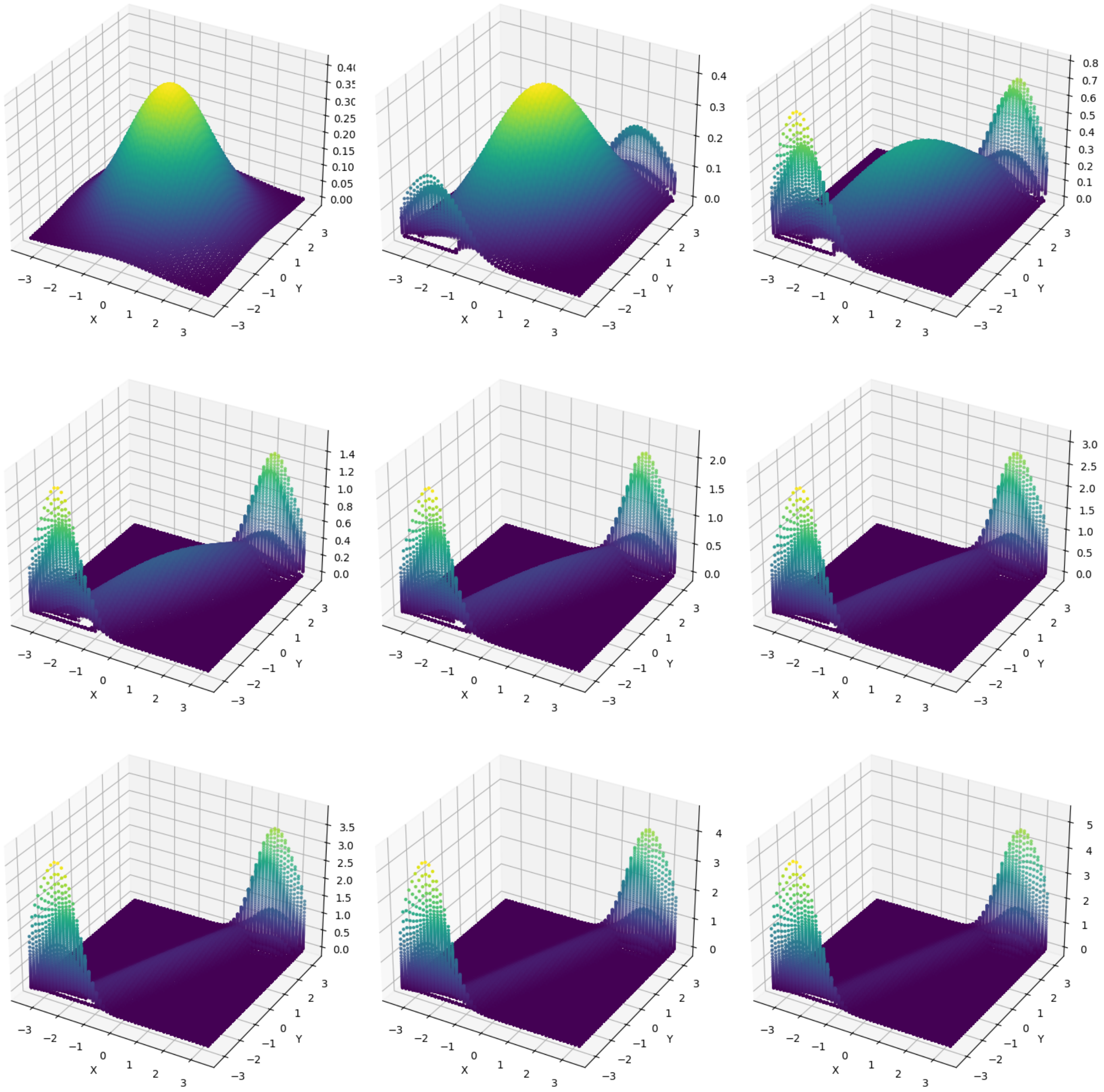
$$\begin{aligned} \rho(\mathbf{x}(t)|\mathbf{y}_{1:t}) &\propto \rho(\mathbf{y}(t)|\mathbf{x}(t))\rho(\mathbf{x}(t)|\mathbf{y}_{1:t-1}) \\ &= \rho(\mathbf{y}(t)|\mathbf{x}(t)) \int_{\mathbb{R}^d} \rho(\mathbf{x}(t)|\mathbf{x}(t-1))\rho(\mathbf{x}(t-1)|\mathbf{y}_{1:t-1})d\mathbf{x}(t-1) \end{aligned}$$

These two equations can then by looped in time to obtain filtered density functions. In the Linear-Gaussian setting these equations reduce to matrix formulae known as the Kalman Filter, however in practice $\mathbf{x}$ and $\mathbf{y}$ may be high dimensional and $\mathbf{f}$ and H are nonlinear. Therefore only approximate solutions can be found.

0.5.7 Utilizing the Fokker-Planck Equation for Filtering

Following the work of Miller 1994, Evensen 1996 [18, 19], when modelling dynamics with random parameter values. The density function associated with what is often a stochastic differential equation, evolves according to the fokker planck equation where the spatial dimensions are equal to that of the state dimension of the SDE. Many applications solve this type of equation through monte-carlo methods, some examples of monte carlo methods used in solving the FPE are particle filters [20]. Within this framework and following the numerical example for the lorenz dynamics - we take the observations of the dynamics as density functions themselves, that is, the state-space distribution Φ rather than singular points - as such in the Kalman Filter [16] and like. By using the Fokker-Planck equation one can directly compute the posterior distribution using Baye's formula. Even in the case where the initial condition ρ_0 is Gaussian in many systems it will not stay Gaussian - one such (simple) example is the Lorenz equations with the initial condition set to a Gaussian distribution. The same conditions as our numerical example are used but with the initial condition $\rho \sim \mathcal{N}(\mathbf{0}, \Sigma)$ where

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$



which does not provide as accurate results than using Φ (this is just a generalisation of a simple histogram) since there are two peaks whereas in the system there is only a single attracting set. However the most likely result is correct. The main point, however, is that the distribution is not Gaussian. In models where there are more than one maximum - least squares methods, such as the Kalman Filter [16] which assumes a linear - gaussian setting, may become useless. This is because a difference of the two peaks is often made. The isolation of a single point of most likely probability may not be found or at most will not reflect the true result - this will often result in 0 probability of finding anything.

0.6 Stochastic Data Assimilation

We now give a data assimilation construction based on the FPE and the assumptions made in this passage. We begin with an initial probability distribution ρ and integrate the fokker-planck equation. When data becomes available we use Bayes Theorem to construct a posterior PDF. At some time t data becomes available through the observation operator H. Here the prior PDF will be the solution to the Fokker-Planck equation. This describes the most likely position to find an object at at particular time t. The relation from 5.6 becomes

$$\begin{aligned}\rho(\mathbf{x}(t)|\mathbf{y}_{1:t}) \propto \rho(\mathbf{y}(t)|\mathbf{x}(t))\rho(\mathbf{x}(t)|\mathbf{y}_{1:t-1}) &= \rho(\mathbf{y}(t)|\mathbf{x}(t))\frac{\partial \rho}{\partial t} \\ &= \rho(\mathbf{y}(t)|\mathbf{x}(t))\left(-\nabla \cdot (\rho \mathbf{f}) + \frac{1}{2}\nabla^2 \cdot (\Gamma \rho)\right)\end{aligned}$$

where the conditional probability density $\rho(\mathbf{y}(t)|\mathbf{x}(t))$ is given by the relation

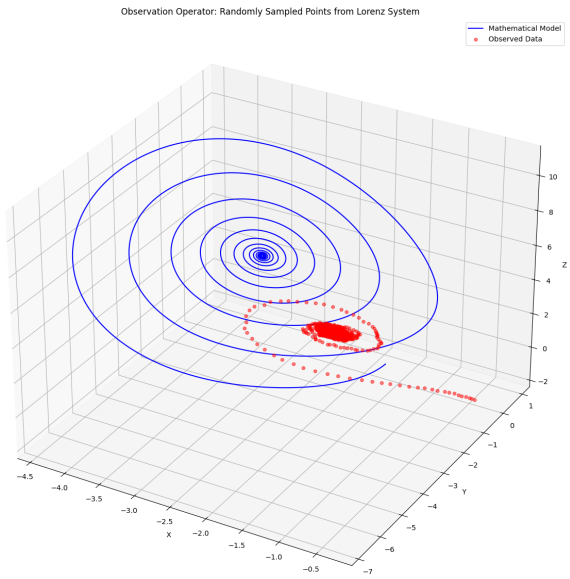
$$\rho\left(H\left(\mathbf{x}(t)\right) - \mathbf{x}(t)\right) = \rho(\mathbf{y}(t) - \mathbf{x}(t))$$

This represents the conditional probability density function of the discrepancy between the predicted state $\mathbf{x}$ based on the system model and the actual measured state $H(\mathbf{x})$ at time t. Therefore, the likelihood function quantifies how likely it is to observe the actual measurement given a particular state hypothesis $\mathbf{x}$ (our mathematical model). In other words, it measures the agreement between the predicted measurement and the observed measurement. In the case where ξ and η are Gaussian Noise then $\eta \sim \mathcal{N}(\mu, \mathbf{\Sigma})$

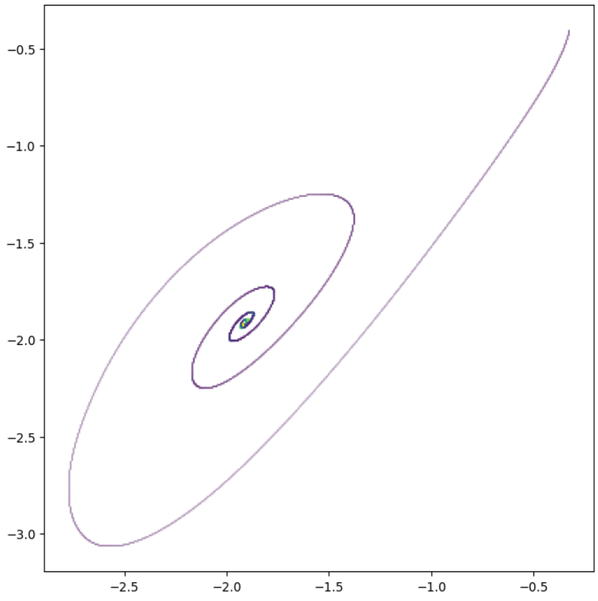
$$\rho(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\mathbf{\Sigma})}} \exp\left\{\frac{1}{2}\left[H\left(\mathbf{x}(t)\right) - \mathbf{x}(t)\right]^{\mathsf{T}} \mathbf{\Sigma}^{-1}\left[H\left(\mathbf{x}(t)\right) - \mathbf{x}(t)\right]\right\}$$

0.7 Numerical Example: Stochastic Data Assimilation for Lorenz 1963

Here we provide a simulation of a Lorenz model observations of a different, but similar Lorenz model with stochastic forcing to simulate observation error. The stochastic lorenz model is similar to that in the previous simulation for the Fokker-Planck equations. Let L_1 be a deterministic model for the Lorenz system with model parameters given by $\sigma_1 = 3.765432$, $\rho_1 = 17.654323456$ and $\beta_1 = 8/3$ with initial conditions $\dot{x}_{1,0} = -3.345$, $\dot{y}_{1,0} = -3.2$ and $\dot{z}_{1,0} = 1$. Now suppose parameters σ_1, ρ_1 and β_1 along with the initial conditions $\dot{x}_{1,0}, \dot{y}_{1,0}, \dot{z}_{1,0}$ are different in the observed dynamics. We wish to update our probability distribution function $\rho(\mathbf{x})$ to better represent the observed dynamics, given the data from our model. For L_1 we have solved the fokker-planck equation in the previous numerical example to obtain a prior distribution. Define L_2 as the observed system - with initial conditions $\sigma_2 = 7.9654345876543$, $\rho_2 = 6.765434567$ and $\beta_2 = 3.9432$ and initial conditions $\dot{x}_{2,0} = -0.8$, $\dot{y}_{2,0} = -1$ and $\dot{z}_{2,0} = -3.7654$. The observations are randomly sampled from a sarcastically forced Lorenz system L_2 with noise intensity $d_n = 5$ and Gaussian Noise.



The samples of the trajectory are taken every 10 steps and a total of 1000 samples of the observed trajectory are taken. The phase space distribution of points for L_2 is seen to focus at $(-1.8, -1.8, z)$.

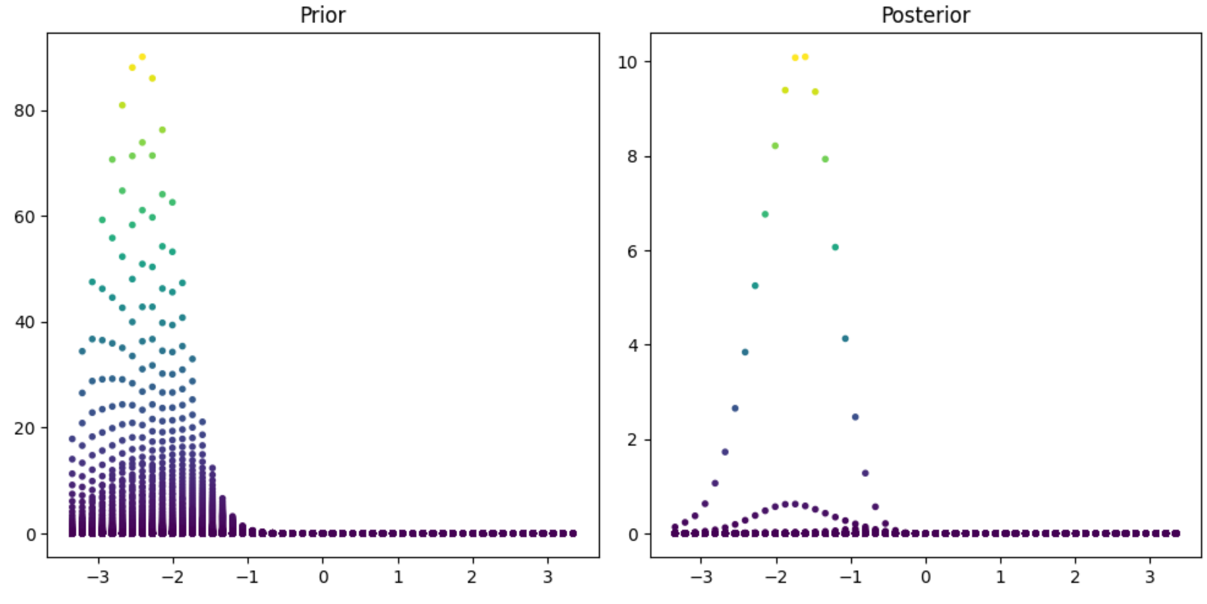


We compute the posterior with the expression

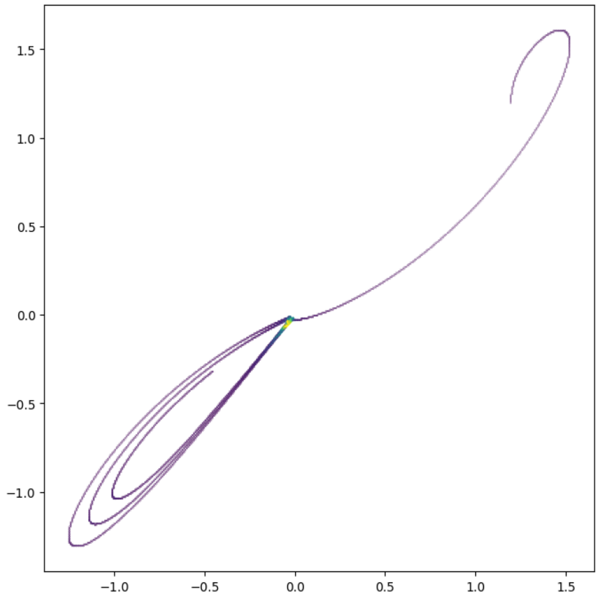
$$\begin{aligned}
\rho(\mathbf{y}(t)|\mathbf{x}(t))\frac{\partial \rho}{\partial t} &= \frac{1}{Z}\rho(\mathbf{y}(t)|\mathbf{x}(t))\left(-\nabla\cdot(\rho\mathbf{f})+\frac{1}{2}\nabla^2\cdot(\Gamma\rho)\right)\\
&= \frac{1}{Z}\rho(\mathbf{y}(t)-\mathbf{x}(t))\left(-\nabla\cdot(\rho\mathbf{f})+\frac{1}{2}\nabla^2\cdot(\Gamma\rho)\right)\\
&= \frac{1}{Z}\left|\exp\left(-\left(\Phi_{L_2}-\Phi_{L_1}\right)\right)\right|^2\left(-\nabla\cdot(\rho\mathbf{f})+\frac{1}{2}\nabla^2\cdot(\Gamma\rho)\right)\\
&= \frac{1}{Z}\left|\exp\left(-\left[\Phi_{L_2}-\left(-\nabla\cdot(\rho\mathbf{f})+\frac{1}{2}\nabla^2\cdot(\Gamma\rho)\right)\right]\right)\right|^2\left(-\nabla\cdot(\rho\mathbf{f})+\frac{1}{2}\nabla^2\cdot(\Gamma\rho)\right)\\
&= \frac{1}{Z}\left|\exp\left(-\left(\Phi_{L_2}-\frac{\partial\rho}{\partial t}\right)\right)\right|^2\frac{\partial\rho}{\partial t}
\end{aligned}$$

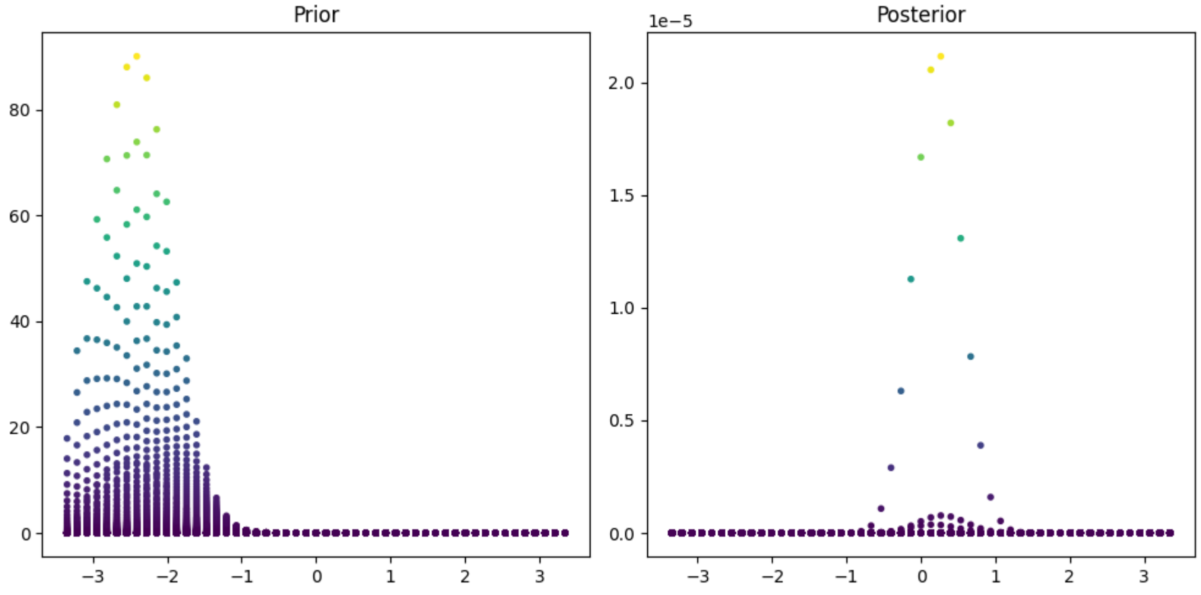
This represents the relative importance or weight assigned to the observed values of the phase space distribution of the observed values Φ_{L_2} in updating the posterior distribution. If the observed values are highly consistent with the predicted values Φ_{L_1}, the exponential term would be close to 1, indicating high importance. Conversely, if there is a significant discrepancy between observed and predicted values, the exponential term will decrease to 0, suggesting lower importance. The absolute value of the exponential term emphasizes the regions where the predicted state (prior PDF) is more consistent with the observations (smaller difference leads to a larger term after squaring). Multiplying by $\partial_t\rho$ introduces a weighting factor based on the existing probability density. Areas with higher prior probability will have a larger influence on the update. Due to the Gaussian nature of the noise in this example. The FPE at each time step includes the prior information about the dynamics of the model up to but not including the time t_k.

$$\begin{aligned}
\rho(\mathbf{x}(t)|\mathbf{y}_{1:t_k}) &= \frac{1}{Z}\left|\exp\left(-\left(\Phi_{L_2}-\frac{\partial\rho}{\partial t_k}\right)\right)\right|^2\frac{\partial\rho}{\partial t_k}\\
&= \frac{1}{(2\pi)^{d/2}\sqrt{\det(\mathbf{I}_3)}}\left|\exp\left(-\left(\Phi_{L_2}-\frac{\partial\rho}{\partial t_k}\right)\right)\right|^2\frac{\partial\rho}{\partial t_k}
\end{aligned}$$



which has successfully located the most probable state of the observed system given variable model parameters and initial conditions. Thus this Lorenz system can be tracked in various scenarios. A further example is given with L_2 model parameters given by $\sigma_2 = 7.9654345876543$, $\rho_2 = 6.765434567$ and $\beta_2 = 0.1$ with initial conditions $\dot{x}_{2,0} = 3$, $\dot{y}_{2,0} = 3$ and $\dot{z}_{2,0} = 3$.





Part III

Supervised Learning of Operators

The objective of supervised learning in dynamical systems is to determine the underlying mapping $\Psi : \mathcal{U} \rightarrow \mathcal{V}$ from samples of data $\{u_n, \Psi(u_n)\}$ where u_n is sampled from a distribution of points for a dynamical model.

0.8 Operator Learning

The success of deep learning in various fields of study [21] has spiked interest in applying deep learning to scientific problems. Scientific Machine Learning [21] combines traditional modeling of phenomena through Partial Differential Equations (PDEs) with observation data and Neural Network based Machine Learning. *Operator Learning* [22] aims to discover or approximate an unknown operator $\mathcal{A}$, which takes the form of a solution operator Ψ to a differential equation. The Solution Operator for a given partial differential equation is the mapping from the coefficients, initial value or boundary values to the corresponding solution. Therefore, finding solution operators is of particular importance when the solution of the same PDE is needed for different values for the initial conditions or different configurations for the boundary values [22].

In other words, the variable inputs are directly mapped to the corresponding solutions for the whole underlying PDE, rather than specific solutions given specific conditions, this removes the need to resolve the PDE for a different set of inputs. [21]

0.8.1 Mathematical Formulation of Operator Learning

Following Boull´e and Townsend and Li et al [cite, cite], we define an operator $\mathfrak{L}$ to be the Fokker-Planck equation and we aim to learn the solution operator $\Psi_{\mathfrak{L}}$ from data. Given pairs of data $(\rho(\mathbf{x}, t), \rho(\mathbf{x}, t+1))$ with $\rho(\mathbf{x}, t) \in \mathcal{U}$ and $\rho(\mathbf{x}, t+1) \in \mathcal{V}$, which are function spaces on a spatial domain $\Omega \subseteq \mathbb{R}^d$.

The operator takes the function $\rho(\mathbf{x}, t)$ as an initial state and tells us the future state $\rho(\mathbf{x}, t+1)$. $\mathcal{U}$ represents the space of all possible initial probability density functions. $\mathcal{V}$ represents the space of all possible future probability density functions. Both $\mathcal{U}$ and $\mathcal{V}$ are defined on a spatial domain Ω in d-dimensional space like a specific region in which the object moves. We define the Fokker-Planck Operator as a function space mapping such that

$$\Psi_{\mathfrak{L}} : \mathcal{U} \rightarrow \mathcal{V}$$
$$\Psi_{\mathfrak{L}}\left(\rho(\mathbf{x}, t)\right) = \rho(\mathbf{x}, t+1)$$

The idea is to find an approximation to $\Psi_{\mathfrak{L}}$, we denote the approximation as $\hat{\Psi}_{\mathfrak{L}}$ - this means that for any new data $p(\mathbf{x}, t)$, which is not used to discover the solution operator, we are left with $\hat{\Psi}_{\mathfrak{L}}\left(p(\mathbf{x}, t)\right) \approx \Psi_{\mathfrak{L}}\left(p(\mathbf{x}, t)\right)$ where $p(\mathbf{x}, t)$ satisfies

$$\frac{\partial p}{\partial t} = -\nabla \cdot (p\mathbf{f}) + \frac{1}{2}\nabla^2 \cdot (\Gamma p)$$
$$\mathfrak{L} = -\nabla \cdot (\cdot\mathbf{f}) + \frac{1}{2}\nabla^2 \cdot (\Gamma \cdot)$$

in machine learning literature the data p is said to be unseen to the neural network, since it was not trained on the data. This ensures the learned mapping is general enough so that the neural network is not over fitting to specific instances of the Fokker-Planck equation, that is, it solve specific type of Fokker-Planck equation but does not solve another. We do this by showcasing different types of Fokker-Planck equations (characterized by different drift and diffusion terms), we want it to learn the underlying principles rather than memorize specific examples. This prevents over fitting, where the neural network would perform well on the training data but fail on unseen data with slightly different dynamics.

The neural operator is a generalisation of a neural network, the inputs and outputs are functions defined on function spaces, not vectors on vector spaces - that is, finite-dimensional Euclidean spaces. This has major benefits because once the solution operator $\Psi_{(.)}$ is learned then one hopes that an entire family of differential equations are then learned. In the examples which have been presented up until now, to solve for many instances of the FPE the solver had to be iterated and the equation resolved and set at specific resolutions each time. This is cumbersome and takes much time - notwithstanding the often small and crude resolution which is tractable for many.

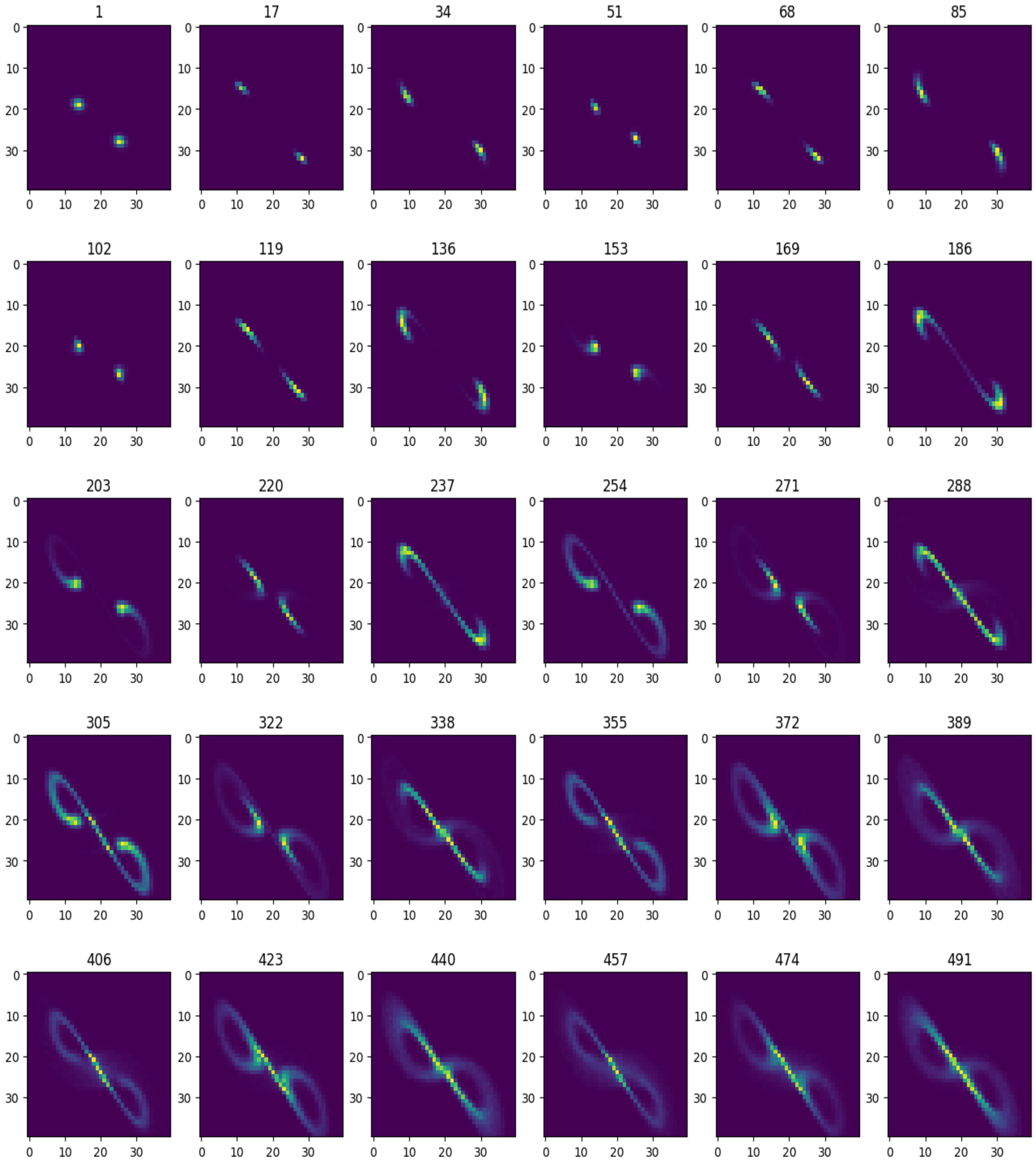
The data is used to train the model $\hat{\Psi}_{\mathfrak{L}}$, we measure error in the training process

$$\left\|\Psi(u) - \hat{\Psi}_{\mathfrak{L}}(u)\right\| = \int_{\Omega} \left|\mathscr{L}\left((u(\mathbf{x}), t)\right) - \mathfrak{L}(u(\mathbf{x}), t)\right|^2 d\mathbf{x}dt$$

which we wish to minimize such that

$$\theta^* = \arg\min_{\theta\in\Theta} \left\|\Psi(u) - \hat{\Psi}_{\mathfrak{L}}(u)\right\|$$

where $\mathscr{L}\left((u(\mathbf{x}), t)\right)$ is the sequentially approximated fokker-planck equation at each time step. Numerical solutions to the FPE are available for the Lorenz Model, they can be used as training data for the network along with samples of Lorenz system trajectories generated through simulations and are used to test the network.



The code implementation is provided at `https://github.com/AdamJamesSheppard/`

Appendix A

Further Mathematical Definitions

A.1 Analysis and Measures

Definition A.1. Let a set $\Omega \neq \emptyset$ and $\mathcal{P}(\Omega) = \{A : A \subset \Omega\}$ be the power set of Ω. A set $\mathcal{F} \subset \mathcal{P}(\Omega)$ is called an *algebra* if

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F}$ where this implies $A^c \in \mathcal{F}$ together
3. For two sets A and B then $A \cup B \in \mathcal{F}$

Definition A.2. An algebra $\mathcal{F}$ is called a $\sigma - algebra$ if for a sequence of sets $A_n \in \mathcal{F}$ where $n \geq 1$ then

$$\bigcup_{n \geq 1} A_n \in \mathcal{F}$$

and all $A_n \subset A_{n+1}$ and $|A_n| \leq |A_{n+1}|$.

Definition A.3. A class of subsets $\mathcal{A} = \{A_1, ..., A_m | A_m \subset \Omega; m < n\}$, where each A_j is a subset of Ω then the $\sigma-$algebra generated by $\mathcal{A}$ is defined as

$$\sigma \langle \mathcal{A} \rangle = \bigcap_{\mathcal{A} \subset \mathcal{F}_j} \mathcal{F}_j$$

the intersection yields the smallest σ-algebra containing all subsets in $\mathcal{A}$ and $\mathcal{F}_j \subset \mathcal{F}$ and is a σ-algebra on Ω.

Definition A.4. A set $E \subseteq \mathbb{R}$ is *open* if for every $x \in E$ there exists some $\epsilon(x) > 0$ such that $\{y : x - \epsilon < y < x + \epsilon\} \subseteq E$.

Definition A.5. A set $F \subseteq \mathbb{R}$ is *closed* if F^c is open (see A.Def 4)

Definition A.6. A class $\mathbb{O} = \{O_1, O_2, O_3, O_4\}$ of open subsets of $\mathbb{R}$ where

1. $O_1 = \{(a_k, b_k) : -\infty \leq a_i < b_i \leq \infty;\ 1 \leq i \leq k\}$
2. $O_2 = \{(-\infty, x_k) : x_k \in \mathbb{R}\}$
3. $O_3 = \{(a_k, b_k) : a_i, b_i \in \mathbb{Q}; a_i < b_i;\ 1 \leq i \leq k\}$
4. $O_4 = \{(-\infty, x_k) : x_k \in \mathbb{Q}\}$

generates a σ-algebra $\mathcal{B}(\mathbb{R}) = \sigma\langle\mathbb{O}\rangle$ called the *borel* σ-algebra generated by the open subsets on $\mathbb{R}$.

Remark A.1. Definition (A.6) extends to $\mathbb{R}^n$ by taking the cartesian product of a euclidean space such that

$$\begin{aligned}
O'_1 &= \left\{ \bigtimes_{k=1}^{n} (a_k, b_k) : -\infty \leq a_i < b_i \leq \infty; 1 \leq i \leq k \leq n \right\} \\
O'_2 &= \left\{ \bigtimes_{k=1}^{n} (-\infty, x_k) : x_k \in \mathbb{R}, k \leq n \right\} \\
O'_3 &= \left\{ \bigtimes_{k=1}^{n} (a_k, b_k) : a_i, b_i \in \mathbb{Q}; a_i < b_i;\ 1 \leq i \leq k \leq n \right\} \\
O'_4 &= \left\{ \bigtimes_{k=1}^{n} (-\infty, x_k) : x_k \in \mathbb{Q}, k \leq n \right\}
\end{aligned}$$

and so then $\mathbb{O}' = \{O'_1, O'_2, O'_3, O'_4\}$ and $\mathcal{B}(\mathbb{R}^n) = \sigma\langle\mathbb{O}'\rangle$ and in terms of Borel σ-algebras defined on $\mathbb{R}$ we have

$$\mathcal{B}(\mathbb{R}^n) = \bigtimes_{k=1}^{n} \mathcal{B}(\mathbb{R})$$

Definition A.7. A set Ω and a σ-algebra $\mathcal{F}$ form are a pair $(\Omega, \mathcal{F})$ called a *measurable space*

Definition A.8. A *measure* on a measurable space $(\Omega, \mathcal{F})$ is a mapping $\mu : \mathcal{F} \to [0, \infty]$ such that

1. $\mu(\emptyset) = 0$
2. $\mathcal{F}$ is a σ-algebra on Ω

and if for a sequence of sets $A_n \in \mathcal{F}$ where $A_n \subset A_{n+1}$ and $|A_n| \leq |A_{n+1}|$ then

$$\mu\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$$

Definition A.9. A triple $(\Omega, \mathcal{F}, \mu)$ such that for any $A_i, A_j \in \mathcal{F}$ then $A_i \cap A_j = \emptyset$ then $(\Omega, \mathcal{F}, \mu)$ is called a *measure space.*

Definition A.10. For a measure space $(\Omega, \mathcal{F}, \mu)$ if

1. $\mu(\Omega) < \infty$ then μ is a *finite* measure.
2. there exists a sequence of sets $A_n \subset \mathcal{F}$ such that

   $$\mathcal{F} = \bigcup_{n\geq 1} A_n$$

   and $\mu(A_n) < \infty$ for all n then μ is $\sigma - finite.$

Definition A.11. A measure space $(\Omega, \mathcal{F}, \mathbb{P})$ is a *probability space* with *probability measure* $\mathbb{P}$ whenever $\mathbb{P}$ is a finite measure on the measurable space $(\Omega, \mathcal{F})$ and $\mathbb{P}(\Omega) = 1$ and $\mathbb{P} : \mathcal{F} \to [0, 1]$.

[1] Ioannis Karatzas and Steven E. Shreve. Brownian Motion and Stochastic Calculus. Springer Verlag, second edition, 1991. ISBN 3-387-97655-8.

[2] Bernt Øksendal. Stochastic Dierential Equations. Springer Verlag, 6th edition, 2005. ISBN 3-540-25662-8

[3] Andrew H. Jazwinski. Stochastic Processes and Filtering Theory. Dover Publications INC, rst edition, 1970. ISBN 13: 978-0-48646274-5

[4] Ioannis Karatzas and Steven E. Shreve. Brownian Motion and Stochastic Calculus. Springer Verlag, second edition, 1991. ISBN 3-387-97655-8

[5]Patrick Nima Raanes Improvements to Ensemble Methods for Data Assimilation in the Geosciences https://ora.ox.ac.uk/objects/uuid:9f9961f0-6906-4147-a8a9-ca9f2d0e4a12/files/m91718d0ba191be2ed1e0e0b7231986b2

[6] Schmallfuß, B. The random attractor of the stochastic Lorenz system. Z. angew. Math. Phys. 48, 951–975 (1997). https://doi.org/10.1007/s000330050074

[7]M. S. Alnaes, A. Logg, K. B. Ølgaard, M. E. Rognes and G. N. Wells. Unified Form Language: A domain-specific language for weak formulations of partial differential equations, ACM Transactions on Mathematical Software 40 (2014). [arXiv] [doi.org/10.1145/2566630]

[8]M. W. Scroggs, J. S. Dokken, C. N. Richardson, and G. N. Wells. Construction of arbitrary order finite element degree-of-freedom maps on polygonal and polyhedral cell meshes, ACM Transactions on Mathematical Software 48(2) (2022) 18:1–18:23. [arXiv] [doi.org/10.1145/3524456]

[9]I. A. Baratta, J. P. Dean, J. S. Dokken, M. Habera, J. S. Hale, C. N. Richardson, M. E. Rognes, M. W. Scroggs, N. Sime, and G. N. Wells. DOLFINx: The next generation FEniCS problem solving environment, preprint (2023). [doi.org/10.5281/zenodo.10447666]

[10]M. W. Scroggs, I. A. Baratta, C. N. Richardson, and G. N. Wells. Basix: a runtime finite element basis evaluation library, Journal of Open Source Software 7(73) (2022) 3982. [doi.org/10.21105/joss.03982]

[11]Zheng, WZ., Liang, Y. & Huang, JP. Equilibrium state and non-equilibrium steady state in an isolated human system. Front. Phys. 9, 128–135 (2014). https://doi.org/10.1007/s11467-013-0337-5

[12]Kolmogorov, Andrei (1931). "Über die analytischen Methoden in der Wahrscheinlichkeitstheorie" [On Analytical Methods in the Theory of Probability]. Mathematische Annalen (in German). 104 (1): 415–458 [pp. 448–451]. doi:10.1007/BF01457949. S2CID 119439925.

[13] Journal of Approximation Theory 95, 90100 (1998) On Krein's Theorem for Indeterminacy of the Classical Moment Problem Henrik L. Pedersen

[14] 1942, Extrapolation, Interpolation and Smoothing of Stationary Time Series. A war-time classified report nicknamed "the yellow peril" because of the color of the cover and the difficulty of the subject. Published postwar 1949 MIT Press. http://www.isss.org/lumwiener.htm Archived 2015-08-16 at the Wayback Machine])

[15]Kalman, R. E.; A New Approach to Linear Filtering and Prediction Problems, Transactions of the ASME - Journal of Basic Engineering Vol. 82: pag. 35-45 (1960).

[16]T. Cipra & A. Rubio; Kalman filter with a non-linear non-Gaussian observation relation, Springer (1991).

[17]Stochastic differential equations occurring in the estimation of continuous parameter stochastic processes Theory of Probability and its Applications, 1969, Volume 14, Issue 4, Pages 567–594 DOI: https://doi.org/10.1137/1114076

[18] Stochastic Methods for Sequential Data Assimilation in Strongly Nonlinear Systems DOI: https://doi.org/10.1175/152
0493(2001)129<1194:SMFSDA>2.0.CO;2

[19] Data assimilation into nonlinear stochastic models ROBERT N. MILLER, EVERETT F. CARTER Jr., SALLY T. BLUE First published: 19 September 2002 https://doi.org/10.1034/j.1600-0870.1999.t01-2-00002.x

[20]Wills, Adrian G.; Schön, Thomas B. (3 May 2023). "Sequential Monte Carlo: A Unified Review". Annual Review of Control, Robotics, and Autonomous Systems. 6 (1): 159–182. doi:10.1146/annurev-control-042920-015119. ISSN 2573-5144. S2CID 255638127.

[21] arXiv:2312.14688 [math.NA]

[22] arXiv:2108.08481 [cs.LG]