Review of probability theory

Can you predict where a leaf falling from a tree will land? Will there be clouds above Santa Cruz tomorrow at noon? Being scientists, we know that there are physical laws and models we could integrate in time which may provide an answer to such questions. For the falling leaf, we have the equations of fluid mechanics coupled with the equations describing the leaf mechanics (fluid-structure interaction). For the weather forecast in Santa Cruz, we have a quite complicated (usually data-driven) model for dynamics of the atmosphere. However, even if we firmly believe that our equations truthfully represent physical reality, i.e., that there is no *model uncertainty*, we still have some problem when making inferences on the two systems mentioned above. In the case of the leaf falling from the tree, we do not know the exact shape of the leaf, nor the distribution of mass within the leaf, nor whether there is a tiny wind gust pushing the leaf in a direction we did not expect, or having it flipping in a way we did not anticipate. We can of course try to control some of these uncertainties, e.g., by designing a *sterilized experiment* in which we are reasonably sure that there is no wind gust and we know "exactly" the geometry, mechanics, and mass distribution of the leaf. Would the result of such an experiment be useful to make inferences about the behavior of the falling leaf in real world? Perhaps not.

Alternatively, we could study the system for which we have available equations and physical laws using techniques that allow us to account for uncertainties in the initial condition, boundary conditions, forcing terms, geometry,and physical parameters.

The most common approach to study uncertainty propagation, and perhaps the first one that was ever developed, is *random sampling*. In this approach we basically study the response of the system, e.g., the trajectory of the leaf and where it lands, corresponding to randomly sampled realizations of the uncertain parameters and stochastic processes driving the system. Such parameters and processes can be modeled as random variables, random functions or random fields. Computing the solution to such *stochastic models* by sampling involves solving the ODE/PDE system many times, so that a sufficiently large ensemble of solutions is available to compute statistics such as mean, standard deviations, and even probability distribution functions. There are many different types of sampling methods that were developed for this purpose. For instance, Monte Carlo methods and their variants (quasi-MC, multi-level MC, etc.), sparse grids, probabilistic collocation methods, etc. Sampling method are often classified as *non-intrusive*. This means that we do not need to modify the equations of our model to perform uncertainty analysis, but simply sample them many times for different conditions.

Another approach to compute the statistics of a given model problem (set of equations describing a physical system, neural network, etc) in the presence of uncertainties is to represent to output of the model relative to a set of stochastic basis functions, e.g., multivariate polynomials of random variables with given probability distributions. This approach is known as *polynomial chaos* (PC) [23], and has many different variants (generalized PC, multi-element PC). The method allows us to compute the solution to a model problem with a small number of random variables and often exhibits exponential convergence rate. Polynomial chaos and related methods based on series expansions of the model problem are often classified as *intrusive methods*. The adjective "intrusive" emphasizes the fact that the equations of motion to propagate uncertainty are problem-dependent and require an ad-hoc derivation and corresponding coding.

A third class of methods relies on transforming the model problem from the state space to probability space and solve for the probability density function of the solution. An example of such transformation is the Liouville equation (linear hyperbolic PDE) for the probability density function of the solution of a nonlinear dynamical system evolving from a random initial state. Another example is the Fokker-Plank equation governing the PDF of the state of a nonlinear dynamical system driven by random (white) noise. In the context of partial differential equations (infinite-dimensional dynamical systems), the PDF equations corresponding to the solution of nonlinear PDEs evolving from random initial states are *functional differ-*

ential equations [20, 18, 2]. Probability density function methods can also be used also to model and study neural nets (neural nets are essentially discrete dynamical systems [21]). Moreover, all Bayesian inference approaches, e.g., Gaussian regression, probabilistic graphical models, and data assimilation techniques, heavily rely probability density function methods.

PDF methods are also very attractive for systems with *unknown governing equations* (if such equations even exist!), or systems for which governing equations can be discovered only "locally" and in an approximate form. Examples of such systems are mathematical models of brain, large random networks of interacting individuals, the mechanical behavior random heterogeneous materials, disease propagation, stock market models. In all these cases it not straightforward to derive a computational model that accurately describes the system in all its features, and can be used to accurately forecast quantities of interest. Recent advances in data-driven modeling and artificial intelligence, open the possibility to discover model and equations from data. Of course, dealing with model uncertainty on the top of uncertainty in operating conditions, parameters, forcing terms, etc., opens a whole new dimension to the problem of modeling and prediction.

It also raises deep philosophical questions regarding the appropriateness of the mathematics we are using to build our models, and therefore the validity of our computations.

Probability space

There is a well-developed mathematical theory that allows us to describe randomness in the world we live in, or at least the way we perceive it. Such theory is known as probability theory [15]. The proper mathematical foundations of probability theory are quite abstract and technical, as they involve rather advanced concepts of *measure theory* [9]. However, for our purposes it possible to avoid most technicalities and have a version of probability theory that allows for computation, and can be digested by the most (including myself). Let me describe hereafter the basic ingredients of such theory.

To formally describe the outcome of an "experiment" from a mathematical viewpoint it is convenient to define the *probability space* $(\Omega, \mathcal{F}, P)$ which consists of the following items:

- Ω (sample space): the set of all possible outcomes of the experiment
- $\mathcal{F}$ (event space): set of events, en event being a set defined as union or intersection of elements in the sample space.
- P (probability measure): this function assigns each event in $\mathcal{F}$ a probability, which is a number between 0 and 1.

Example 1: Suppose the experiment is rolling a fair dice with 6 faces once. In this case we can define the sample space as

$$\Omega = \{1, 2, 3, 4, 5, 6\} \qquad \text{(sample space)}. \tag{1}$$

The definition of the event space depends on what we are interested in. In particular, we may be interested in the following events:

$$\mathcal{F} = \{\emptyset, \Omega, \underbrace{\{1, 3, 5\}}_{\text{odd}}, \underbrace{\{2, 4, 6\}}_{\text{even}}\}. \tag{2}$$

These events can be phrased as: "rolling the dice produces no number" (event $\emptyset$); "rolling the dice returns any number between 1 and 6" (event Ω); "rolling the dice gives an even number" (event $\{2, 4, 6\}$), "rolling the dice returns an odd number" (event $\{1, 3, 5\}$). Clearly, we can assign probabilities to these events as:

$$P(\emptyset) = 0, \qquad P(\Omega) = 1, \qquad P(\{1, 3, 5\}) = \frac{1}{2}, \qquad P(\{2, 4, 6\}) = \frac{1}{2}. \tag{3}$$

Note that in this case assigning probabilities is rather straightforward as we can imagine the process of rolling a dice, and its outcome quite easily. In a similar way, we can assign, e.g., the probability of winning various prizes in the Powerball or the Mega-Millions (assuming the lottery is fair). A rather different story is when we are asked to assign probabilities to complex processes influenced by many variables, e.g., where the leaf falling from a tree is going to land.

Example 2: Let $(\theta(\omega), r(\omega))$ be the polar coordinated identifying where the leaf falling off a tree is going to land. Suppose that $r = 0$ identifies the center of the tree. Clearly $(\theta(\omega), r(\omega))$ is a vector with two random components. In this case, the outcome of the experiment are realizations of two real random variables (coordinates $(\theta(\omega), r(\omega))$ of the leaf after it lands). We can define the following set of events (distance from the tree):

$$\mathcal{F} = \{\emptyset, \Omega, \underbrace{\{\omega : r(\omega) \leq 1\}}_{\text{event 1}}, \underbrace{\{\omega : 1 < r(\omega) \leq 2\}}_{\text{event 2}}, \underbrace{\{\omega : r(\omega) > 2\}}_{\text{event 3}}\}. \tag{4}$$

At this point we need to assign probabilities[1] to each event in (4), which can be done, e.g., by running a very complicated fluid dynamics model (repeated simulations), or by observing many many leaves falling off a tree.

Example 3: Consider an infinite (uncountable) collection of continuous functions $X(t;\omega)$ (stochastic process) defined in the temporal interval $[0, T]$. Let the sample space Ω be the collection of such functions and consider the event space $\mathcal{F}$

$$\mathcal{F} = \{\emptyset, \Omega, \underbrace{\{\omega : X(t;\omega) < 1\}}_{\text{event 1}}, \underbrace{\{\omega : X(t;\omega) \geq 1\}}_{\text{event 2}}\}. \tag{5}$$

In other words, here we are interested in two events only, namely whether the stochastic process $X(t, \omega)$ is (for all $t \in [0, T]$) strictly smaller than one, or larger or equal to one. We can assign a probability to each event in $\mathcal{F}$, e.g., as

$$P(\emptyset) = 0, \qquad P(\Omega) = 1, \qquad P(\{\omega : X(t;\omega) < 1\}) = a, \qquad P(\{\omega : X(t;\omega) \geq 1\}) = 1 - a, \tag{6}$$

where $a \in [0, 1]$. Again, the way a is computed depends on the statistical characterization of the process $X(t;\omega)$. In other words, the calculation leading to $P(\{\omega : X(t;\omega) < 1\})$ may involve quite a lot of operations. Alternatively, the probability of an event $E \in \mathcal{F}$ can be estimated using a frequency approach, i.e., $P(E) \simeq n_E/n$, where n_E is the number E occurs over n trials.

σ-algebra. As we shall see hereafter, in order to perform set operations and corresponding operations on probabilities we need to make sure that $\mathcal{F}$ has the structure of a σ-algebra on Ω. A σ-algebra on Ω is a collection of subsets of Ω that is closed under complement, countable unions, and countable intersections. In other words,

$$A, B \in \mathcal{F} \quad \Rightarrow \quad \begin{cases} A \cap B \in \mathcal{F} \\ A \cup B \in \mathcal{F} \\ A^c, B^c \in \mathcal{F} \quad \text{(complement of } A \text{ and } B\text{, i.e., } A^c = \Omega \setminus A) \end{cases} \tag{7}$$

From this conditions it also follows that $\emptyset, \Omega \in \mathcal{F}$. Moreover, if $\{A_i\}_{i=1}^{\infty} \in \mathcal{F}$ then

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}, \qquad \bigcap_{i=1}^{\infty} A_i \in \mathcal{F} \qquad \text{(countable union and intersection)}. \tag{8}$$

[1] For a thorough discussion on the meaning of probability and how to assign probabilities see [15, Chapters 1-3].

Examples of σ-algebras:

- Consider the sample space $\Omega = \{a, b, c\}$. The power set of Ω, i.e., the combination of all possible elements of Ω (including the empty set), is a σ-algebra.

$$2^\Omega = \{\emptyset, a, b, c, \{a, b\}, \{a, c\}, \{b, c\}, \underbrace{\{a, b, c\}}_{\Omega}\} \qquad \text{(power set)}. \tag{9}$$

  The cardinality of the power set, i.e., the number of elements of the set 2^Ω is equal to $2^{\#\Omega}$ (where $\#$ denotes the number of elements of a set). In the specific case of (9) we have $\#\Omega = 3$, and therefore $\#2^\Omega = 2^3 = 8$.

- If the sample space Ω is countably infinite (i.e., the elements of Ω can be put in a correspondence with $\mathbb{N}$) then the power set 2^Ω is isomorphic to $\mathbb{R}$, i.e., it is an uncountable set.

- If the sample space Ω is uncountably infinite, e.g., $\Omega = [0, 1]$ then any σ-algebra $\mathcal{F}$ on Ω can be represented as a sub-algebra of the power set 2^Ω (Stone's representation theorem [9]). This is why the σ-algebra $\mathcal{F}$ on an uncountably infinite sample space Ω is often written as a subset of the power set 2^Ω, i.e., $\mathcal{F} \subseteq 2^\Omega$.

- The σ-algebra on $\Omega = \mathbb{R}$ is the σ-algebra of the collection of all open subsets of $\mathbb{R}$. Such σ-algebra necessarily contains all open sets, all closes sets, and all (countable) unions and intersections of open and closed sets. Such σ-algebra is a sub-algebra of the power set 2^Ω.

Probability measure. The probability function

$$P : \mathcal{F} \to [0, 1] \tag{10}$$

assigns to each event A in the σ-algebra F a number $P(A) \in [0, 1]$. In other words, $P(A)$ measures the likelihood that A occurs. The probability function P satisfies the properties of a *measure* (hence the name probability measure[2]):

1. $P(\emptyset) = 0$.
2. $P(\Omega) = 1$.
3. For all countable collections of *disjoint* sets $A_i \in \mathcal{F}$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \tag{11}$$

4. For all $A, B \in \mathcal{F}$,

$$P\left(A \cup B\right) = P(A) + P(B) - P(A \cap B). \tag{12}$$

From these properties it follows that of the event $B \in \mathcal{F}$ is a subset of $A \in \mathcal{F}$ then $A = B \cup (B^c \cap A)$, which implies that (note that B and $B^c \cap A$ are disjoint)

$$P(A) = P(B) + P(B^c \cap A) \geq P(B). \tag{13}$$

[2]In real analysis, the pair $(\Omega, \mathcal{F})$ is called *measurable space* [9]. The elements of $\mathcal{F}$, i.e., the events, are called *measurable sets*. The triple $(\Omega, \mathcal{F}, P)$ is called *probability space*, which is essentially a measurable space in which we define a probability measure.

Frequency interpretation of the probability measure: Suppose that in an experiment the event A shows up n_A times out of n trials. If we define the empirical distribution

$$\mu_n(A) = \frac{n_A}{n} \tag{14}$$

then

$$P(A) = \lim_{n\to\infty} \mu_n(A). \tag{15}$$

Random variables

Let $(\Omega, \mathcal{F}, P)$ be a probability space. A real-valued random variable $X(\omega)$ is a measurable map from the sample space Ω into $\mathbb{R}$, i.e.,

$$X : \Omega \to \mathbb{R}. \tag{16}$$

The *distribution function* of the random variable $X(\omega)$ is defined as

$$F(x) = P(\underbrace{\{\omega : X(\omega) \leq x\}}_{\text{event}}) \qquad x \in \mathbb{R}. \tag{17}$$

The distribution function represents the measure of the set (event) $\{\omega \in \Omega : X(\omega) \leq x\}$, i.e., the probability that $X(\omega)$ is smaller than a given real number x. By using the properties of the probability measure P it is straightforward to conclude that:

1. $F(-\infty) = 0$,
2. $F(\infty) = 1$,
3. $F(x)$ is non-decreasing, i.e., $x_1 < x_2 \quad \Rightarrow F(x_1) \leq F(x_2)$,
4. $P(\{\omega : X(\omega) > x\}) = 1 - F(x)$,
5. $F(x)$ is continuous from the right, i.e.,

$$\lim_{\epsilon\to 0^+} F(x+\epsilon) = F(x), \tag{18}$$

6. $F(x)$ is not continuous from the left (for discrete random variables),
7. $P(\{\omega : a < X(\omega) \leq b\}) = F(b) - F(a)$
8. $P(\{\omega : a \leq X(\omega) \leq b\}) = F(b) - \lim_{\epsilon\to 0^+} F(a+\epsilon)$.

The proof of 1.-8. can be found in [15, Chapter 4].

If $F(x)$ is continuous in x then we say that the random variable $X(\omega)$ is *continuous*. If $F(x)$ is a staircase function then the random variable $X(\omega)$ is *discrete*. $F(x)$ is discontinuous and not staircase, then we say that $X(omega)$ is *mixed*.

Frequency interpretation of the distribution function $F(x)$: Suppose we perform an experiment n-times and observe n realization of the random variable $X(\omega)$, say $\{X(\omega_1), \ldots, X(\omega_n)\}$. Let us place all these numbers on the x axis of a Cartesian plane, and form a staircase function, where each step at $X(\omega_i)$ has height $1/n$. Then the staircase function $F_n(x)$ converges to $F(x)$ in the limit $n \to \infty$.

Probability density function. The probability density function (PDF) $p(x)$ of the random variable $X(\omega)$ is (technically speaking) the Radon–Nikodym derivative[3] (assuming it exists) of the probability measure P. The existence of the Radon–Nikodym derivative allows us to write the cumulative distribution function (17) as

$$F(x) = \int_{-\infty}^{x} p(y)dy. \tag{21}$$

Equivalently $p(x)$ can be interpreted as the (weak) derivative of $F(x)$, i.e.,

$$p(x) = \frac{dF(x)}{dx}. \tag{22}$$

By taking the limit of Lebesgue-integrable Dirac delta sequences, we can make sense of Radon–Nikodym PDFs converging to Dirac deltas. This is useful when dealing with the PDF of deterministic (non-random) variables, or discrete random variables. For example,

$$p(x) = \delta(x-a) \qquad \text{(PDF of the random variable } X(\omega) = a \text{ for all } \omega \in \Omega), \tag{23}$$

and

$$p(x) = \sum_{i=1}^{N} p_i \delta(x - x_i) \qquad \text{(PDF of a discrete random variable with range } \{x_1, \ldots, x_n\}). \tag{24}$$

In particular, the PDF of a fair dice with 6 faces is

$$p(x) = \frac{1}{6}\sum_{i=1}^{6} \delta(x-i). \tag{25}$$

By using the properties of the cumulative distribution function $F(x)$ it is straightforward to derive the following properties for the PDF

$$p(x) \geq 0 \quad \text{(positivity)}, \qquad \int_{-\infty}^{\infty} p(x)dx = 1 \quad \text{(normalization)}. \tag{26}$$

Other properties are

$$P\left(\{\omega : x_1 < X(\omega) \leq x_2\}\right) = \int_{x_1}^{x_2} p(x)dx, \qquad P\left(\{\omega : x < X(\omega) \leq x + dx\}\right) = p(x)dx \tag{27}$$

Frequency interpretation of PDFs: Suppose we sample the random variable $X(\omega)$ n times and find that $n_{\Delta x}$ samples fall between x and $x + \Delta x$. By using equation (27), and the frequency interpretation of probability we conclude that

$$p(x)\Delta x \simeq \frac{n_{\Delta x}}{n} \quad \Rightarrow \quad p(x) \simeq \frac{1}{\Delta x}\frac{n_{\Delta x}}{n}. \tag{28}$$

[3] A probability measure P on the measurable space $(\Omega, \mathcal{F})$ is said to be *absolutely continuous* with respect to another measure ν if for all events $E \in \mathcal{F}$ such that $P(E) = 0$ we have $\nu(E) = 0$. In other words, P is absolutely continuous with respect to ν if all impossible events (measured relative to P) are also impossible relative to ν. This is denoted as $\nu \ll P$. Consider, in particular, the Lebesgue measure $d\nu = dx$. The Radon-Nikodym theorem says that if P is absolutely continuous with respect to the Lebesgue measure, then there exists a unique function $p(x)$ such that

$$P(E) = \int_E p(x)dx. \tag{19}$$

Setting the event E in (19) as

$$E = \{w : X(\omega) \leq x\} \in \mathcal{F} \tag{20}$$

yields equation (21).

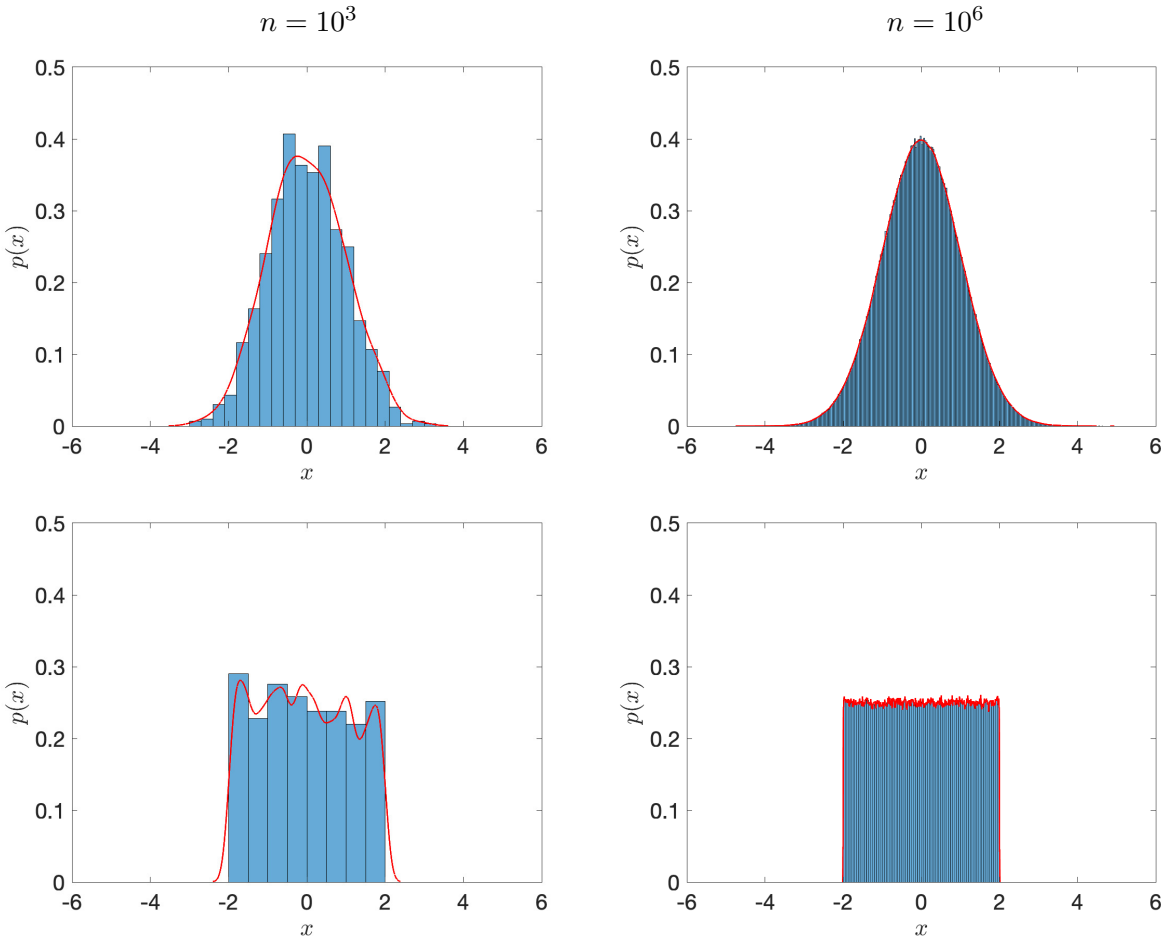

Figure 1: Estimation of the PDF of a Gaussian random variable (first row) and a uniform random variable (second row) using the frequency approach, i.e., formula (28), and the kernel density estimate discussed in [4] (red line) . We plot results for a different number of samples n.

Hence, by dividing the support of the random variable $X(\omega)$ into bins and counting the number of samples within each bin allows us to estimate the PDF of $X(\omega)$ in a rather straightforward way. This is at the basis of the Monte-Carlo estimation method for random variables. There are of course more effective methods to estimate the PDF of one random variable from data (see, e.g., [4]). In figure 1 we estimate the PDF of a Gaussian random variable using frequency approach, i.e., equation (28), and the kernel density estimation method discussed in [4].

Examples of one-dimensional PDFs:

- Gaussian (continuous):

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad x \in \mathbb{R}. \tag{29}$$

- Uniform (continuous):

$$p(x) = \frac{1}{b-a}, \qquad x \in [a, b]. \tag{30}$$

- Binomial (discrete):

$$p(x) = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \delta(x-i), \qquad p \in]0,1[, \qquad x \geq 0. \tag{31}$$

- Poisson (discrete):

$$p(x) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \delta(x-k), \qquad \lambda \in]0,\infty[, \qquad x \geq 0. \tag{32}$$

Functions of one random variable. In this section we discuss how to compute the probability density function of a random variable $Y(\omega)$ defined as a deterministic nonlinear function of another random variable $X(\omega)$. To this end, let $(\Omega, \mathcal{F}, P)$ be a probability space,

$$g : \mathbb{R} \to \mathbb{R} \tag{33}$$

a deterministic function,

$$X, Y : \Omega \to \mathbb{R} \tag{34}$$

random variables. Suppose we are given the PDF $p_X(x)$ of $X(\omega)$, and that

$$Y(\omega) = g(X(\omega)) \tag{35}$$

for all $\omega \in \Omega$. What is PDF $Y(\omega)$? Since X and Y are defined on the same probability space we have

$$F_Y(y) = P\left(\{\omega : Y(\omega) \leq y\}\right) = P\left(\{\omega : g(X(\omega)) \leq y\}\right). \tag{36}$$

Therefore, to determine the distribution function $F_Y(y)$ we just need to measure the set

$$B_y = \{\omega : g(X(\omega)) \leq y\} \tag{37}$$

for each y in the set of $g(\mathscr{R}(X))$ (where $\mathscr{R}(X)$ denotes the range of the random variable X). The set B_y is shown in Figure 2 (in yellow) for a prototype function $g(x)$ and a specific value of y. Clearly, the distribution function $F_Y(y)$ must be defined case-by-case. With reference to Figure 2 we have

$$F_Y(y) = F_X(x_1(y)) + 1 - F_X(x_2(y)), \tag{38}$$

where $x_1(y)$ and $x_2(y)$ are the branches of the inverse function $g^-1(y)$. The function (38) represents the distribution function of Y in terms of cumulative distribution function of X, which we know.

With the cumulative distribution function of Y available, it is straightforward to compute the PDF of Y, by taking the (weak) derivative of $F_Y(y)$. This is formalized in the following theorem

Theorem 1. Let X be a random variable with PDF $p_X(x)$, $g \in C^1(\mathbb{R})$ a continuously differentiable function. Then the PDF of $Y = g(X)$ is given by

$$p_Y(y) = \sum_{i=1}^{r} \frac{p_X(x_i(y))}{|g'(x_i(y))|}, \tag{39}$$

where $x_i(y)$ $(i = 1, \ldots, r)$ are the real roots of the equation $g(x) = y$, and $g'(x_i(y))$ is assumed to be non-zero[4].

[4]If $g'(x_i(y)) = 0$ then formula (39) does not apply, and we need to resort to a different method. For example we can use the distribution function approach outlined in Figure 2, i.e., we could measure sets depending on y with the probability measure P and connect such set to the distribution function of X.

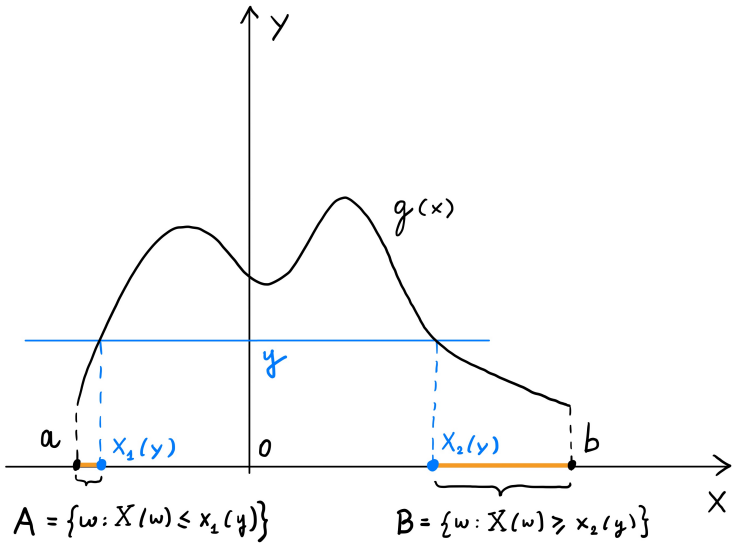

Figure 2: Sketch of the set B_y defined in equation (37) (yellow lines). The random variable X is compactly supported in $[a, b]$. The distribution function of the random variable $Y(\omega) = g(X(\omega))$ evaluated at y is the measure of the set $B_y = A \cup B$ (union of the two yellow lines), i.e., $F_Y(y) = F_X(x_1(y)) + 1 - F_X(x_2(y))$.

Proof. We prove the theorem using Fourier transforms[5]. Let

$$\phi_Y(a) = \int_{-\infty}^{\infty} e^{iay} p_Y(y) dy = \int_{-\infty}^{\infty} e^{iag(x)} p_X(x) dx. \tag{40}$$

Taking the inverse Fourier transform yields

$$p_Y(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{ia(g(x)-y)} p_X(x) dx da. \tag{41}$$

Next, recall that

$$\delta\left(g(x) - y\right) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ia(g(x)-y)} da. \tag{42}$$

Substituting (42) into (41) yields (see also [10])

$$p_Y(y) = \int_{-\infty}^{\infty} \delta\left(g(x) - y\right) p_X(x) dx. \tag{43}$$

At this point we use the well-known identity[6]

$$\delta\left(g(x) - y\right) = \sum_{i=1}^{r} \frac{\delta\left(x - x_i(y)\right)}{|g'\left(x_i(y)\right)|}, \tag{44}$$

where $x_i(y)$ are the real roots of the $y = g(x)$ for each $y \in \mathbb{R}$. A substitution of (44) into (43) yields (39). This completes the proof.

□

[5]The Fourier transform of a the probability density function $p_X(x)$ is known as *characteristic function* of the random variable $X(\omega)$ (see Eq. (90)).

[6]The identity (44) if and only if $g'(x_i(y)) \neq 0$.

Examples of probability density mappings: Let X be a random variable with probability density function $p_X(x)$. In the following examples we derive the PDF of $Y = g(X)$ for a few prototype $g(x)$.

- Consider the random variable $Y(\omega) = X(\omega)^2$. The mapping $y = g(x) = x^2$ between the random variables X and Y can be inverted (with real roots) for all $y \geq 0$. This yields

$$x_1(y) = \sqrt{y}, \qquad x_2(y) = -\sqrt{y} \qquad y \geq 0. \tag{45}$$

  By using Theorem 1 we immediately obtain

$$p_Y(y) = \frac{1}{2\sqrt{y}} \left[p_X(\sqrt{y}) + p_X(-\sqrt{y}) \right]. \tag{46}$$

  For instance, if $p_X(x)$ is Gaussian , i.e.,

$$p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{47}$$

  then

$$p_Y(y) = \frac{e^{-y/2}}{\sqrt{2\pi y}} \qquad (\chi^2\text{-distribution}). \tag{48}$$

  Similarly, if X is uniformly distributed in $[-1, 1]$ then[7]

$$p_Y(y) = \frac{1}{2\sqrt{y}} \quad \text{for all } 0 \leq y \leq 1. \tag{49}$$

- Consider the random variable $Y(\omega) = e^{tX(\omega)}$, where $t \geq 0$ is a real parameter. The mapping $y = g(x) = e^{tx}$ can be inverted (with unique solution) for all $y > 0$ as

$$x = \frac{\log(y)}{t} \qquad y > 0. \tag{50}$$

  The derivative of $g(x)$ is $g'(x) = te^{tx}$. Therefore

$$p_Y(y) = \frac{1}{ty} p_X\left(\frac{\log(y)}{t} \right) \qquad y > 0. \tag{51}$$

Application to dynamical systems. Let us briefly discuss two applications of the PDF mapping technique to simple one-dimensional dynamical systems.

- Consider the following Cauchy problem for one ODE evolving from a random initial state

$$\begin{cases} \dfrac{dx}{dt} = f(x) \\ x(0) = X(\omega) \end{cases} \tag{52}$$

  We know from AM 214 that if f is continuously differentiable in x then the system generates a smooth flow map $x(t) = x(t, X(\omega))$ (differentiable in X) that takes any initial state $X(\omega)$ (at $t = 0$) and maps it to the corresponding solution at time t. Given the PDF of the initial condition $p_X(x)$ we can compute the PDF of $x(t)$ as

$$p(x, t) = \int_{-\infty}^{\infty} \delta\left(x - x(t, y) \right) p_X(y) dy. \tag{53}$$

[7] Recall that a uniform PDF in $[-1, 1]$ is $p_X(x) = 1/2$ for all $x \in [-1, 1]$.

A convenient way to actually compute such PDF is by sampling, i.e., compute sample paths of $x(t)$ corresponding to samples of $X(\omega)$. However, if the flow map is available analytically then we can also compute (53) analytically. To this end, consider the system

$$\begin{cases} \dfrac{dx}{dt} = x^2 \\ x(0) = X(\omega) \end{cases} \tag{54}$$

We known that the analytical solution (flow map) is

$$x(t, X) = \frac{X(\omega)}{1 - tX(\omega)}. \tag{55}$$

Suppose that $X(\omega)$ is uniformly distributed in $[-1, 0]$ so that the flow map exists for all $t \geq 0$ (no blow-up). What is then the PDF of $x(t, X)$ at each fixed time t? Clearly, we can invert

$$g(x) = \frac{x}{1 - tx} = y \tag{56}$$

uniquely for each $t \geq 0$ ($x \leq 0$) as

$$x(1 + ty) = y \quad \Rightarrow \quad x(y) = \frac{y}{1 + ty}. \tag{57}$$

The first derivative of (56) with respect to x evaluated at the unique root $x(y) = y/(1 + ty)$ is

$$g'(x) = \frac{1}{(1 - tx(y))^2} = (1 + ty)^2. \tag{58}$$

At this point we use Theorem 1 to conclude that the PDF of the solution to the ODE (54) at each fixed time t is

$$p(x, t) = \frac{1}{(1 + tx)^2} p_X \left(\frac{x}{1 + tx} \right). \tag{59}$$

In particular, if p_X is the PDF of a uniform random variable in $[-1, 0]$ then the support of $p(x, t)$ is defined by the condition

$$-1 \leq \frac{x}{1 + tx} \leq 0 \quad \Rightarrow -\frac{1}{(1 + t)} \leq x \leq 0. \tag{60}$$

Hence, as t goes to infinity the support of $p(x, t)$ shrinks to 0 and $p(x, t)$ converges to a Dirac delta function at $x = 0$. Note that for each fixed t we have that the normalization condition of the PDF $p(x, t)$ is satisfied. In fact,

$$\int_{-\infty}^{\infty} p(x, t) dx = \int_{-1/(1+t)}^{0} \frac{1}{(1 + tx)^2} \underbrace{1}_{p_X\left(\frac{x}{1+tx}\right)} dx = 1. \tag{61}$$

- Next, consider the linear decay problem

$$\begin{cases} \dfrac{dx}{dt} = \xi(\omega) x \\ x(0) = 1 \end{cases} \tag{62}$$

where $\xi(\omega)$ is a random variable with known probability density $p_\xi(x)$. The analytical solution to (62) is

$$x(t, \xi(\omega)) = e^{t\xi(\omega)}. \tag{63}$$

By using equation (51), we immediately conclude that the probability density of the solution to (62) is

$$p(x,t) = \frac{1}{tx} p_X\left(\frac{\log(x)}{t}\right) \qquad x > 0. \tag{64}$$

For instance, if p_X is a uniform PDF in $[-2, 0]$ then the support of $p(x,t)$ is defined by

$$-2t \leq \log(x) \leq 0 \quad \Rightarrow \quad e^{-2t} \leq x \leq 1. \tag{65}$$

At $t = 0$ the PDF of the solution is supported only at one point, i.e., $x = 1$. Indeed,

$$p(x,0) = \delta(x-1) \quad \text{(deterministic initial condition)}. \tag{66}$$

For $t > 0$ the PDF of the solution to (62) is[8]

$$p(x,t) = \frac{1}{2tx} \quad \text{for} \quad e^{-2t} \leq x \leq 1. \tag{68}$$

Data-driven identification of the PDF of the initial state. What is the probability density of the initial state $p(x,0)$ that generates an envelope of trajectories that is as close as possible to a measured quantity of interest $h(x(t))$? This is an inverse problem that can be solved by minimizing a performance metric, i.e., a dissimilarity measure between the measurements at various times and the envelope of trajectories, over the parameters representing the initial probability density function.

Liouville equation. The PDF of the solution to the Cauchy problem (52) satisfies the following linear hyperbolic conservation law (see Appendix A of the present course note)

$$\frac{\partial p(x,t)}{\partial t} + \frac{\partial}{\partial x}\left(f(x)p(x,t)\right) = 0, \qquad p(x,0) = p_X(x). \tag{69}$$

This equation is known as Liouville equation. It is straightforward to show by using the method of characteristics that (59) is the solution of Liouville equation (69) for $f(x) = x^2$, i.e., for the dynamical system (54). In Appendix A, we prove that the joint probability density function of the phase space variables of any n-dimensional nonlinear dynamical system

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}(\boldsymbol{x}), \qquad \boldsymbol{x}(0) = \boldsymbol{X}(\omega) \tag{70}$$

evolving from a random initial state satisfies $\boldsymbol{X}(\omega)$ satisfies the Liouville equation

$$\frac{\partial p(\boldsymbol{x},t)}{\partial t} + \nabla \cdot \left(\boldsymbol{f}(\boldsymbol{x})p(\boldsymbol{x},t)\right) = 0. \tag{71}$$

To solve (71) one could propagate characteristic curves from the support of random initial state $p(\boldsymbol{x},0)$, or use more sophisticated methods, e.g., numerical tensor methods [7, 8] or physics-informed neural network techniques [16].

Sampling from arbitrary one-dimensional PDFs. Let $X(\omega)$ be a uniform random variable in $[0,1]$. We would like to find a mapping $g(X)$ such that the (continuous) random variable $Y(\omega) = g(X)$ has a

[8] Note that the PDF (68) integrates to one. In fact

$$\int_{-\infty}^{\infty} p(x,t)dx = \int_{e^{-2t}}^{1} \frac{1}{tx} \underbrace{\frac{1}{2}}_{p_X\left(\frac{\log(x)}{t}\right)} dx = 1. \tag{67}$$

desired probability density $p_Y(y)$. With such mapping g available we can transform each sample of $X(\omega)$ to a sample of the PDF p_Y, hence constructing a sampler for $Y(\omega)$. As we shall see hereafter, if we denote by $F_Y(y)$ the cumulative distribution of the continuous random variable Y (the random variable we are interested in sampling) then the mapping g is simply the inverse of F_Y, i.e., $Y(\omega) = F_Y^{-1}(X(\omega))$.

Lemma 1. Let $X(\omega)$ be a uniform random variable in $[0,1]$. Consider a second random variable $Y(\omega)$ with PDF p_Y and cumulative distribution function

$$F_Y(y) = \int_{-\infty}^{y} p_Y(x)dx \tag{72}$$

The random variable $Y = F_Y^{-1}(X)$ has cumulative distribution function $F_Y(y)$.

Proof. Suppose that F_Y is invertible. Let us show that the random variable $F_Y^{-1}(X)$ has indeed cumulative distribution function $F_Y(y)$. By definition,

$$\begin{aligned} F_Y(y) &= P\left(\{\omega : Y(\omega) \leq y\}\right) \\ &= P\left(\{\omega : F_Y^{-1}(X(\omega)) \leq y\}\right) \\ &= P\left(\{\omega : X(\omega) \leq F_Y(y)\}\right) \qquad (F_Y \text{ invertible and nondecreasing}) \\ &= F_X(F_Y(y)) \\ &= F_Y(y). \end{aligned} \tag{73}$$

In fact, since $X(\omega)$ is uniform in $[0,1]$ we have $F_X(x) = x$ for all $x \in [0,1]$.

□

Expectation, moments and cumulants. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $X : \Omega \to \mathbb{R}$ a random variable with cumulative distribution function $F_X(x)$ and PDF $p_X(x)$. For any function $g(X)$ we define the *expectation* of $g(X)$ as [9]

$$\mathbb{E}\left\{g(X)\right\} = \int_{-\infty}^{\infty} g(x)dF_X(x) = \int_{-\infty}^{\infty} g(x)p_X(x)dx. \tag{75}$$

Clearly, if $Y(\omega) = g(X(\omega))$ is a random variable with PDF $p_Y(y)$, we can equivalently express the expectation as

$$\mathbb{E}\left\{g(X)\right\} = \mathbb{E}\left\{Y\right\} = \int_{-\infty}^{\infty} ydF_Y(y) = \int_{-\infty}^{\infty} yp_Y(y)dy. \tag{76}$$

In particular, if we set $g(X) = X^k$ then $\mathbb{E}\left\{X^k\right\}$ are called *moments*[10] of the random variable X

$$\mathbb{E}\left\{X^k\right\} = \int_{-\infty}^{\infty} x^k dF_X(x) = \int_{-\infty}^{\infty} x^k p_X(x)dx. \tag{77}$$

[9] We do not need to assume the existence of the PDF to define the expectation operator. In fact, a more general expression for (75) is

$$\mathbb{E}\left\{g(X(\omega))\right\} = \int_{\Omega} g(X(\omega))dP(\omega). \tag{74}$$

[10] There are random variables for which moments do not exist. An example is the Cauchy random variable. Random variables with compactly supported range have all moments. For such compactly supported random variables it is always possible to reconstruct the PDF p_X from the knowledge of its moments or cumulants. In other words, the so-called *moment problem* has a unique solution for compactly supported PDFs.

The first few moments of a random variable X are

$$\mathbb{E}\left\{X\right\} = \int_{-\infty}^{\infty} x p_X(x) dx \qquad \text{(mean)}, \tag{78}$$

$$\mathbb{E}\left\{X^2\right\} = \int_{-\infty}^{\infty} x^2 p_X(x) dx \qquad \text{(second-order moment)}, \tag{79}$$

$$\mathbb{E}\left\{X^3\right\} = \int_{-\infty}^{\infty} x^3 p_X(x) dx \qquad \text{(third-order moment)}. \tag{80}$$

The moments of random variable are the coefficients of the power series expansion of the so-called *moment generating function*

$$M(a) = \mathbb{E}\left\{e^{aX(\omega)}\right\} \tag{81}$$

In fact,

$$M(a) = M(0) + \underbrace{\frac{dM(0)}{da}}_{\mathbb{E}\{X\}} a + \frac{1}{2} \underbrace{\frac{d^2M(0)}{da^2}}_{\mathbb{E}\{X^2\}} a^2 + \cdots . \tag{82}$$

In general,

$$\mathbb{E}\left\{X^k\right\} = \frac{d^k M(0)}{da^k}. \tag{83}$$

A function related to the moment generating function is the *cumulant generating function*

$$\Psi(a) = \log(M(a)). \tag{84}$$

The coefficients of the power series expansion of $\Psi(a)$ are called *cumulants* of the random variable $X(\omega)$

$$\Psi(a) = \Psi(0) + \underbrace{\frac{d\Psi(0)}{da}}_{\mathbb{E}\{X\}} a + \frac{1}{2} \underbrace{\frac{d^2\Psi(0)}{da^2}}_{\mathbb{E}\{X^2\}-\mathbb{E}\{X\}^2} a^2 + \cdots \tag{85}$$

The cumulants of a random variable X are often denotes as $\left\langle X^k \right\rangle_c$. For example, we have

$$\left\langle X \right\rangle_c = \mathbb{E}\left\{X\right\}, \tag{86}$$

$$\left\langle X^2 \right\rangle_c = \mathbb{E}\left\{X^2\right\} - \mathbb{E}\left\{X\right\}^2, \tag{87}$$

$$\left\langle X^3 \right\rangle_c = \mathbb{E}\left\{X^3\right\} - 3\mathbb{E}\left\{X\right\}\mathbb{E}\left\{X^2\right\} + 2\mathbb{E}\left\{X\right\}^3, \tag{88}$$

$$\cdots$$

The quantity

$$\left\langle X^2 \right\rangle_c = \mathbb{E}\left\{X^2\right\} - \mathbb{E}\left\{X\right\}^2, \tag{89}$$

is the *variance* of the random variable X. Finally, we define the the *characteristic function* of the random variable $X(\omega)$ as

$$\phi(a) = \mathbb{E}\left\{e^{iaX(\omega)}\right\} \tag{90}$$

where i is the imaginary unit. We have seen this function already, i.e., in the proof of Theorem 1. The characteristic function is the Fourier transform of the probability density function $p(x)$. It is straightforward to show that

$$\mathbb{E}\left\{X^k\right\} = \frac{1}{i^k} \frac{d^k \phi(0)}{da^k}. \tag{91}$$

By expanding the complex exponential function in a power series, and using the definition of cumulants we obtain the following *cumulant expansion* of $\phi(a)$ (see, e.g., [12])

$$\phi(a) = \exp\left[\sum_{j=1}^{\infty} \left\langle X^j \right\rangle_c \frac{(ia)^j}{j!}\right]. \tag{92}$$

Example: The characteristic function of a Gaussian random variable with mean μ and variance σ^2 is

$$\phi(a) = e^{i\mu a - \sigma^2 a^2/2}. \tag{93}$$

This expression can be derived by the taking the Fourier transform of (29), or by using (92). In fact, for Gaussian random variables we have that only the first two cumulants are non-zero, i.e.,

$$\langle X \rangle_c = \mathbb{E}\{X\} = \mu, \tag{94}$$

$$\left\langle X^2 \right\rangle_c = \mathbb{E}\left\{X^2\right\} - \mathbb{E}\{X\}^2 = \sigma^2, \tag{95}$$

$$\left\langle X^k \right\rangle_c = 0 \quad \text{for all } k \geq 3. \tag{96}$$

Substituting these expressions into (92) yields (93).

Random vectors

Let $(\Omega, \mathcal{F}, P)$ be a probability space. A real-valued random vector $\boldsymbol{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))$ is a measurable map from Ω into $\mathbb{R}^n$, i.e.,

$$\boldsymbol{X} : \Omega \to \mathbb{R}^n. \tag{97}$$

Each component $X_i(\omega)$ of the random vector $\boldsymbol{X}(\omega)$ is a real-valued random variable. The *distribution function* of the random vector $\boldsymbol{X}(\omega)$ is defined as

$$F(x_1, \ldots, x_n) = P(\underbrace{\{\omega : X_1(\omega) \leq x_1\} \cap \cdots \cap \{\omega : X_n(\omega) \leq x_n\}}_{\text{element of } \mathcal{F} \text{ (event) defined as intersection of events}}). \tag{98}$$

As before, if P is absolutely continuous with respect to the Lebesgue measure $dx_1 \cdots dx_n$ then there exists a (Lebesgue integrable) probability density function[11] $p(x_1, \ldots, x_n)$ such that

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} p(y_1, \ldots, y_n) dy_1 \cdots dy_n. \tag{99}$$

Equivalently, we can express $p(x_1, \ldots, x_n)$ as a (weak) derivative of $F(x_1, \ldots, x_n)$ as

$$p(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n}. \tag{100}$$

The multivariate distribution function F and associated probability density function p satisfy similar properties as the properties we have seen for one one random variable (see [15] for details).

Frequency interpretation of the joint PDF: Suppose we observe realizations of a random vector $\boldsymbol{X}(\omega)$ with only two components, i.e., $X_1(\omega)$ and $X_2(\omega)$. By using (98)-(99), we have

$$P(\{\omega : x_1 \leq X_1(\omega) \leq x_1 + \Delta x_1\} \cap \{\omega : x_2 \leq X_2(\omega) \leq x_2 + \Delta x_2\}) \simeq p(x_1, x_2) \Delta x_1 \Delta x_2. \tag{101}$$

[11] As before, the probability density function $p(x_1, \ldots, x_n)$ is the Radon-Nikodym derivative of the probability measure P relative to the Lebesgue measure $dx_1 \cdots dx_n$.

Let us partition the tensor product space $\mathbb{R}^2$ with an evenly-spaced grid of width Δx_1 (along x_1) and Δx_2 (along x_2). Suppose we observe n realizations of the random vector $\boldsymbol{X}(\omega) = (X_1(\omega), X_2(\omega))$, and suppose that $n_A < n$ instances satisfy the condition

$$\{x_1 \leq X_1(\omega) \leq x_1 + \Delta x_1\} \quad \text{and} \quad \{x_2 \leq X_2(\omega) \leq x_2 + \Delta x_2\}. \tag{102}$$

Then from (101) we obtain the PDF estimate

$$p(x_1, x_2) \simeq \frac{1}{\Delta x_1 \Delta x_2} \frac{n_A}{n}. \tag{103}$$

More efficient and accurate methods to estimate the PDF from data are based on kernels [4] (see Figure 4)

Marginal probability density and marginal distribution. Let $\boldsymbol{X}(\omega) = (X_1(\omega), X_2(\omega))$ be a random vector with joint distribution function function $F(x_1, x_2)$. The distribution of the random variable $X_1(\omega)$ can be obtained from $F(x_1, x_2)$ simply by sending x_2 to infinity, i.e.,

$$F(x_1) = \lim_{x_2 \to \infty} F(x_1, x_2). \tag{104}$$

In fact,

$$\lim_{x_2 \to \infty} F(x_1, x_2) = P(\{\omega : X_1(\omega) \leq x_1\} \cap \{\omega : X_2(\omega) \leq \infty\}) = P(\{\omega : X_1(\omega) \leq x_1\}) = F(x_1). \tag{105}$$

We can write the last equation in terms of PDFs as

$$\lim_{x_2 \to \infty} \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} p(y_1, y_2) dy_1 dy_2 = \int_{-\infty}^{x_1} p(y_1) dy_1. \tag{106}$$

Since x_1 is arbitrary, it follows from (106) that

$$p(x_1) = \int_{-\infty}^{\infty} p(x_1, x_2) dx_2 \qquad \text{(marginalization rule)}. \tag{107}$$

Moreover, we have $F(\infty, \infty) = 1$, i.e.,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x_1, x_2) dx_1 dx_2 = 1 \qquad \text{(normalization condition)}. \tag{108}$$

It is straightforward to extend these formulas to distribution functions and PDFs in more than two variables. For example, if $\boldsymbol{X}(\omega) = (X_1(\omega), X_2(\omega), X_3(\omega), X_4(\omega))$ is a four-dimensional random vector with distribution function $F(x_1, \ldots, x_4)$ and PDF $p(x_1, \ldots, x_4)$, then we can obtain the joint distribution function and the joint PDF of X_2 and X_3, respectively, as

$$F(x_2, x_3) = F(\infty, x_2, x_3, \infty), \qquad p(x_2, x_3) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x_1, x_2, x_3, x_4) dx_1 dx_4. \tag{109}$$

Example (Gaussian distribution): Consider the multivariate Gaussian PDF

$$p(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} e^{-(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})/2}, \tag{110}$$

where

$$\boldsymbol{x}^T = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}, \tag{111}$$

$$\boldsymbol{\mu}^T = \begin{bmatrix} \mathbb{E}\{X_1\} & \dots & \mathbb{E}\{X_n\} \end{bmatrix} \qquad \text{(mean)}, \tag{112}$$

$$\Sigma_{ij} = \mathbb{E}\{X_i X_j\} - \mathbb{E}\{X_i\}\,\mathbb{E}\{X_j\} \qquad \text{(covariance matrix)}. \tag{113}$$

It is straightforward to show that all marginal PDF and distribution functions are still Gaussians of the form (110).

Independence. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Two events $A \in F$ and $B \in \mathcal{F}$ are said to be *independent* if the probability of their intersection (that means the probability that both events A and B happen) equals the product of their probabilities, i.e.,

$$A, B \in \mathcal{F} \quad \text{independent} \quad \Leftrightarrow \quad P(A \cap B) = P(A)P(B). \tag{114}$$

Consider now a random vector $\boldsymbol{X}(\omega) = (X_1(\omega), X_2(\omega))$ with components $X_1(\omega)$ and $X_2(\omega)$. We say that the random variables $X_1(\omega)$ and $X_2(\omega)$ are statistically independent if

$$P(\underbrace{\{\omega : X_1(\omega) \le x_1\}}_{\text{event } A} \cap \underbrace{\{\omega : X_2(\omega) \le x_2\}}_{\text{event } B}) = P(\{\omega : X_1(\omega) \le x_1\})P(\{\omega : X_2(\omega) \le x_2\}), \tag{115}$$

for all $x_1, x_2 \in \mathbb{R}$. Equation (115) can be written in terms of the cumulative distribution function as

$$F(x_1, x_2) = F(x_1)F(x_2). \tag{116}$$

This also implies that the joint PDF of X_1 and X_2 (if it exists) is simply the product of the PDF of X_1 and the PDF of X_2, i.e.,

$$p(x_1, x_2) = p(x_1)p(x_2). \tag{117}$$

These formulas can be generalized to n independent random variables as

$$F(x_1, \dots, x_n) = F(x_1) \cdots F(x_n), \qquad p(x_1, \dots, x_n) = p(x_1) \cdots p(x_n). \tag{118}$$

Examples:

- Jointly uniform random vector. Let $\boldsymbol{X}$ be a n-dimensional random vector with zero-mean i.i.d. (independent identically distributed) uniform components in [-1,1]. The joint PDF of $\boldsymbol{X}$ is

$$p(x_1, \dots, x_n) = \begin{cases} \dfrac{1}{2^n} & (x_1, \dots, x_n) \in [-1, 1]^n \\ 0 & \text{otherwise} \end{cases} \tag{119}$$

- Jointly normal random vector. Let $\boldsymbol{X}$ be a n-dimensional random vector with zero-mean i.i.d. Gaussian components with variance equal to one. The joint PDF of $\boldsymbol{X}$ is

$$p(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2}} e^{-\boldsymbol{x}^T \boldsymbol{x}/2} \qquad \boldsymbol{x} \in \mathbb{R}^n. \tag{120}$$

  Clearly, from equation (110) we see that Gaussian random variables are independent if and only if

$$\mathbb{E}\{X_i X_j\} = \mathbb{E}\{X_i\}\mathbb{E}\{X_j\} \quad \text{for } i \neq j. \tag{121}$$

  In general, if (121) is satisfied then we say that X_i and X_j are *uncorrelated*. Lack of correlation is a much weaker statement than independence, yet sufficient to claim independence for Gaussian random variables.

Conditional distribution function and conditional PDF. Conditional probability is a measure of the probability of an event A occurring, given that another event B has already occurred. Suppose that the two aforementioned events belong to the σ-algebra $\mathcal{F}$ of a probability space $(\Omega, \mathcal{F}, P)$. Then the probability of A under the condition B is defined as[12]

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{122}$$

Note that the conditional probability is non-zero if A and B are intersecting. Also note that if B is a subset of A then $P(A|B) = 1$.

Clearly, if A and B are independent events then by equation (114) we have that $P(A \cap B) = P(A)P(B)$. This implies that if A and B are independent then $P(A|B) = P(A)$. In other words, B has no effect whatsoever on the probability of A occurring. Moreover, $P(A \cap B) \leq P(B)$ and therefore we always have that $P(A|B) \leq 1$.

In the context of random vectors with multiple components, we may be interested in determining the conditional probability of an event involving one component, given that another event involving another component has already occurred. This yields the concept of conditional distribution function and conditional probability density. Let us first clarify these concepts for a random vector with only two components $\boldsymbol{X}(\omega) = (X_1(\omega), X_2(\omega))$. By using the definition of the cumulative distribution function (98) we obtain (see [15, Ch. 7])

$$F(x_1|x_2) = \frac{F(x_1, x_2)}{F(x_2)} \quad \Leftrightarrow \quad F(x_1, x_2) = F(x_1|x_2)F(x_2). \tag{123}$$

The determination of the conditional density of $X_1(\omega)$ assuming $X_2(\omega) = x_2$, i.e., a specific value of $X_2(\omega)$ is of particular interest. This density cannot be derived directly from (122) because, in general, the event $X_2(\omega) = x_2$ has zero probability. However, one can make sense of such conditional probability by taking a suitable limit. Specifically, consider

$$P(\{X_1(\omega) \leq x_1\} \cap \{x_2 < X_2(\omega) \leq x_2 + \Delta x_2\}) = F(x_1, x_2 + \Delta x_2) - F(x_1, x_2) \tag{124}$$

and

$$P(\{x_2 < X_2(\omega) \leq x_2 + \Delta x_2\}) = F(x_2 + \Delta x_2) - F(x_2). \tag{125}$$

In (124) it is understood that $F(x_1, x_2)$ is the joint distribution function of (X_1, X_2), while in (125) $F(x_2)$ denotes the distribution function of X_2 alone. Clearly, for small Δx_2

$$F(x_1, x_2 + \Delta x_2) - F(x_1, x_2) \simeq \Delta x_2 \int_{-\infty}^{x_1} p(y_1, x_2) dy_1, \tag{126}$$

and

$$F(x_2 + \Delta x_2) - F(x_2) \simeq p(x_2)\Delta x_2. \tag{127}$$

By differentiating (123) with respect to x_1, and taking into account (126)-(127) yields the conditional PDF

$$p(x_1|X_2 = x_2) = \frac{\Delta x_2 \dfrac{\partial}{\partial x_1} \displaystyle\int_{-\infty}^{x_1} p(y_1, x_2) dy_1}{\Delta x_2 p(x_2)}, \tag{128}$$

[12] An example of conditional probability could be the following:

- Event A: "Daniele's team scores a goal".
- Event B: "Daniele takes a shot at the goal".

The conditional probability $P(A|B)$, i.e., the probability that Daniele's team scores a goal, conditional to Daniele taking a shot equals the probability that Daniele takes a shot and scores a goal, divided by the probability that Daniele takes a shot.

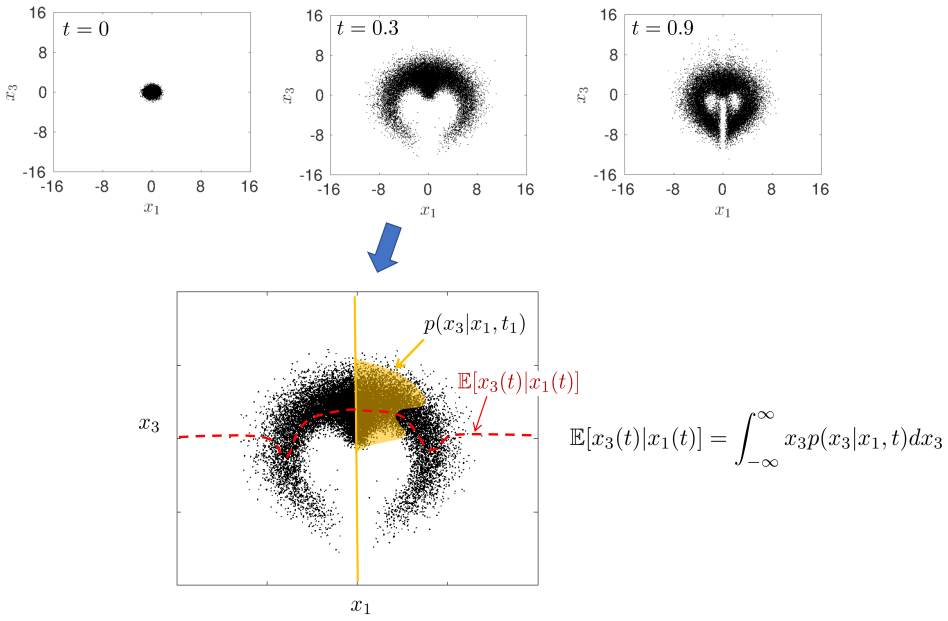

Figure 3: Point clouds representing the joint PDF of the phase variables $x_1(t)$ and $x_3(t)$ of the Kraichnan-Orzag system at different times, i.e., $p(x_3, x_1, t)$. Shown is the procedure to compute the conditional PDF $p(x_3|x_1, t)$ and the corresponding conditional mean $\mathbb{E}\{X_3|X_1 = x_1\}$.

i.e.,

$$p(x_1|X_2 = x_2) = \frac{p(x_1, x_2)}{p(x_2)} \quad \text{(conditional PDF)}. \tag{129}$$

In summary, to compute the conditional PDF, $p(x_1|X_2 = x_2)$ we literally take a section of the joint $p(x_1, x_2)$ for some fixed value of x_2 and then rescale the function we obtain by the number $p(x_2)$, i.e., the one-dimensional PDF of $p(x)$ of $X_2(\omega)$ evaluated at $x = x_2$. This procedure is illustrated in Figure 3 for a PDF represented in terms of a point cloud.

Equation (129) can be written as

$$p(x_1, x_2) = p(x_1|x_2)p(x_2) = p(x_2|x_1)p(x_1) \tag{130}$$

which yields the identities

$$p(x_2) = \int_{-\infty}^{\infty} p(x_2|x_1)p(x_1)dx_1, \qquad p(x_1) = \int_{-\infty}^{\infty} p(x_1|x_2)p(x_2)dx_2. \tag{131}$$

The conditional probability density rule can be generalized to multiple random variables. For instance, if $p(x_1, x_2, x_3, x_4)$ denotes the joint PDF of four random variables then

$$p(x_1, x_2, x_3, x_4) = p(x_1|x_2, x_3, x_4)p(x_2, x_3, x_4) = p(x_1|x_2, x_3, x_4)p(x_2|x_3, x_4)p(x_3|x_4)p(x_4). \tag{132}$$

Moreover, conditional probability densities satisfy the *marginalization rule*. For instance

$$p(x_1, x_3|x_4, x_5) = \int_{-\infty}^{\infty} p(x_1, x_2, x_3|x_4, x_5)dx_2. \tag{133}$$

This property follows directly from the definition of conditional probability density (129).

Expectation, joint moments, and joint cumulants. Let $\boldsymbol{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))$ be a random vector defined on the probability space $(\Omega, \mathcal{F}, P)$. For any measurable function $g(X_1, \ldots, X_n)$ we define the expectation[13] as

$$\mathbb{E}\left\{g(X_1, \ldots, X_n)\right\} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) p(x_1, \ldots, x_n) dx_1 \cdots dx_n. \tag{135}$$

In particular, if $g(X_1, \ldots, X_n) = X_1^{k_1} \cdots X_n^{k_n}$ then

$$\mathbb{E}\left\{X_1^{k_1} \cdots X_n^{k_n}\right\} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1^{k_1} \cdots x_n^{k_n} p(x_1, \ldots, x_n) dx_1 \cdots dx_n \qquad \text{(joint moments)} \tag{136}$$

The correlation matrix[14] and the covariance matrix are defined as (see, e.g., (110))

$$\mathbb{E}\left\{X_i X_j\right\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_i x_j p(x_i, x_j) dx_i dx_j \qquad \text{(correlation matrix)}, \tag{138}$$

$$\mathbb{E}\left\{(X_i - \mu_i)(X_j - \mu_j)\right\} = \mathbb{E}\left\{X_i X_j\right\} - \mu_i \mu_j \qquad \text{(covariance matrix)}. \tag{139}$$

where $\mu_i = \mathbb{E}\left\{X_i\right\}$ (mean of X_i).

Remark: We say that two random variables $X_i(\omega)$ and $X_j(\omega)$ are *uncorrelated* if

$$\mathbb{E}\left\{X_i X_j\right\} = \mathbb{E}\left\{X_i\right\} \mathbb{E}\left\{X_j\right\}. \tag{140}$$

Independent random variables are always uncorrelated. In fact, let $p(x_i, x_j)$ be the joint PDF of X_i and X_j. We know that if X_i and X_j are independent then $p(x_i, x_j)$ can be factorized as

$$p(x_i, x_j) = p(x_i) p(x_j). \tag{141}$$

A substitution of (141) into (138) immediately yields (140).

We define the *moment generating function* of the random vector $\boldsymbol{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))$ as

$$m(a_1, \ldots, a_n) = \mathbb{E}\left\{e^{a_1 X_1 + \cdots + a_n X_n}\right\}. \tag{142}$$

[13] Note that the expectation $\mathbb{E}\{\cdot\}$ is a linear operator from a space of functions, e.g., the space of real-valued functions that are measurable with respect $p(x_1, \ldots, x_n)$. Also, we do not need to assume the existence of the PDF to define the expectation operator. In fact, a more general expression for (135) is

$$\mathbb{E}\left\{g(X_1, \ldots, X_n)\right\} = \int_{\Omega} g(X_1(\omega), \ldots, X_n(\omega)) dP(\omega). \tag{134}$$

[14] Note that (138) follows from (136) using the marginalization property of the PDF. For instance

$$\begin{aligned}
\mathbb{E}\left\{X_1 X_2\right\} &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1 x_2 p(x_1, \ldots, x_n) dx_1 \cdots dx_n \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 \left(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x_1, \ldots, x_n) dx_3 \cdots dx_n \right) dx_1 dx_2 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 p(x_1, x_2) dx_1 dx_2.
\end{aligned} \tag{137}$$

.

It is straightforward to show that

$$\frac{\partial m(0,\dots,0)}{\partial a_i} =\mathbb{E}\{X_i\}, \tag{143}$$

$$\frac{\partial^2 m(0,\dots,0)}{\partial a_j \partial a_i} =\mathbb{E}\{X_i X_j\} \tag{144}$$

$$\frac{\partial^3 m(0,\dots,0)}{\partial a_j \partial a_i \partial a_k} =\mathbb{E}\{X_i X_j X_k\},$$

$$\dots \tag{145}$$

Hence, the partial derivatives of the moment generating function evaluated at zero represent the joint moments of the components of random vector $\boldsymbol{X}$. Clearly, if $m(a_1,\dots,a_n)$ admits a convergent power series expansion at 0 then all joint moments exist.

The *cumulant generating function* of the random vector $\boldsymbol{X}(\omega) = (X_1(\omega),\dots,X_n(\omega))$ is defined as

$$\Psi(a_1,\dots,a_n) = \log(m((a_1,\dots,a_n))). \tag{146}$$

It is straightfoward to show that

$$\frac{\partial \Psi(0,\dots,0)}{\partial a_i} =\mathbb{E}\{X_i\}, \tag{147}$$

$$\frac{\partial^2 \Psi(0,\dots,0)}{\partial a_j \partial a_i} =\mathbb{E}\{X_i X_j\} - \mathbb{E}\{X_i\}\mathbb{E}\{X_j\}, \tag{148}$$

$$\frac{\partial^3 \Psi(0,\dots,0)}{\partial a_j \partial a_i \partial a_k} =\mathbb{E}\{X_i X_j X_k\} - \mathbb{E}\{X_i\}\mathbb{E}\{X_j X_k\} - \mathbb{E}\{X_j\}\mathbb{E}\{X_i X_k\} - \mathbb{E}\{X_k\}\mathbb{E}\{X_i X_j\}$$
$$+ 2\mathbb{E}\{X_i\}\mathbb{E}\{X_j\}\mathbb{E}\{X_k\},$$
$$\dots$$

The quantities at the right hand side are known as *joint cumulants* of the random variables $(X_1,\dots,X_n)$. The cumulants are often denoted as $\langle X_i X_j \cdots \rangle_c$ (see, e.g., [12])

$$\begin{aligned} \langle X_i \cdots \rangle_c =&\mathbb{E}\{X_i\}, \\ \langle X_i X_j \cdots \rangle_c =&\mathbb{E}\{X_i X_j\} - \mathbb{E}\{X_i\}\mathbb{E}\{X_j\}, \\ &\dots. \end{aligned} \tag{149}$$

The *characteristic function* of the random vector $\boldsymbol{X}(\omega) = (X_1(\omega),\dots,X_n(\omega))$ is defined as

$$\phi(a_1,\dots,a_n) = \mathbb{E}\left\{ e^{i(a_1 X_1 + \cdots + a_n X_n)} \right\}. \tag{150}$$

Note that the characteristic function is the Fourier transform of the joint probability density function $p(x_1,\dots,x_n)$ and therefore it essentially carries the same information. The joint moments of $\boldsymbol{X}$ can be computed as

$$\mathbb{E}\left\{ X_1^{k_1} \cdots X_n^{k_n} \right\} = \frac{1}{i^{k_1+\cdots+k_n}} \frac{\partial^{k_1+\cdots+k_n} \phi(0,\dots,0)}{\partial^{k_1} a_1 \cdots \partial^{k_n} a_n}. \tag{151}$$

It is interesting to notice that the marginalization operation we have seen for the PDF, e.g.,

$$p(x_1) = \int_{-\infty}^{\infty} p(x_1, x_2, \dots, x_n) dx_2 \cdots dx_n \tag{152}$$

turns out to be simplified quite substantially in Fourier space. Indeed

$$\phi(a_1) = \phi(a_1, 0, \ldots, 0) = \mathbb{E}\left\{e^{ia_1X_1 + i0X_2 \cdots + i0X_n}\right\}. \tag{153}$$

By using the well known series expansion of the complex exponential, it is possible to show that (see, e.g., [12])

$$\phi(a_1, a_2, \ldots, a_n) = \exp\left[\sum_{\nu_1,\ldots,\nu_n=0}^{\infty} \langle X_1^{\nu_1} \cdots X_n^{\nu_n}\rangle_c \prod_{k=1}^{n} \frac{(ia_k)^{\nu_k}}{\nu_k!}\right] \tag{154}$$

where the series at the exponent excludes the case $\nu_1 = \cdots = \nu_n = 0$. For example,

$$\phi(a_1, a_2) = \phi(a_1)\phi(a_2) \exp\left[\sum_{j,k=1}^{\infty} \left\langle X_1^j X_2^k \right\rangle_c \frac{(ia_1)^j (ia_2)^k}{j!k!}\right], \tag{155}$$

where we used (92). Clearly, if X_1 and X_2 are independent we have $\left\langle X_1^j X_2^k \right\rangle_c = 0$ for all i and j and therefore (155) reduces to

$$\phi(a_1, a_2) = \phi(a_1)\phi(a_2). \tag{156}$$

Clearly, this equation is the Fourier transform of the PDF $p(x_1, x_2) = p(x_1)p(x_2)$, and shows that if X_1 and X_2 are independent both the joint PDF and the joint characteristic function can be factorized as a product of one-dimensional functions.

Conditional expectation. Let $\boldsymbol{X}(\omega)$ and $\boldsymbol{Y}(\omega)$ be two random vectors defined on the probability space $(\Omega, \mathcal{F}, P)$. The *conditional mean* of $u(\boldsymbol{X}(\omega))$ (u is an arbitrary measurable function) assuming $\boldsymbol{Y}(\omega) = \boldsymbol{y}$ is defined as[15]

$$\mathbb{E}\{g(\boldsymbol{X})|\boldsymbol{Y} = \boldsymbol{y}\} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x}, \tag{157}$$

where

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{y})} \tag{158}$$

is the conditional probability density of $\boldsymbol{X}(\omega)$ given $\boldsymbol{Y}(\omega) = \boldsymbol{y}$. Note that the $\mathbb{E}\{g(\boldsymbol{X})|\boldsymbol{Y} = \boldsymbol{y}\}$ is a function of $\boldsymbol{y}$. The conditional mean defined in equation (157) allows us to write the conditional moments of a random variable or a random vector, given information on another random vector. For example, the conditional mean and conditional correlation of $\boldsymbol{X}$ given $\boldsymbol{Y}(\omega) = \boldsymbol{y}$ are defined as

$$\mathbb{E}\{X_i|\boldsymbol{Y} = \boldsymbol{y}\} = \int_{-\infty}^{\infty} x_i p(x_i|\boldsymbol{y}) dx_i, \tag{159}$$

$$\mathbb{E}\{X_i X_j|\boldsymbol{Y} = \boldsymbol{y}\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_i x_j p(x_i, x_j|\boldsymbol{y}) dx_i dx_j. \tag{160}$$

The conditional mean of a system with two random variables is visualized in Figure 3.

By combining (158), (157) and (135) we see that

$$\mathbb{E}\{g(\boldsymbol{X})\} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbb{E}\{g(\boldsymbol{X})|\boldsymbol{Y} = \boldsymbol{y}\} p(\boldsymbol{y}) d\boldsymbol{y}. \tag{161}$$

In this sense, $\mathbb{E}\{g(\boldsymbol{X})|\boldsymbol{Y} = \boldsymbol{y}\}$ can be interpreted as a random variable, i.e., a scalar function of the random variable $\boldsymbol{Y}$ which, if averaged over $p(\boldsymbol{y})$, yuelds exactly $\mathbb{E}\{g(\boldsymbol{X})\}$.

[15] The conditional mean in equation (157) is often written as $\mathbb{E}\{g(\boldsymbol{X})|\boldsymbol{Y}\}$.

Joint PDF of m functions of n random variables. Let $\boldsymbol{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))$ be a random vector with joint probability density function $p(x_1, \ldots, x_n)$. Define

$$\begin{cases} Y_1 = g_1(X_1, \ldots, X_n) \\ \quad \vdots \\ Y_m = g_m(X_1, \ldots, X_n) \end{cases} \tag{162}$$

What is the joint probability density function of the random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$? Note that m can be smaller, equal or larger than n. These cases need to be handled differently.

- If $n = m$ and $\{g_1, \ldots, g_m\}$ are distinct functions we proceed as in Theorem 2 below.
- If $m < n$ and $\{g_1, \ldots, g_m\}$ are distinct functions we can add $m - n$ equations to complement the system so that we have n independent equations in n variables:

$$\begin{cases} Y_1 = g_1(X_1, \ldots, X_n) \\ \quad \vdots \\ Y_m = g_m(X_1, \ldots, X_n) \\ Y_{m+1} = X_{m+1} \\ \quad \vdots \\ Y_n = X_n \end{cases} \tag{163}$$

  Once the joint PDF of $Y_1, \ldots, Y_n$ is known (using Theorem 2 below) then we can marginalize it with respect to $(y_{m+1}, \ldots, y_n)$ to obtain $p(y_1, \ldots, y_m)$ as

$$p(y_1, \ldots, y_m) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(y_1, \ldots, y_m, y_{m+1}, \ldots y_n) dy_{m+1} \cdots dy_n. \tag{164}$$

- If we have more equations than variables (i.e. $m > n$) then the computation of the joint PDF of $(Y_1, \ldots, Y_m)$ is not as straightforward as above. Consider for example the mapping $Y_1(\omega) = X(\omega)$ and $Y_2(\omega) = X^2(\omega)$. Here we have two functions of the same random variable. Note also that $Y_2 = Y_1^2$. It can be shown that the joint PDF of $Y_1 = X$ and $Y_2 = X^2$ is

$$p(y_1, y_2) = p_X(y_1)\delta(y_2 - y_1^2), \tag{165}$$

  where p_X is the PDF of X and $\delta(\cdot)$ is the Dirac delta function.

Theorem 2. Let $\boldsymbol{x}_k(\boldsymbol{y})$ $(k = 1, \ldots, r)$ be the zeros of the nonlinear system of equations $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$ defined in (162) (for $n = m$) or in (163) (for $m < n$). The joint PDF of $Y_1, \ldots, Y_n$ is given by

$$p_{\boldsymbol{Y}}(\boldsymbol{y}) = \sum_{i=1}^{r} \frac{p_{\boldsymbol{X}}(\boldsymbol{x}_i(\boldsymbol{y}))}{|J(\boldsymbol{x}_i(\boldsymbol{y}))|}, \tag{166}$$

where J is the Jacobian determinant[16] associated with the mapping $\boldsymbol{g}(\boldsymbol{x})$ evaluated at $\boldsymbol{x}_i(\boldsymbol{y})$ (assumed non-zero).

[16] In (166) it is assumed that

$$J(\boldsymbol{x}_i(\boldsymbol{y})) = \det \left[\frac{\partial \boldsymbol{g}(\boldsymbol{x})}{\partial \boldsymbol{x}} \right]_{\boldsymbol{x} = \boldsymbol{x}_i(\boldsymbol{y})} \neq 0. \tag{167}$$

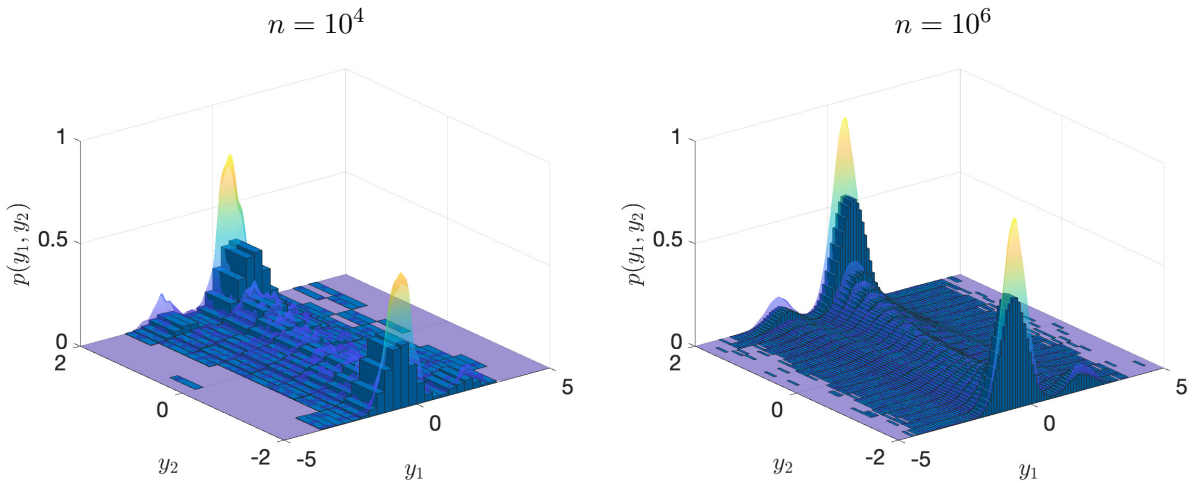

Figure 4: Estimation of the joint PDF of the random variables $Y_1 = X_1$ and $Y_2 = 2\sin(2X_1 + X_2)$ where X_1 and X_2 and independent Gaussians with zero mean and variance one. We show the results we obtain with the frequency approach, i.e., formula (103) and the 2D kernel density estimation method discussed in [4] (transparent surface plot). We plot results for a different number of samples n.

The proof of this theorem is provided in [15, Chapter 8].

Example: Consider the mapping

$$Y_1 = X_1^2 \qquad Y_2 = X_1 + X_2. \tag{168}$$

Suppose we know the joint PDF of X_1 and X_2. What's the joint PDF of Y_1 and Y_2? The following mapping from (X_1, X_2) to (Y_1, Y_2) can be inverted as

$$\begin{cases} y_1 = x_1^2 \\ y_2 = x_1 + x_2 \end{cases} \quad \Rightarrow \quad \begin{cases} x_1 = \pm\sqrt{y_1} \\ x_2 = y_2 \mp \sqrt{y_1} \end{cases} . \tag{169}$$

The Jacobian determinant of (169) is easily obtained as

$$J(x_1, x_2) = \det \begin{bmatrix} 2x_1 & 0 \\ 1 & 1 \end{bmatrix} = 2x_1. \tag{170}$$

Hence, by applying Theorem 2, we obtain the following joint PDF of Y_1 and Y_2 is

$$p_{\boldsymbol{Y}}(y_1, y_2) = \frac{1}{2\sqrt{y}} \left[p_{\boldsymbol{X}}(\sqrt{y_1}, y_2 - \sqrt{y_1}) + p_{\boldsymbol{X}}(-\sqrt{y_1}, y_2 + \sqrt{y_1}) \right] \qquad y_1 \geq 0. \tag{171}$$

Example: Consider the mapping

$$Y_1(\omega) = X_1 \qquad Y_2(\omega) = 2\sin\left(2X_1(\omega) + X_2(\omega)\right), \tag{172}$$

where X_1 and X_2 and independent Gaussians with zero mean and variance one. In Figure 4 we estimate the joint PDF of Y_1 and Y_2 using the frequency approach approach, i.e., formula (103), and the 2D kernel density estimation method discussed in [4].

Alternative methods to compute the joint PDF of functions of random vectors. There are alternative equivalent methods to compute the joint PDF $(Y_1, \ldots, Y_m)$, given the joint PDF $(Y_1, \ldots, Y_n)$,

e.g., methods based on the Dirac delta function [10] or methods based on the joint characteristic function. With reference to the previous example we have the joint characteristic function

$$\phi_{\boldsymbol{Y}}(a_1,a_2)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{ia_1x_1^2+ia_2(x_1+x_2)}p(x_1,x_2)dx_1dx_2. \tag{173}$$

Clearly, if $\phi_{\boldsymbol{Y}}(a_1,a_2)$ can be computed then we can simply inverse Fourier transform it to obtain the joint PDF of (Y_1,Y_2). By using Dirac delta functions we can represent directly the joint PDF of the random variable

$$Y(\omega)=g(X_1(\omega),\ldots,X_n(\omega)), \tag{174}$$

as

$$\begin{aligned}p(y)=&\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\delta(y-g(x_1,\ldots,x_n))p(x_1,\ldots,x_n)dx_1\cdots dx_n &(175)\\ =&\frac{1}{2\pi}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}e^{ia(y-g(x_1,\ldots,x_n))}p(x_1,\ldots,x_n)dx_1\cdots dx_nda. &(176)\end{aligned}$$

Example: Let $Y_1=X$ and $Y_2=X^2$ (two functions of one random variable). What is the joint PDF of Y_1 and Y_2? The mapping (162) yields a Jacobian determinant that is zero, and therefore the mapping it is not invertible. This implies that theorem (2) cannot be applied. However, using the characteristic function approach we obtain

$$\phi(a_1,a_2)=\int_{-\infty}^{\infty}e^{ia_1x+ia_2x^2}p_X(x)dx. \tag{177}$$

Taking the inverse Fourier transform yields,

$$\begin{aligned}p(y_1,y_2)=&\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{ia_1(x-y_1)+ia_2(x^2-y_2)}p_X(x)dxda_1da_2\\ =&\frac{1}{2\pi}\int_{-\infty}^{\infty}\delta(x-y_1)e^{ia_1(x^2-y_2)}p_X(x)dxda_2\\ =&\delta(y_1^2-y_2)p_X(y_1). &(178)\end{aligned}$$

Remark: If $(X_1,\ldots,X_n)$ are independent random variables and $(g_1,\ldots,g_n)$ are n functions from $\mathbb{R}$ into $\mathbb{R}$, then $Y_1=g_1(X_1)$, ..., $Y_n=g_n(X_n)$ are independent random variables. It is straightforward to prove this statement using the Dirac delta function representation (or the characteristic function) of PDF mapping [10]. To this end, let

$$Y_i(\omega)=g_i(X_i(\omega)). \tag{179}$$

We have

$$\begin{aligned}p(y_1,\ldots,y_n)=&\int_{-\infty}^{\infty}\prod_{j=1}^{n}\delta(y_j-g_j(x_j))p(x_1,\ldots,x_n)dx_1\cdots dx_n\\ =&\prod_{j=1}^{n}\int_{-\infty}^{\infty}\delta(y_j-g_j(x_j))p(x_j)dx_j\\ =&p(y_1)\cdots p(y_n). &(180)\end{aligned}$$

Remark: The PDF of the sum of independent random variables is the *convolution* the PDF of each variable. For example, let

$$Y = X_1 + X_2 + X_3 \tag{181}$$

be the sum of three independent random variables X_1, X_2 and X_3, with PDFs $p_1(x_1)$, $p_2(x_2)$ and $p_3(x_3)$ respectively. By using (175) we obtain

$$\begin{aligned} p(y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta(y - x_1 - x_2 - x_3) p(x_1, x_2, x_3) dx_1 dx_2 dx_3 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta(x_1 - y + x_2 + x_3) p_1(x_1) p_2(x_2) p_3(x_3) dx_1 dx_2 dx_3 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_1(x_2 + x_3 - y) p_2(x_2) p_3(x_3) dx_2 dx_3 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_1(x_1 - y) p_2(x_1 - x_3) p_3(x_3) dx_1 dx_3. \end{aligned} \tag{182}$$

In the last equality we considered the mapping $x_1 = x_2 + x_3$ as a coordinate change from x_1 to x_2 with parameter x_3. Note that the process of computing the PDF of the sum of independent random variables can be also seen as a hierarchical process in which we proceed with two variables at a time To this end, we first compute the PDF of $Z = X_2 + X_3$ as

$$p_Z(z) = \int_{-\infty}^{\infty} p_2(z - x_3) p_3(x_3) dx_3. \tag{183}$$

Clearly, Z is independent of X_1 and therefore the PDF of $Y = Z + X_1$ is

$$p_Y(y) = \int_{-\infty}^{\infty} p_1(y - x_1) p_Z(x_1) dx_1. \tag{184}$$

A substitution of (183) into (184) yields (182).

Lebesgue spaces of random variables. The expectation operator $\mathbb{E}\{\cdot\}$ is a linear integral operator over a probability measure. Such an operator can be used to define norms and eventually inner products in spaces of random variables. For example,

$$\mathbb{E}\left\{|X|^q\right\} = \int_\Omega |X(\omega)|^q dP(\omega) \qquad q \in \mathbb{N} \tag{185}$$

is essentially a weighted q norm. The space of random variables satisfying $\mathbb{E}\left\{|X|^q\right\} < \infty$ is denoted as $L^q(\Omega, \mathcal{F}, P)$, in analogy with the classical Lebesgue space for functions. The case $q = 2$ is of particular importance as it has the structure of a Hilbert space. Specifically, for any two random variables in $L^2(\Omega, \mathcal{F}, P)$ we have the inner product

$$\mathbb{E}\left\{XY\right\} = \int_\Omega X(\omega) Y(\omega) dP(\omega) \tag{186}$$

and the norm

$$\mathbb{E}\left\{X^2\right\} = \int_\Omega X(\omega)^2 dP(\omega). \tag{187}$$

The inner product (186) allows us to define orthogonal random variables. Specifically, $X(\omega)$ and $Y(\omega)$ are orthogonal in $L^2(\Omega, \mathcal{F}, P)$ if they are uncorrelated, i.e., $\mathbb{E}\left\{XY\right\} = 0$. Also, $X(\omega)$ and $Y(\omega)$ are orthonormal if they are orthogonal and have norm equal to one, i.e., $\mathbb{E}\left\{X^2\right\} = \mathbb{E}\left\{Y^2\right\} = 1$.

Application to dynamical systems

Consider the following linear dynamical system

$$\begin{cases} \dot{x}(t) + \xi(\omega)x(t) = 0 \\ x(0) = x_0(\omega) \end{cases} \tag{188}$$

where $\xi(\omega)$ and $x_0(\omega)$ are independent random variables. Specifically $\xi(\omega)$ is uniformly distributed in $[0,1]$, while $x_0(\omega)$ is Gaussian random variable with mean zero and variance one. As is well-known, the analytical solution of (188) is

$$x(t;\omega) = x_0(\omega)e^{-t\xi(\omega)}. \tag{189}$$

Let us compute the mean, the second-order moment and the auto-correlation function of the solution $x(t;\omega)$, i.e., $\mathbb{E}\{x(t;\omega)\}$, $\mathbb{E}\{x(t;\omega)^2\}$, and $\mathbb{E}\{x(t;\omega)x(t'\omega)\}$ versus time. We have

$$\mathbb{E}\left\{x(t;\omega)\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x_0 e^{-x_0^2/2}dx_0 \int_0^1 e^{-t\xi}d\xi = 0, \tag{190}$$

$$\mathbb{E}\left\{x(t;\omega)^2\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x_0^2 e^{-x_0^2/2}dx_0 \int_0^1 e^{-2t\xi}d\xi = \frac{1}{2t}\left(1-e^{-2t}\right), \tag{191}$$

$$\mathbb{E}\left\{x(t;\omega)x(t';\omega)\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x_0^2 e^{-x_0^2/2}dx_0 \int_0^1 e^{-(t+t')\xi}d\xi = \frac{1}{t+t'}\left(1-e^{-(t+t')}\right). \tag{192}$$

The one-time probability density function of $x(t;\omega)$ can be easily computed by using the Dirac delta function approach [10]. Indeed,

$$\begin{aligned} p(x,t) =& \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\int_0^1 \delta\left(x - x_0e^{-\xi t}\right)e^{-x_0^2/2}dx_0d\xi \\ =& \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\int_0^1 \frac{\delta\left(x_0 - xe^{\xi t}\right)}{e^{-\xi t}}e^{-x_0^2/2}dx_0d\xi \qquad (193) \\ =& \frac{1}{\sqrt{2\pi}}\int_0^1 e^{\xi t-(xe^{\xi t})^2/2}d\xi. \qquad (194) \end{aligned}$$

Now consider the change of variables

$$u = \frac{xe^{\xi t}}{\sqrt{2}} \quad \Rightarrow \quad d\xi = \frac{\sqrt{2}}{\xi t}e^{-\xi t}du. \tag{195}$$

A substitution of (195) into (194) yields

$$\begin{aligned} p(x,t) =& \frac{1}{\xi t\sqrt{\pi}}\int_{x/\sqrt{2}}^{xe^t/\sqrt{2}} e^{-u^2}du \qquad (196) \\ =& \frac{1}{\xi t\sqrt{\pi}}\left[\text{erf}\left(\frac{xe^t}{\sqrt{2}}\right) - \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right]. \qquad (197) \end{aligned}$$

Liouville equation approach: We can transform the linear system (188) involving one random variable at the right hand side to an equivalent 2D linear system evolving from a random initial state (an no random variables at the right hand side). To this end, we notice that

$$\begin{cases} \dot{x}(t) + yx(t) = 0 \\ \dot{y}(t) = 0 \\ x(0) = x_0(\omega) \\ y(0) = \xi(\omega) \end{cases} \tag{198}$$

is completely equivalent to (188). In this setting, we can derive a linear transport equation for the joint PDF of $x(t;\omega)$ and $y(t;\omega)$, i.e., $x(t;\omega)$ and $\xi(\omega)$. Such PDF equation takes the form

$$\begin{cases} \dfrac{\partial p(x,y,t)}{\partial t} + \dfrac{\partial}{\partial x}\left(xyp(x,y,t)\right) + \dfrac{\partial}{\partial y}\left(xyp(x,y,t)\right) = 0 \\ p(x,y,0) = p_{x_0}(x)p_{\xi}(y) \end{cases} \tag{199}$$

It can be verified by a direct substitution that the solution the initial value problem (199) is

$$p(x,y,t) = \frac{1}{\sqrt{2\pi}} e^{yt-(xe^{yt})^2/2}. \qquad y \in [0,1], \qquad x \in \mathbb{R}. \tag{200}$$

Note that the joint PDF (200) was already obtained in equation (194), right before marginalizing with respect to ξ.

Data-driven identification of random dynamical systems. A system with random parameters and/or random initial states generates an envelope of trajectories that depends on the joint PDF of the random variables driving the system. It is possible to identify such joint PDF from data, e.g., by minimizing a performance metric, i.e., a dissimilarity measure (e.g., a Wasserstein norm) between the measurements of a quantity of interest at various times and the envelope of trajectories, over the degrees of freedom representing the joint probability density function. of the random variables. In this way, we are essentially trying to reduce *model uncertainty* by shrinking a continuous trajectory tube generated by a random dynamical system of the form

$$\begin{cases} \dfrac{d\boldsymbol{x}(t)}{dt} = \boldsymbol{f}(\boldsymbol{x}(t);\omega) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0(\omega) \end{cases} \tag{201}$$

around measurements of some phase space function $\boldsymbol{h}(\boldsymbol{x}(t))$. Note that $\boldsymbol{f}(\boldsymbol{x}(t);\omega)$ is a *random vector field* and $\boldsymbol{x}_0(\omega)$ is a random initial state. We will shall see hereafter that $\boldsymbol{f}(\boldsymbol{x}(t);\omega)$ can be represented in a Karhunen-Loève expansion

$$f_i(\boldsymbol{x}(t);\omega) = \sum_{k=1}^{\infty} \xi_i^k(\omega)\theta_i^k(\boldsymbol{x}) \qquad i = 1,\ldots,n \tag{202}$$

where $\xi_i^k(\omega)$ are uncorrelated random variables and $\theta_i^k(\boldsymbol{x})$ are orthonormal basis functions. Other representations of $f_i(\boldsymbol{x}(t);\boldsymbol{\xi}(\omega))$ can be built, e.g., using tensor expansions in weighted L^2 spaces, e.g., functional tensor train [7, 3, 14]. The minimization procedure discussed above essentially identifies the degrees of freedom of the joint probability density function of $\boldsymbol{x}_0(\omega)$ and $\boldsymbol{\xi}(\omega)$, i.e., $p(\boldsymbol{x}_0,\boldsymbol{\xi})$, either in the form of a *sampler*, e.g., using Wasserstein generative neural networks [1], or the actual multivariate function $p(\boldsymbol{x}_0,\boldsymbol{\xi})$.

Random processes and random fields

Let $\Omega,\mathcal{F},P)$ be a probability space. A real valued stochastic process in the time interval $[0,T]$ is a mapping

$$X : \Omega \times [0,T] \to \mathbb{R}. \tag{203}$$

The process can be continuous in time (e.g., Brownian motion) discontinuous in time (e.g., telegrapher's random process), or time-discrete, e.g., represented by a sequence of random variables $X(t_j;\omega)$ $j = 1,\ldots,n$.

Remark: The notion of continuity we know for real valued functions can be generalized substantially when dealing with stochastic processes. We have, for example,

- Continuity in probability:

$$\lim_{s\to t} P(\{\omega : |X(t;\omega) - X(s;\omega)| > \epsilon\}) = 0 \quad \text{for all } \epsilon > 0. \tag{204}$$

- Mean-square continuity:

$$\lim_{s\to t} \mathbb{E}\{|X(t;\omega) - X(s;\omega)|^2\} = 0. \tag{205}$$

- Continuity in distribution:

$$\lim_{s\to t} F(x,s) = F(x,t) \qquad (F(x,t) \text{ distribution function of } X(t;\omega). \tag{206}$$

$$\text{Continuity in mean-square} \quad \Rightarrow \quad \text{continuity in probability} \quad \Rightarrow \quad \text{continuity in distribution.}$$

Continuity in probability follows from mean-square continuity[17] thanks to the Markov's inequality

$$P(\{\omega : |X(t;\omega) - X(s;\omega)| > \epsilon\}) \leq \frac{1}{\epsilon^2}\mathbb{E}\{|X(t;\omega) - X(s;\omega)|^2\} \qquad \forall t,s \in [0,T]. \tag{208}$$

Other properties of $X(t;\omega)$ very much depend on the way we characterize the process, i.e., the set of rules and specifications that allow us to fully characterize the process. Clearly, $X(t;\omega)$ is a random variable for each fixed t. This means that $X(t;\omega)$ admits a distribution function

$$F(x,t) = P\left(\{\omega : X(t;\omega) \leq x\}\right), \tag{209}$$

and eventually a probability density function

$$p(x,t) = \frac{dF(x,t)}{dx}. \tag{210}$$

With $F(x,t)$ or $p(x,t)$ available we can compute the statistical moments at time t, e.g.,

$$\mathbb{E}\left\{X(t;\omega)^k\right\} = \int_{-\infty}^{\infty} x^k p(x,t)dx, \quad k \in \mathbb{N}. \tag{211}$$

The PDF $p(x,t)$, however, provides very limited statistical information about the process $X(t;\omega)$. In fact, it does not allow us to compute any joint statistics at different times, for example the autocorrelation function

$$\mathbb{E}\{X(t;\omega)X(s;\omega)\} = \int_{-\infty}^{\infty} x_1 x_2 p(x_1,x_2,t_1,t_2)dx_1dx_2, \tag{212}$$

where $p(x_1,x_2,t_1,t_2)$ is the joint probability density function of the random variables $X(t_1;\omega)$ and $X(t_2;\omega)$ (t_1 and t_2 here can vary in $[0,T]$). A straightforward generalization of this line of thinking leads us to construct the joint PDF of $\{X(t_1;\omega),\ldots,X(t_n;\omega)\}$ for an increasing number of distinct time instants

[17] Mean square continuity also implies that the mean process $\mathbb{E}\{X(t;\omega)\}$ is continuous in t. In fact, using the inequality $|\mathbb{E}\{X\}|^2 \leq \mathbb{E}\{X^2\}$ we obtain

$$|\mathbb{E}\{X(t;\omega) - X(s;\omega)\}| \leq \sqrt{\mathbb{E}\{|X(t;\omega) - X(s;\omega)|^2\}}. \tag{207}$$

Similarly, if the process is mean-square continuous then the auto-correlation function $\mathbb{E}\{X(t;\omega)X(s;\omega)\}$ continuous in both s and t.

$t_i \in [0,T]$. Similarly, we can construct the joint characteristic function of the random process $X(t;\omega)$ at distinct time instants $(t_1,\ldots,t_n)$ as

$$\phi(a_1,\ldots,a_n;t_1,\ldots,t_n) = \mathbb{E}\left\{e^{ia_1X(t_1;\omega)+\cdots+ia_nX(t_n;\omega)}\right\}. \tag{213}$$

This expression can be obtained (at least formally) from the so-called Hopf characteristic functional [20, 11] associated with the stochastic process $X(t;\omega)$, i.e.,

$$\Phi([\theta(t)]) = \mathbb{E}\left\{\exp\left(\int_0^T X(\tau;\omega)\theta(\tau)d\tau\right)\right\}, \tag{214}$$

where $\theta(t)$ is a deterministic test function which we are free to choose. For example, if we pick

$$\theta(t) = \sum_{i=1}^n a_i\delta(t-t_i), \tag{215}$$

and substitute it into (214) then we obtain (213). The Hopf functional[18] (214) provides full statistical information about the stochastic process $X(t;\omega)$, including all joint statistical moments, all multi-time PDFs, etc. For instance, the functional derivatives of Φ evaluated at $\theta = 0$ coincide with the statistical moments (see, e.g. [18])

$$\left.\frac{\delta^q\Phi([\theta])}{\delta\theta(t)^q}\right|_{\theta=0} = \frac{1}{i^q}\mathbb{E}\{X(t;\omega)^q\}, \qquad \left.\frac{\delta^{q+p}\Phi([\theta])}{\delta\theta(t)^q\delta\theta(s)^p}\right|_{\theta=0} = \frac{1}{i^{q+p}}\mathbb{E}\{X(t;\omega)^qX(s;\omega)^p\}. \tag{216}$$

In [11] the Hopf functional is determined for various types of stochastic processes.

Remark: To fully characterize a stochastic process it is not necessary to identify or provide the Hopf functional. A stochastic process can be defined in many different ways, some of which are not even explicit. However, if the Hopf characteristic functional is available, then the process is fully specified, perhaps in the most compact possible way (see [13] for applications of Hopf functional methods to turbulence).

Gaussian processes. The Hopf characteristic functional for a Gaussian process is (see, e.g., [11])

$$\Phi([\theta(t)]) = \mathbb{E}\left\{\exp\left(i\int_0^T \mu(\tau)\theta(\tau) - \int_0^T\int_0^T C(\tau,s)\theta(\tau)\theta(s)d\tau ds\right)\right\}, \tag{217}$$

where

$$\mu(t) = \mathbb{E}\{X(t;\omega)\} \qquad \text{(mean)}, \tag{218}$$

$$C(t,s) = \mathbb{E}\{X(t;\omega)X(s;\omega)\} - \mu(t)\mu(s) \qquad \text{(covariance function)}. \tag{219}$$

Higher order moments can be computed using functional differentiation (e.g., (216)), or by noticing that the joint characteristic function the random process $X(t;\omega)$ at an arbitrary number of distinct time instants is

$$\phi(a_1,\ldots,a_n;t_1,\ldots,t_n) = \exp\left(i\sum_{k=1}^n a_k\mu(t_k) - \sum_{k,j=1}^n C(t_k,t_j)a_ka_j\right). \tag{220}$$

[18] Recall that a functional is a mapping from a certain space of functions (or distributions) into the real line or the complex plane. The Hopf functional is a complex-valued nonlinear functional into $\mathbb{C}$.

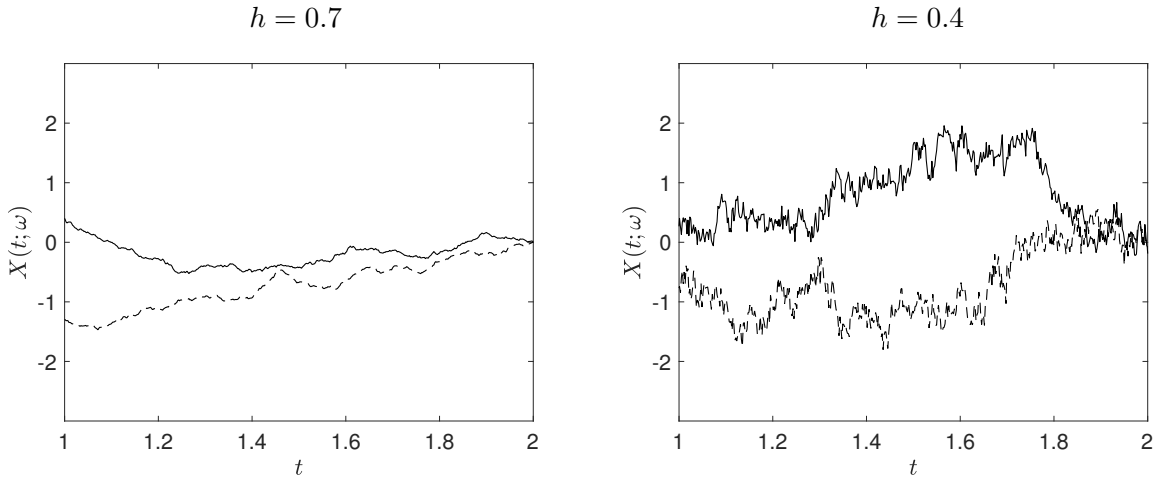

Figure 5: Samples of zero-mean Gaussian process with covariance function (222) and $\sigma = 1$. We show samples corresponding to different values of the Hurst parameter h.

Sampling Gaussian processes: To sample a Gaussian process with mean $\mu(t)$ and covariance function $C(s,t)$ it is sufficient to construct a temporal grid in $[0,T]$ and then sample a Gaussian random vector with mean $\mu_i = \mu(t_i)$, and covariance matrix with entries $C(t_k,t_j)$. To this end, it is sufficient to recall that if $\boldsymbol{X}(\omega)$ is a zero-mean Gaussian random vector (column vector) with independent entries of variance one, and $\boldsymbol{C} = \boldsymbol{R}\boldsymbol{R}^T$ is the Cholesky decomposition of the covariance matrix[19] $\boldsymbol{C}$, then $\boldsymbol{Y} = \boldsymbol{R}\boldsymbol{X}$ is a zero-mean Gaussian random vector with covariance $\boldsymbol{C}$. In fact,

$$\mathbb{E}\{\boldsymbol{Y}(\omega)\boldsymbol{Y}^T(\omega)\} = \mathbb{E}\{\boldsymbol{R}\boldsymbol{X}(\omega)\boldsymbol{X}^T(\omega)\boldsymbol{R}^T\} = \boldsymbol{R}\underbrace{\mathbb{E}\{\boldsymbol{X}(\omega)\boldsymbol{X}^T(\omega)\}}_{\text{(identity matrix)}}\boldsymbol{R}^T = \boldsymbol{C}. \tag{221}$$

In figure 5 we plot a few samples of a Gaussian random process with zero mean and covariance function

$$C(s,t) = \frac{\sigma}{2}\left(|s|^{2h} + |t|^{2h} - |s-t|^{2h}\right), \tag{222}$$

where $0 < h < 1$ is the so-called Hurst parameter. A Gaussian process with covariance function (222) is called fractional Brownian motion.

Gaussian random fields: The procedure we used to sample of Gaussian stochastic process with covariance $C(s,t)$ (e.g., (222)) can be extended to *Gaussian random fields* [17], i.e., random functions defined of a domain $V \subseteq \mathbb{R}^d$. For example, we could sample a zero-mean Gaussian random field $X(\boldsymbol{x};\omega)$ defined on the square domain $V = [0,1]\times[0,1]$ with covariance function

$$C(\boldsymbol{x},\boldsymbol{y}) = \frac{\sigma}{2}\left(\|\boldsymbol{x}\|_2^{2h} + \|\boldsymbol{y}\|_2^{2h} - \|\boldsymbol{x}-\boldsymbol{y}\|_2^{2h}\right), \tag{223}$$

to this end we first construct the covariance matrix $C(\boldsymbol{x}_i,\boldsymbol{x}_j)$ and then use the procedure we used before, i.e.: i) sample a zero-mean i.i.d. Gaussian random variable with variance one at each spatial location $\boldsymbol{x}_i$, and ii) multiply the sample of the random vector constructed in this way by the matrix $\boldsymbol{R}$ obtained by the Cholesky decomposition of the autocovariance function (223). In Figure 6 we provide a few samples of a zero mean Gaussian random field with covariance (223).

[19] The entries of the covariance matrix $\boldsymbol{C}$ are $C(t_i,t_j)$, where $C(t,s)$ is the covariance function of the random process.

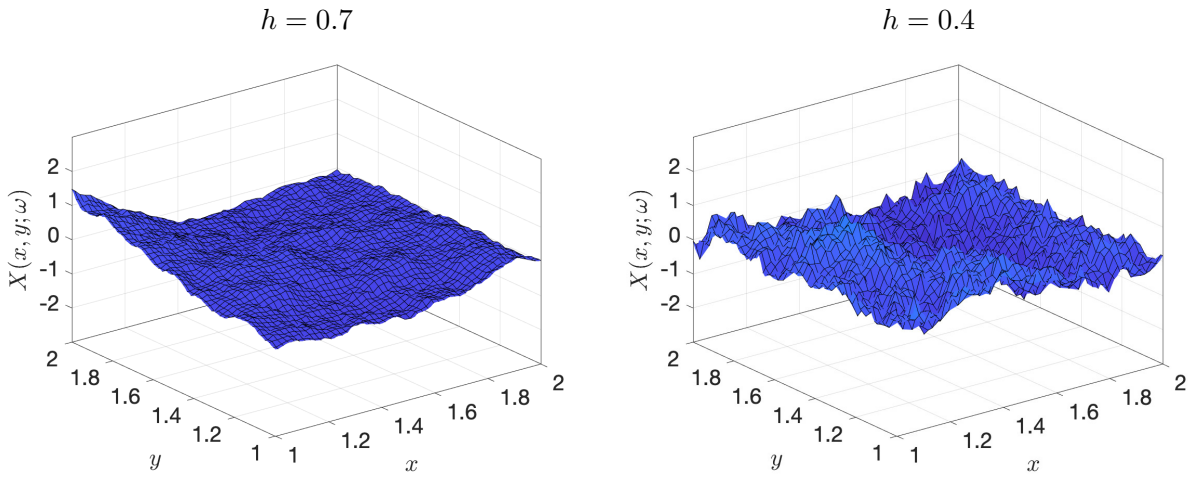

Figure 6: Samples of zero-mean Gaussian random field with covariance function (223) and $\sigma = 1$. We show samples corresponding to different values of the Hurst parameter h.

Discrete Markov processes. Consider a discrete set of distinct temporal time instant, say $\{t_1, \ldots, t_n\}$ and a time-discrete random process which is essentially a collection of random variables

$$X_i(\omega) = X(t_i; \omega), \tag{224}$$

or a collection of random vectors

$$\boldsymbol{X}_i(\omega) = \boldsymbol{X}(t_i; \omega). \tag{225}$$

The random process (224) can be defined in many different ways, for example as a recurrence relation[20]

$$X_{i+1}(\omega) = h(X_i(\omega)) + \xi_i(\omega), \tag{226}$$

where $\xi_i(\omega)$ are random variables and $X_0(\omega)$ is random as well. Similarly, we can define a vector valued discrete process as

$$\boldsymbol{X}_{i+1}(\omega) = \boldsymbol{h}(\boldsymbol{X}_i(\omega)) + \boldsymbol{\xi}_i(\omega), \tag{227}$$

Note that the structure of (227) is the same as a recurrent neural network perturbed by noise [21].

Disregarding how we generate the sequence of random variables $X_0, \ldots, X_n$, in (226), we can characterize the statistics of the process X_i in terms of the joint PDF (assuming it exists) $p(x_n, \ldots, x_0)$. By using the definition of conditional probability density we have

$$p(x_n, \ldots, x_1, x_0) = p(x_n | x_{n-1}, \ldots, x_1, x_0) p(x_{n-1}, \ldots, x_0). \tag{228}$$

If the system is memoryless (or Markovian), we have that the conditional PDF of X_n given the entire history of X_i equals $p(x_n|x_{n-1})$, i.e.,

$$p(x_n | x_{n-1}, \ldots, x_1, x_0) = p(x_n | x_{n-1}). \tag{229}$$

In other words, the PDF of $X_n(\omega)$ conditional to any set of variables $\{X_j(\omega)\}$ with $j < n$ equals to $p(x_n|x_{n-1})$, i.e., it depends only on the value of $X_{n-1}(\omega)$. By applying (229) recursively we obtain

$$p(x_n, x_{n-1}, \ldots, x_1, x_0) = p(x_n|x_{n-1}) p(x_{n-1}|x_{n-2}) \cdots p(x_1|x_0) p(x_0). \tag{230}$$

[20] The discrete process (226) is also called *autoregressive process*.

Hence, the process is fully specified by the transition density $p(x_{k+1}|x_k)$. Denoting by p_{ξ_k} the PDF of ξ_k in (226), and assuming that $\{\xi_1, \ldots, \xi_{n-1}\}$ are statistically independent we have that the transition probability defined by the Markov chain (226) is

$$p(x_{k+1}|x_k) = p_{\xi_k}(x_{x+1} - F(x_k)). \tag{231}$$

Remark: More general auto-regressive random vector processes of the form (227) are the discussed in the book [5]. For example, vector auto-regressive moving-average (VARMA) processes, integrated VARMA (VARIMA) processes, etc.

Markov Chain Monte Carlo (MCMC). Markov Chain Monte Carlo (MCMC) refers to a class of methods that allow us to sample high-dimensional probability density functions [6]. In MCMC we construct a discrete Markov process that has a stationary PDF that coincides with the distribution of interest, i.e., the PDF we'd like to sample from. Hence, simulations of the Markov chain[21] provide samples of the high-dimensional PDF we are interested in, once a transient, i.e., the so-called *burn-in* phase of the chain, is completed. There are several MCMC algorithms to sample from high-dimensional PDFs. Perhaps the simplest ones are the Gibbs sampling and the Metropolis-Hastings algorithms. Let us briefly describe the Gibbs sampling method. To this end, suppose you are given a three-dimensional PDF $p(x_1, x_2, x_3)$ and that the conditional PDFs $p(x_1|x_2, x_3)$, $p(x_2|x_1, x_3)$ and $p(x_3|x_1, x_2)$ are all available[22]. To sample from $p(x_1, x_2, x_3)$ we proceed as follows:

1. Initialize $x_2 = x_2^{(i)}$ and $x_3 = x_3^{(i)}$. Here $x_2^{(i)}$ and $x_3^{(i)}$ are two real numbers. The superscript "i" is an integer number that labels the discrete Markov process

$$\boldsymbol{X}_i(\omega) = \begin{bmatrix} x_1^{(i)}(\omega) & x_2^{(i)}(\omega) & x_3^{(i)}(\omega) \end{bmatrix} \qquad i \in \mathbb{N}. \tag{232}$$

2. Sample a new $x_1^{(i+1)}$ from the one-dimensional conditional PDF $p\left(x_1|x_2^{(i)}, x_3^{(i)}\right)$.

3. With the sample $x_1^{(i+1)}$ available, sample a new $x_2^{(i+1)}$ from the one-dimensional conditional PDF $p\left(x_2|x_1^{(i+1)}, x_3^{(i)}\right)$.

4. With the sample $x_2^{(i+1)}$ available, sample a new $x_3^{(i+1)}$ from the one-dimensional conditional PDF $p\left(x_3|x_1^{(i+1)}, x_2^{(i+1)}\right)$.

5. Update $x_j^{(i)} \leftarrow x_j^{(i+1)}$ for $j = 1, 2, 3$ and go back to point 2.

This algorithm allows us to compute $\boldsymbol{X}_{i+1}$ from $\boldsymbol{X}_i$ by sampling known one-dimensional conditional transition densities. To sample from such arbitrary one-dimensional transition densities we can use different methods. If the inverse cumulative distribution of each conditional PDF is known, then we have seen that it is sufficient to sample a uniform PDF and then map such sample using the inverse cumulative distribution function. Alternatively, we can determine the mapping between uniform random variables and conditionally distributed random variables using *polynomial chaos expansions*. The mapping $\boldsymbol{X}_i \to \boldsymbol{X}_{i+1}$ defines a *random walk* in $\mathbb{R}^3$. The stationary distribution of such random walk coincides with $p(x_1, x_2, x_3)$.

[21] Simulations of a Markov chain are usually performed with the Monte Carlo method, hence the name Markov Chain Monte Carlo.

[22] Recall that to compute the conditional PDF $p(x_1|x_2, x_3)$ we literally set x_2 and x_3 in $p(x_1, x_2, x_3)$ equal to some number, say $x_2 = x_2^*$ and $x_3 = x_3^*$ and then normalize the one-dimensional function $p(x_1, x_2^*, x_3^*)$ so that the integral with respect to x_1 equals one.

In other words, after the burn-in phase is completed, i.e., for sufficiently large i, we have that $X_i(\omega)$ are samples of the joint PDF $p(x_1, x_2, x_3)$.

Karhunen-Loève expansion. Let $X(t;\omega)$ be a zero-mean square-integrable stochastic process defined on the probability space $(\Omega, \mathcal{F}, P)$. "Square-integrable" means that $X(t;\omega)$ has finite second order moment, i.e.,

$$\mathbb{E}\left\{\int_0^T X(t;\omega)^2 dt\right\} < \infty. \tag{233}$$

By using the properties of $L^2(\Omega, \mathcal{F}, P)$ spaces (probability spaces of square integrable random variables), it can be shown that $X(t;\omega)$ admits a series expansion

$$X(t;\omega) = \sum_{k=1}^{\infty} \sqrt{\lambda_k}\xi_k(\omega)\psi_k(t), \tag{234}$$

where $\{\xi_1(\omega), \xi_2(\omega), \ldots\}$ is a set of uncorrelated, i.e., orthonormal, random variables satisfying

$$\mathbb{E}\left\{\xi_i(\omega)\xi_j(\omega)\right\} = \delta_{ij}, \tag{235}$$

and $\{\psi_1(t), \psi_2(t), \ldots\}$ are orthonormal (in $L^2([0,T])$) temporal modes

$$\int_0^T \psi_i(t)\psi_j(t)dt = \delta_{ij}. \tag{236}$$

By using the orthogonality properties (235)-(236), we obtain the so-called dispersion relations[23]

$$\xi_k(\omega) = \frac{1}{\sqrt{\lambda_k}} \int_0^T X(t;\omega)\psi_k(t)dt, \tag{238}$$

$$\psi_k(t) = \frac{1}{\sqrt{\lambda_k}} \mathbb{E}\{\xi_k(\omega)X(t;\omega)\}. \tag{239}$$

A substitution of (238) into (239) yields the eigenvalue problem

$$\tag{240}$$

$$\int_0^T C(t,s)\psi_k(s)ds = \lambda_k^2\psi_k(t). \tag{241}$$

where

$$C(t,s) = \mathbb{E}\left\{X(t;\omega)X(s;\omega)\right\} \tag{242}$$

is the autocorrelation function of the process. In other words, the KL temporal modes are are eigenfunctions of the the auto-correlation function of the process. Since $C(t,s)$ is a Mercer's kernel (continuous symmetric non-negative definite kernel) we have that $\{\psi_k(t)\}$ is a complete orthonormal basis of $L^2([0,T])$.

Example: Let us compute the KL expansion of a stochastic process with *exponential* auto-correlation function

$$C(t,s) = \frac{\sigma^2}{2\tau} e^{-|t-s|/\tau}, \tag{243}$$

[23] It is straightforward to show that (239) follows form the variational principle

$$\min_{\psi_k} E([\psi_1, \psi_2, \ldots]) = \min_{\psi_k} \int_0^T \mathbb{E}\left\{\left|X(t;\omega) - \sum_{k=1}^{\infty} \sqrt{\lambda_k}\xi_k(\omega)\psi_k(t)\right|^2\right\} dt. \tag{237}$$

where τ denotes the correlation time. Note that (243) is an element of a Dirac delta sequence. This implies that

$$\lim_{\tau\to 0}\frac{\sigma^2}{2\tau}e^{-|t-s|/\tau}=\sigma^2\delta(t-s). \tag{244}$$

The eigenvalue problem (243) with $C(t,s)$ defined in(243) admits the analytical solution[24] (see [**?**])

$$\psi_k(t)=\frac{\tau z_k\cos(z_k t)+\sin(z_k t)}{\sqrt{\frac{1}{2}\left(\tau^2 z_k^2+1\right)T+\left(\tau^2 z_k^2-1\right)\frac{\sin(2z_k T)}{4z_k}+\frac{\tau}{2}\left(1-\cos(2z_k T)\right)}}, \tag{248}$$

where z_k are solution of the transcendental equation

$$\left(z_k^2-\frac{1}{\tau}\right)\tan(z_k T)-\frac{2z_k}{\tau}=0, \tag{249}$$

and

$$\lambda_k=\frac{\sigma^2}{\left(z_k^2\tau^2+1\right)} \tag{250}$$

are the KL eigenvalues. The KL eigenvalues become smaller and smaller as z_k increases. The eigenvalue decay is more pronounced for larger correlation lengths τ, while for very small correlation lengths the eigenvalue decay rate is very small, eventually zero for zero correlation length.

For practical purposes, the KL series expansion (234) is usually truncated to a finite number of terms. As we just discussed, the number of terms is inversely proportional to τ: the smaller τ the larger the number of terms. The number of terms M in the KL series expansion (234) is usually chosen by thresholding the relative "energy" of the process as

$$\frac{\sum_{k=1}^{M}\lambda_k}{\sum_{k=1}^{\infty}\lambda_k}\simeq 0.95. \tag{251}$$

This implies that the modes we retain in the series capture about 95% of the process "energy". In Figure 7 we plot samples of the exponentially correlated Gaussian random process

$$X(t;\omega)=\sin(t)+\frac{\sigma}{2\tau}\sum_{k=1}^{M}\sqrt{\lambda}_k\xi_k(\omega)\psi_k(t)\qquad t\in[0,20], \tag{252}$$

for $\tau=1$ and $\tau=0.1$.

[24] To compute the analytical solution of the KL eigenvalue problem (241) with exponential covariance (243) let us first rewrite it as

$$\int_0^T e^{-c|t-s|}\psi_k(s)ds=\hat{\lambda}_k\psi_k(t),\qquad c=\frac{1}{\tau},\qquad \hat{\lambda}_k=\frac{2\tau}{\sigma^2}\lambda_k. \tag{245}$$

Differentiating with respect to t the equivalent expression

$$\int_0^t e^{-c(t-s)}\psi_k(s)ds+\int_t^T e^{c(t-s)}\psi_k(s)ds=\hat{\lambda}_k\psi_k(t) \tag{246}$$

yields the second-order boundary value problem

$$\begin{cases}\dfrac{d^2\psi_k}{dt^2}=\dfrac{c^2\hat{\lambda}_k-2c}{\hat{\lambda}_k}\psi_k(t)\\ \dfrac{d\psi_k(t)}{dt}=c\psi(0)\\ \dfrac{d\psi_k(T)}{dt}=c\psi(T)\end{cases} \tag{247}$$

The solution of the BVP (247) is (248)-(250).

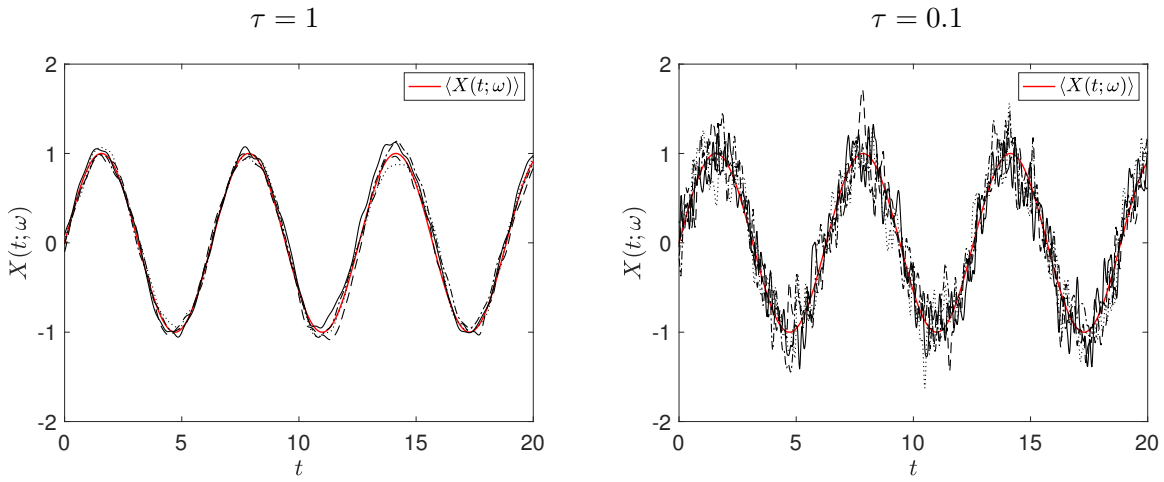

Figure 7: Samples of the exponentially correlated Gaussian random process (252) for different correlation times τ. The KL The mean of the process is shown in red. The truncation threshold for the number of terms M is set at 95% of the energy of the process (see Eq. (251)).

Remark: In the case where (241) cannot be solved analytically, we can resort to numerical method for Fredholm eigenvalue problems, e.g., Finite-difference methods, spectral methods, or Galerkin methods (see e.g., [19]). Of course it is also possible to define KL expansions of *random fields* by simply generalizing the bi-orthogonal series (234) as

$$X(\boldsymbol{x};\omega) = \sum_{k=1}^{\infty} \sqrt{\lambda}_k \xi_k(\omega)\psi_k(\boldsymbol{x}). \tag{253}$$

The computation of the KL expansion follows exactly the same steps as before, i.e., $\psi_k(\boldsymbol{x})$ are solutions to the eigenvalue problem

$$\int_V C(\boldsymbol{x},\boldsymbol{y})\psi_k(\boldsymbol{y})d\boldsymbol{y} = \lambda_k \psi_k(\boldsymbol{x}), \tag{254}$$

where V is some spatial domain.

Remark: To sample realizations of the random process (234) we need to sample the random variables $\{\xi_1,\dots,xi_M\}$. Such random variables are (by construction) orthonormal (see (235)), i.e., they are uncorrelated and have variance equal to one. Clearly, if $\{\xi_1,\dots,xi_M\}$ are jointly Gaussian then we know that the condition (235) is necessary and sufficient for independence. Hence, in the Gaussian case, sampling the joint PDF of $\{\xi_1,\dots,xi_M\}$ reduces to sampling the PDF of an independent set of one-dimensional Gaussian random variables with zero mean and variance one. More generally, if we have available the joint PDF $p(\xi_1,\dots,xi_M)$, e.g., by computing (238), then we can sample it using Markov Chain Monte Carlo (MCMC) methods, e.g., the Metropolis-Hastings algorithm or Gibbs sampling.

Wiener process. The Wiener process is a zero-mean continuous-time random process satisfying the following conditions:

- The increment $X(t+\tau;\omega) - X(t;\omega)$ is a Gaussian random variable with zero mean and variance τ. In other words, the conditional probability density of $X(t+\tau;\omega)$ given $X(t;\omega)$, is Gaussian with zero mean and variance τ.
- The random variables (increments)

$$X(t_1;\omega) - X(t_0;\omega) \qquad \text{and} \qquad X(t_3;\omega) - X(t_2;\omega) \tag{255}$$

are statistically independent for $t_0 < t_1 \leq t_2 < t_3$. In other words, the Wiener process is an *independent increment* process.

- The process $X(t;\omega)$ is continuous with probability one, i.e.,

$$P\left(\{\omega : \lim_{s\to t} |X(s;\omega) - X(t;\omega)|\}\right) = 1 \qquad \text{for all } t \geq 0. \tag{256}$$

This means that almost all (except sets of measure zero) sample paths are continuous in the classical sense, but the process $X(t;\omega)$ is nowhere differentiable. Continuity with probability one implies continuity in probability, and therefore mean square continuity and continuity in distribution.

An very clear description of the Wiener process is provided by Wiener himself in [22, Lecture 1]. The simplest algorithm to sample a Wiener process leverages the fact that the process has Gaussian distributed independent increments. Let $\{t_k\}_{k=1,\ldots,n}$ be n distinct time instants

$$0 = t_0 < t_1 < \ldots < t_n. \tag{257}$$

Then

$$X(t_k;\omega) = \sum_{j=1}^{k} \sqrt{\Delta t_j}\xi_j(\omega) \qquad \Delta t_j = t_j - t_{j-1}, \tag{258}$$

where $\{\xi_j(\omega)\}$ are independent random variables with mean zero and variance 1. A closer look at (258), reveals

$$X(t_1;\omega) = \sqrt{\Delta t_1}\xi_1(\omega), \tag{259}$$

$$X(t_2;\omega) = X(t_1;\omega) + \sqrt{\Delta t_2}\xi_2(\omega) = \sqrt{\Delta t_1}\xi_1(\omega) + \sqrt{\Delta t_2}\xi_2(\omega) \tag{260}$$

$$\ldots$$

Since $X(t_k;\omega)$ is a superimposition of essentially an infinite number of independent random vairable, it is rather straightforward to show that the one time PDF of $X(t;\omega)$ is

$$p(x,t) = \frac{1}{\sqrt{2\pi t}} e^{-x^2/(2t)}, \tag{261}$$

i.e., Gaussian. This equation also follows from the conditional PDF identity

$$p(x,t) = \int_{\infty}^{\infty} p(x,t|y,s)p(y,s)dy \qquad t > s, \tag{262}$$

where $p(x,t|y,s)$ is the transition density[25], and $p(y,s)$ is the PDF of $X(s;\omega)$. If we set $s = 0$ then $p(y,0) = \delta(y)$ and, of course, this yields (261). The auto-correlation function of the Wiener process is

$$C(t,s) = \min(t,s) \tag{264}$$

With the autocorrelation function available we can compute a KL expansion of the Wiener process following the procedure outlined in the previous section. If we consider the time interval $[0,1]$ this yields the eigenvalue problem

$$\int_0^1 C(t,s)\psi_k(s) = \lambda_k \psi_k(t), \tag{265}$$

[25] From the recurrence relation

$$X(t_{k+1};\omega) = X(t_k;\omega) + \sqrt{\Delta t_{k+1}}\xi_{k+1}(\omega) \tag{263}$$

with $\xi_{k+1}(\omega)$ Gaussian with zero mean and variance one we see that the conditional PDF $p(x,t|y,s)$ is Gaussian with zero mean and variance $t - s$.

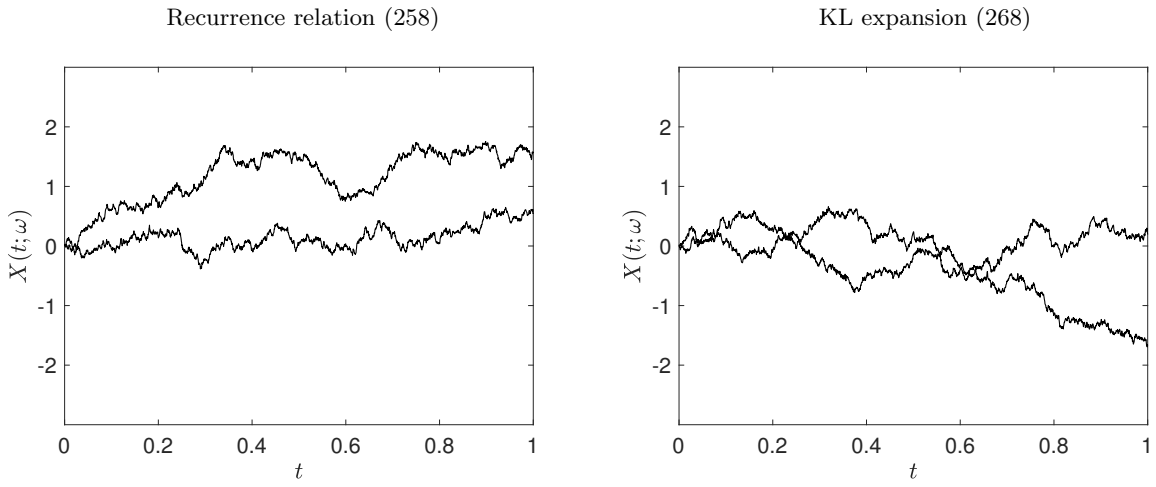

Figure 8: Wiener processes obtained by sampling the Karhunen-Loève expansion (268) with 10^5 terms on a temporal grid with 2000 points in $[0,1]$, and by iterating (258) on the same temporal grid.

the solution of which is

$$\psi_k(t) = \sqrt{2}\sin\left(\left[k-\frac{1}{2}\right]\pi t\right) \quad k=1,2,\dots \tag{266}$$

and

$$\lambda_k = \frac{4}{\pi^2\left(2k-1\right)^2}. \tag{267}$$

Substituting (266) and (267) into (234) yields

$$X(t;\omega) = \sum_{k=1}^{\infty}\frac{2\sqrt{2}}{\pi\left(2k-1\right)}\xi_k(\omega)\sin\left(\left[k-\frac{1}{2}\right]\pi t\right), \tag{268}$$

where $\xi_k(\omega)$ are independent Gaussian random variables with zero mean and variance one (they satisfy (235)). The series expansion in (268) can be eventually truncated to a finite number of terms, depending on the threshold set on the eigenvalues (267) (which decay as $1/k$). In Figure 8 we plot a few samples of the Wiener process we obtain by sampling (268) with 10^5 terms on a temporal grid with 2000 points in $[0,1]$, and the Wiener process we obtain by iterating (258) on the same temporal grid. Note that if $X(t;\omega)$ is a Wiener process in $t\in[0,1]$ then

$$\sqrt{T}X\left(\frac{t}{T};\omega\right) \qquad t\in[0,T] \tag{269}$$

is a Wiener process in $[0,T]$. This expression is obtained by simply changing the variables in the integral equation (265). The expression (269) shows that features of a Wiener process do not change while zooming in or out. In other words, the Wiener process is self-similar.

Appendix A: Derivation of the Liouville equation

Consider the nonlinear dynamical system

$$\begin{cases}\dfrac{d\boldsymbol{x}(t)}{dt} = \boldsymbol{f}(\boldsymbol{x}(t))\\ \boldsymbol{x}(0) = \boldsymbol{x}_0(\omega)\end{cases} \tag{270}$$

where $\boldsymbol{x}_0(\omega)$ is a random vector with known joint probability density function $p_0(\boldsymbol{x})$. We know that if $\boldsymbol{f}(\boldsymbol{x})$ is continuously differentiable in $\boldsymbol{x}$ then (270) admits a smooth flow $\boldsymbol{x}(t,\boldsymbol{x}_0(\omega))$, which is at least continuously differentiable in $\boldsymbol{x}_0$. The flow is also continuously differentiable in t, i.e., $x(t,\boldsymbol{x}_0(\omega))$ is a diffeomorphism in t. We are interested in determining an evolution equation for $p(\boldsymbol{x},t)$, i.e., the probability density function of $\boldsymbol{x}(t,\boldsymbol{x}_0)$ at time t. To this end, consider the characteristic function representation of the PDF $p(\boldsymbol{x},t)$

$$\phi(\boldsymbol{a},t)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}e^{i\boldsymbol{a}\cdot\boldsymbol{x}(t;\boldsymbol{x}_0)}p(\boldsymbol{x}_0)d\boldsymbol{x}_0=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}e^{i\boldsymbol{a}\cdot\boldsymbol{x}}p(\boldsymbol{x},t)d\boldsymbol{x} \tag{271}$$

Differentiating with respect to t yields

$$\begin{aligned}\frac{\partial\phi(\boldsymbol{a},t)}{\partial t}&=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}i\boldsymbol{a}\cdot\frac{\partial\boldsymbol{x}(t,\boldsymbol{x}_0)}{\partial t}e^{i\boldsymbol{a}\cdot\boldsymbol{x}(t;\boldsymbol{x}_0)}p(\boldsymbol{x}_0)d\boldsymbol{x}_0\\&=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}i\boldsymbol{a}\cdot\boldsymbol{f}\left(\boldsymbol{x}(t,\boldsymbol{x}_0)\right)e^{i\boldsymbol{a}\cdot\boldsymbol{x}(t;\boldsymbol{x}_0)}p(\boldsymbol{x}_0)d\boldsymbol{x}_0\\&=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}i\boldsymbol{a}\cdot\boldsymbol{f}\left(\boldsymbol{x}\right)e^{i\boldsymbol{a}\cdot\boldsymbol{x}}p(\boldsymbol{x},t)d\boldsymbol{x}\\&=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial}{\partial\boldsymbol{x}}\left(e^{i\boldsymbol{a}\cdot\boldsymbol{x}}\right)\cdot\boldsymbol{f}\left(\boldsymbol{x}\right)p(\boldsymbol{x},t)d\boldsymbol{x}\\&=-\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}e^{i\boldsymbol{a}\cdot\boldsymbol{x}}\nabla\cdot\left(\boldsymbol{f}\left(\boldsymbol{x}\right)p(\boldsymbol{x},t)\right)d\boldsymbol{x}.\quad\text{(integrating by parts)}\end{aligned} \tag{272}$$

By using (271) and (272) we obtain

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}e^{i\boldsymbol{a}\cdot\boldsymbol{x}}\left[\frac{\partial p(\boldsymbol{x},t)}{\partial t}+\nabla\cdot\left(\boldsymbol{f}\left(\boldsymbol{x}\right)p(\boldsymbol{x},t)\right)\right]d\boldsymbol{x}=0,\quad\text{for all }\boldsymbol{a}\in\mathbb{R}^n, \tag{273}$$

which implies that the function between square bracket must be equal to zero for all $\boldsymbol{x}$ and all t, i.e.,

$$\frac{\partial p(\boldsymbol{x},t)}{\partial t}+\nabla\cdot\left(\boldsymbol{f}\left(\boldsymbol{x}\right)p(\boldsymbol{x},t)\right)=0\qquad\text{(Liouville equation).} \tag{274}$$

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv:1701.07875*, pages 1–30, 2017.

[2] M. J. Beran. *Statistical continuum theories*. New York: Interscience Publishers, 1968.

[3] D. Bigoni, A. P. Engsig-Karup, and Y. M. Marzouk. Spectral tensor-train decomposition. *SIAM J. Sci. Comput.*, 38(4):A2405–A2439, 2016.

[4] Z. I. Botev, J. F. Grotowski, and D. P. Kroese. Kernel density estimation via diffusion. *Annals of Statistics*, 38(5):2916–2957, 2010.

[5] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. Wiley, 2008.

[6] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC, 2011.

[7] A. Dektor, A. Rodgers, and D. Venturi. Rank-adaptive tensor methods for high-dimensional nonlinear pdes. *Journal of Scientific Computing*, 88(36):1–27, 2021.

[8] A. Dektor and D. Venturi. Dynamically orthogonal tensor methods for high-dimensional nonlinear PDEs. *J. Comput. Phys.*, 404:109125, 2020.

[9] G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley, second edition, 2007.

[10] A. I. Khuri. Applications of Dirac's delta function in statistics. *Int. J. Math. Educ. Sci. Technol.*, 35(2):185–195, 2004.

[11] V. I. Klyatskin. *Dynamics of stochastic systems*. Elsevier Publishing Company, 2005.

[12] R. Kubo. Generalized cumulant expansion method. *Journal of the Physical Society of Japan*, 17(7):1100–1120, 1962.

[13] A. S. Monin and A. M. Yaglom. *Statistical Fluid Mechanics, Volume II: Mechanics of Turbulence*. Dover, 2007.

[14] I. V. Oseledets. Tensor-train decomposition. *SIAM J. Sci. Comput.*, 33(5):2295—-2317, 2011.

[15] A. Papoulis. *Probability, random variables and stochastic processes*. McGraw-Hill, third edition, 1991.

[16] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.*, 378:606–707, 2019.

[17] H. Rue and L. Held. *Gaussian Markov random fields*. Chapman & Hall/CRC, 2005.

[18] D. Venturi. The numerical approximation of nonlinear functionals and functional differential equations. *Physics Reports*, 732:1–102, 2018.

[19] D. Venturi, M. Choi, and G. E. Karniadakis. Supercritical quasi-conduction states in stochastic Rayleigh-Bénard convection. *Int. J. Heat and Mass Transfer*, 55(13-14):3732–3743, 2012.

[20] D. Venturi and A. Dektor. Spectral methods for nonlinear functionals and functional differential equations. *Res. Math. Sci.*, 8(27):1–39, 2021.

[21] D. Venturi and X. Li. The Mori-Zwanzig formulation of deep learning. *ArXiv*, (2209.05544):1–40, 2022.

[22] N. Wiener. *Nonlinear problems in random theory.* MIT Press, 1966.

[23] D. Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach.* Princeton University Press, 2010.