Probability spaces and random variables

There is a well-developed and powerful mathematical theory that allows us to describe randomness in phenomena we observe and experience every day. Such theory is called probability theory and it is at the basis of all UQ techniques we will discuss in this course. The proper mathematical setting of probability theory is quite abstract and technical, as is involves rather advanced mathematical concepts such as *measure theory* [4]. However, we will try to avoid most technicalities whenever possible, and have a more practical version of probability theory that allows to perform computations.

Probability space

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

- Ω (sample space): the set of all possible outcomes of the experiment
- $\mathcal{B}$ (event space): set of possible events, en event being a set defined as union or intersection of elements in the sample space. As we shall see hereafter, such elements are usually defined by some condition.
- P (probability measure): this function assigns each event in $\mathcal{B}$ a probability, which is a number between 0 and 1.

Example 1: Let us roll a fair dice with 6 faces once. In this case we can define the sample space as

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

In particular, we may be interested in the following events[1]

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

These events can be phrased as: "rolling the dice produces no number" (event $\emptyset$); "rolling the dice returns a 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 $\{1,3,5\}$). As we shall see here hereafter, the events (2) are all subsets of a so-called σ-algebra $\mathcal{B}(\Omega)$ We can assign probabilities to the events in (2) 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 process, as we can imagine rolling a dice, and how its outcomes affect probability. Similarly we can consider the process of rolling two dices (or the same dice twice). The sample space in this case will be the set of elements

$$\Omega = \{\{1,1\},\{1,2\},\ldots,\{5,6\},\{6,6\}\} \tag{4}$$

As before, we can define a set of events we are interested in, and assign corresponding probabilities. In a similar way, we can assign, e.g., the probability of winning various prizes in the the Mega-Millions lottery (assuming the lottery is fair).

Example 2: Let (θ, r) 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 (θ, r) is a vector with two random components. In this case, the outcome of the "experiment" are realizations of two real random variables

[1] As we will see, the set $\mathcal{B}$ defined in (2) is a σ-algebra.

(polar coordinates (θ, r) of the location where the leaf lands). The sample space in this case can be thought of as a Cartesian product of two intervals

$$\Omega = [0, 2\pi[\times[0, \infty[\tag{5}$$

We can define the following events (distance from the tree)[2]:

$$\mathcal{B} = \{\emptyset, \Omega, \underbrace{\{r \leq 1\}}_{\text{event 1}}, \underbrace{\{1 < r \leq 2\}}_{\text{event 2}}, \underbrace{\{r > 2\}}_{\text{event 3}}\}. \tag{6}$$

At this point we can assign probabilities[3] to each event in (6), which can be done, e.g., by observing many leaves falling off a tree, or by running a fluid dynamics model (repeated simulations).

Example 3: Consider an infinite (uncountable) collection of real-valued continuous functions $X(t;\omega)$ (stochastic process) defined in the temporal interval $[0, T]$. In this case the sample can be thought of as the set of all possible sample paths $X(t;\omega)$. An event, can be defined, e.g., by singling out of the sample space ensemble of paths satisfying a certain condition. For instance, we can define an event by requiring that for some fixed $t > s > 0$

$$E_1 = \{\omega : X(t;\omega) < 1\} \cap \{\omega : X(s;\omega) < 1\}. \tag{7}$$

Alternatively, we can require that for all $t \in [0, T]$

$$E_2 = \{\omega : X(t;\omega) \geq 1\} \tag{8}$$

Both events are (measurable) subsets of the sample space (space of continuous functions defined in $[0, T]$) The probability the events E_1 or E_2 may be estimated using a frequency approach, i.e., $P(E_i) \simeq n_{E_i}/n$, where n_{E_i} is the number events E_i that occurs over n trials (assuming we can observe $X(t;\omega)$).

Remark: In all three examples we defined a sample space that coincides with the range of a discrete random variable (Example 1), a continuous two-dimensional random vector (Example 2), or a continuous random process (Example 3). The sample space does not need to be the actual space in which we observe outcomes of a random variables or process, but rather it can be just a space in which we pick labels for a certain random quantity. For instance, it can be shown that the set of sample paths of a Wiener process (continuous random process) can be put in a correspondence with $\Omega = [0, 1]$ (sample space), meaning that we can set up a map that labels each sample path of the random process with a real number in $[0, 1]$ (see [9]).

The event space $\mathcal{B}$. As we shall see hereafter, in order to perform set operations and corresponding operations on probabilities we need to make sure that the event space $\mathcal{B}(\Omega)$ has the structure of a σ-algebra on Ω. Broadly speaking, 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,

$$A, B \in \mathcal{B} \quad \Rightarrow \quad \begin{cases} A \cap B \in \mathcal{B} \\ A \cup B \in \mathcal{B} \\ A^c, B^c \in \mathcal{B} \quad \text{(complement of } A \text{ and } B\text{, i.e., } A^c = \Omega \setminus A) \end{cases} \tag{9}$$

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

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}, \qquad \bigcap_{i=1}^{\infty} A_i \in \mathcal{B} \qquad \text{(countable union and intersection).} \tag{10}$$

[2] As we shall see, the set $\mathcal{B}$ derived in (6) is a σ-algebra

[3] For a thorough discussion on the meaning of probability and how to assign probabilities see [7, Chapters 1-3].

Note that by this definition, the simplest σ-algebra we can possibly think of is

$$\mathscr{B} = \{\emptyset, \Omega\} \qquad \text{(trivial } \sigma\text{-algebra)}. \tag{11}$$

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.

  $$2^{\Omega} = \{\emptyset, a, b, c, \{a, b\}, \{a, c\}, \{b, c\}, \underbrace{\{a, b, c\}}_{\Omega}\} \qquad \text{(power set)}. \tag{12}$$

  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 (12) 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 on Ω can be represented as a sub-algebra of the power set 2^{Ω} (Stone's representation theorem [4]). This is why the σ-algebra $\mathscr{B}$ on an uncountably infinite sample space Ω is often written as a subset of the power set 2^{Ω}, i.e., $\mathscr{B}(\Omega) \subseteq 2^{\Omega}$.

- 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[4], and all (countable) unions and intersections of open and closed sets. Such σ-algebra is a sub-algebra of the power set 2^{Ω}.

- Let H be a Hilbert space of functions. We can define a Borel σ-algebra on $\mathscr{B}(H)$ as the collection of all open subsets of H (subsets of functions in H). $\mathscr{B}(H)$ defines which subsets of the Hilbert space are "measurable".

Borel σ-algebras. The last two examples of σ-algebras, i.e., $\mathscr{B}(\mathbb{R})$ and $\mathscr{B}(H)$ are Borel σ-algebras. A Borel σ-algebra on a space S (such as the real line $\mathbb{R}$ or the Hilbert space H) is generated by the collection of all open sets of S. It includes all sets that can be constructed from open sets using countable unions, countable intersections, and complements. Clearly the trivial σ-algebra (11) is not a Borel σ-algebra, as it does not include any subset of Ω.

[4] Recall that the complement of an element in $\mathscr{B}$ still belong to the $\mathscr{B}$. Hence if we pick $A =]-\infty, 1[\cup]2, \infty[\in \mathscr{B}$ (union of two open sets), then $A^c = [1, 2] \in \mathscr{B}$ (closed interval).

Probability measure. The probability function

$$P : \mathcal{B} \to [0,1] \tag{13}$$

assigns to each event A (i.e., a set) in the σ-algebra $\mathcal{B}$ a number $P(A) \in [0,1]$. In other words, $P(A)$ measures the likelihood that the event represented by the set A occurs. The probability function P satisfies the properties of a *measure* (hence the name probability measure[5]):

1. $P(\emptyset) = 0$.
2. $P(\Omega) = 1$.
3. For all countable collections of *disjoint* sets $A_i \in \mathcal{B}$
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \tag{14}$$
4. For all $A, B \in \mathcal{B}$,
$$P\left(A \cup B\right) = P(A) + P(B) - P(A \cap B). \tag{15}$$

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

$$P(A) = P(B) + P(B^c \cap A) \geq P(B). \tag{16}$$

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

$$\mu_n(A) = \frac{n_A}{n} \tag{17}$$

then

$$P(A) = \lim_{n\to\infty} \mu_n(A). \tag{18}$$

Random variables

Let $(\Omega, \mathcal{B}, P)$ 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.,

$$X : \Omega \mapsto \mathbb{R}. \tag{19}$$

The mapping (19) induces a probability measure (known as "push-forward measure")

$$P_X(A) = \mathcal{B}(\mathbb{R}) \mapsto [0,1], \tag{20}$$

where $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$. The measure $P_X(A)$ is defined as[6]

$$P_X(A) = P\left(\{\omega \in \Omega : X(\omega) \in A\}\right) \quad \text{for all} \quad A \in \mathcal{B}(\mathbb{R}). \tag{21}$$

[5] In real analysis, the pair $(\Omega, \mathcal{B})$ is called *measurable space* [4]. The elements of $\mathcal{B}$, i.e., the events, are called *measurable sets*. The triple $(\Omega, \mathcal{B}, P)$ is called *probability space*, which is essentially a measurable space in which we define a probability measure.

[6] The set $\{\omega \in \Omega : X(\omega) \in A\}$ at the right hand side of (54) is called the *pre-image* of the set $A \in \mathbb{R}$ under the mapping X, and denoted by $X^{-1}(A)$.

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

$$F(x) = P(\underbrace{\{\omega : X(\omega) \leq x\}}_{\text{event}}) = P_X\left(]-\infty, x]\right) \qquad x \in \mathbb{R}. \tag{22}$$

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. As such it may seem that $F(x)$ does not carry sufficient statistical information about the random variable X to enable us to measure other sets. However, given that P is defined on a Borel σ-algebra, we can see that F actually allows us to measure most sets of interest in the range of X. An exception is perhaps the measure of the intersection of two non-overlapping sets, which is of course zero and cannot be written in terms of $F(x)$.

By using the axiomatic properties of the probability measure P it can be shown 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. $P\left(\{\omega : X(\omega) > x\}\right) = 1 - F(x)$,
5. $F(x)$ is continuous from the right, i.e.,
$$\lim_{\epsilon \to 0^+} F(x+\epsilon) = F(x), \tag{23}$$
6. $F(x)$ is not continuous from the left (for discrete random variables),
7. $P\left(\{\omega : a < X(\omega) \leq b\}\right) = F(b) - F(a)$
8. $P\left(\{\omega : a \leq X(\omega) \leq b\}\right) = F(b) - \lim\limits_{\epsilon \to 0^+} F(a+\epsilon)$.

The proof of 1.-8. can be found in [7, 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 (PDF). The probability density function (PDF) $p(x)$ of the random variable $X(\omega)$ is (technically speaking) the Radon–Nikodym derivative[7] (assuming it exists) of the probability

[7] A probability measure P on the measurable space $(\Omega, \mathcal{B})$ is said to be *absolutely continuous* with respect to another measure ν if for all events $E \in \mathcal{B}(\Omega)$ such that $P(E) = 0$ we have $\nu(E) = 0$. 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

$$P(E) = \int_E p(x)dx. \tag{24}$$

Setting the event E in (24) as

$$E = \{w : X(\omega) \leq x\} \in \mathcal{B}(\Omega) \tag{25}$$

yields equation (26).

measure P with respect to the Lebesgue measure. The existence of the Radon–Nikodym derivative allows us to write the cumulative distribution function (22) as

$$F(x) = \int_{-\infty}^{x} p(y)dy. \tag{26}$$

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

$$p(x) = \frac{dF(x)}{dx}. \tag{27}$$

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

$$p(x) = \delta(x-a) \qquad \text{(PDF of the random variable } X(\omega) = a \text{ for all } \omega \in \Omega), \tag{28}$$

and

$$p(x) = \sum_{i=1}^{N} p_i \delta(x - x_i) \qquad \text{(PDF of a discrete random variable with range } \{x_1, \dots, x_n\}). \tag{29}$$

For example, the PDF of a fair dice with 6 faces is

$$p(x) = \frac{1}{6}\sum_{i=1}^{6} \delta(x-i). \tag{30}$$

By using the properties of the cumulative distribution function $F(x)$ it can be shown that the PDF satisfies the following properties

1. $p(x) \geq 0$ (positivity),
2. $\displaystyle\int_{-\infty}^{\infty} p(x)dx = 1$ (normalization),
3. $\displaystyle P\left(\{\omega : x_1 < X(\omega) \leq x_2\}\right) = \int_{x_1}^{x_2} p(x)dx$,
4. $P\left(\{\omega : x < X(\omega) \leq x + dx\}\right) = p(x)dx$.

Frequency interpretation of the PDF: 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 property 4. above , and the frequency interpretation of probability we conclude that

$$p(x)\Delta x \simeq \frac{n_{\Delta x}}{n} \quad \Rightarrow \quad p(x) \simeq \frac{1}{\Delta x}\frac{n_{\Delta x}}{n}. \tag{31}$$

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., [1]). In Figure 2 we estimate the PDF of a Gaussian random variable using frequency approach, i.e., equation (31), and the kernel density estimation method discussed in [1].

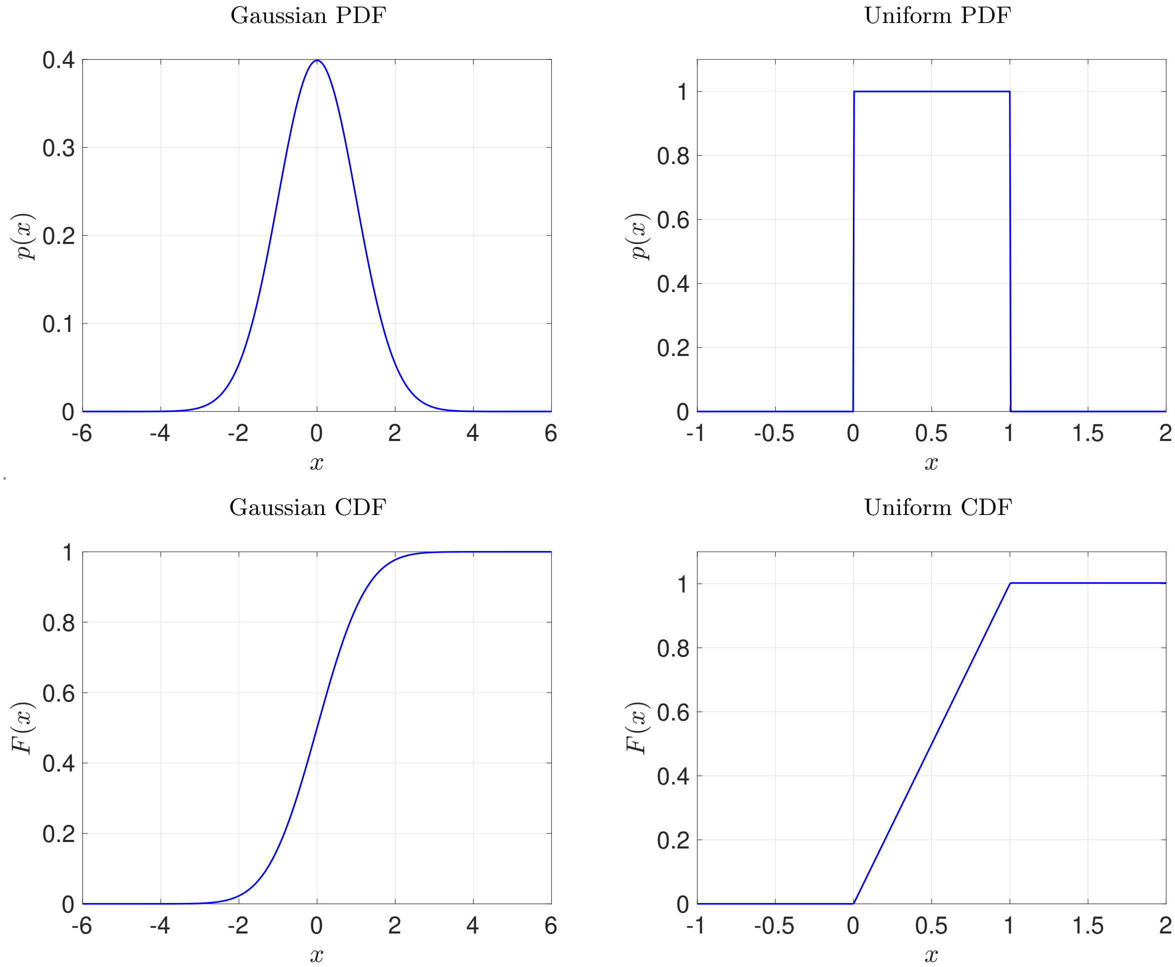

Figure 1: PDFs and CDFs of Gaussian (left) and uniform (right) random variables.

Examples of one-dimensional PDFs and CDFs

- Gaussian (continuous):

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad F(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \qquad x \in \mathbb{R}. \tag{32}$$

- Uniform (continuous):

$$p(x) = \frac{1}{b-a}, \qquad F(x) = x, \qquad x \in [a,b]. \tag{33}$$

- Binomial (discrete):

$$p(x) = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \delta(x-i), \qquad p \in]0,1[, \qquad x \geq 0. \tag{34}$$

- Poisson (discrete):

$$p(x) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \delta(x-k), \qquad \lambda \in]0,\infty[, \qquad x \geq 0. \tag{35}$$

See Figure 1.

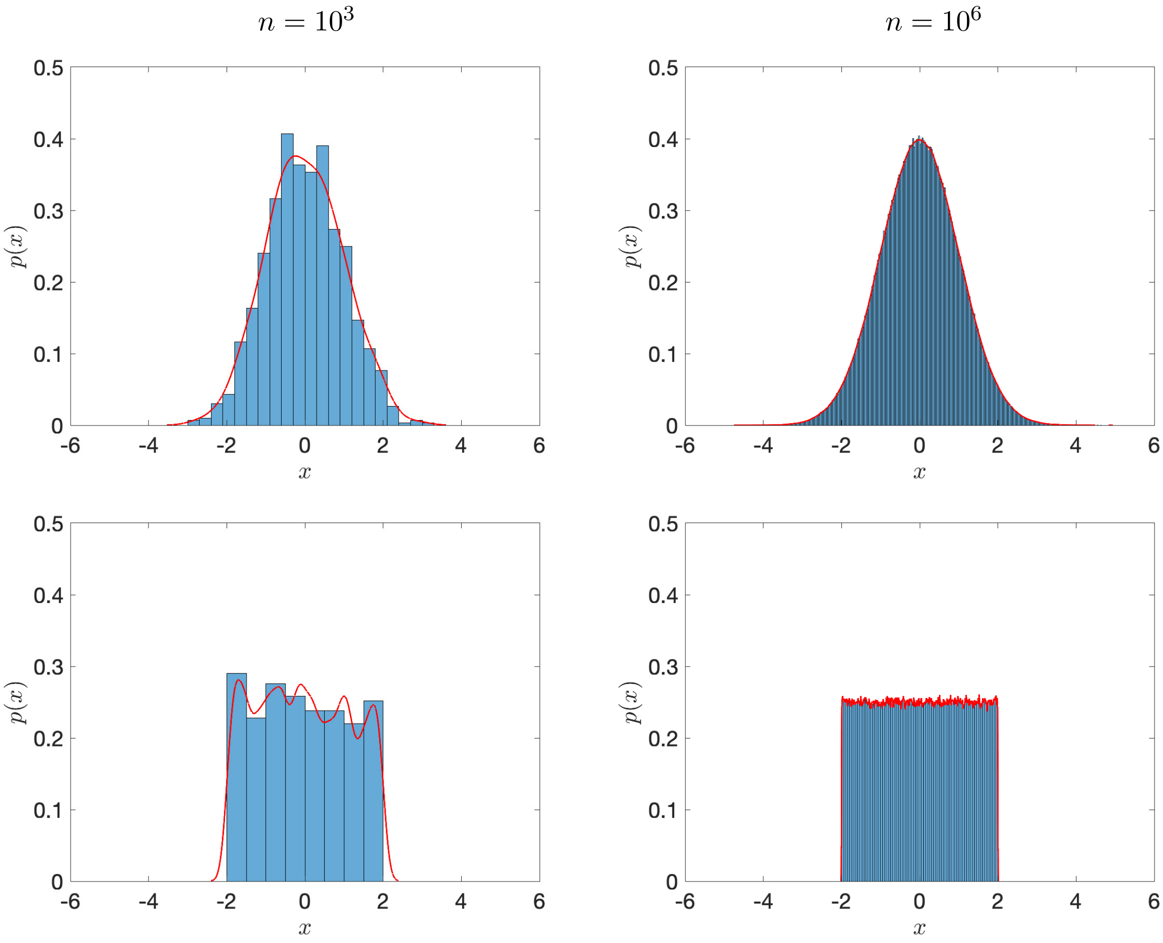

Figure 2: Estimation of the PDF of a Gaussian random variable (first row) and a uniform random variable (second row) using the frequency approach, i.e., formula (31), and the kernel density estimate discussed in [1] (red line) . We plot results for a different number of samples n.

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, suppose we are give a deterministic function

$$g : \mathbb{R} \to \mathbb{R} \tag{36}$$

such that

$$Y(\omega) = g(X(\omega)) \tag{37}$$

for all $\omega \in \Omega$, and suppose we know the PDF of X, $p_X(x)$. What is PDF of $Y(\omega)$?

Since X and Y are defined on the same probability space we have

$$F_Y(y) = P\left(\{\omega : Y(\omega) \leq y\}\right) = P\left(\{\omega : g(X(\omega)) \leq y\}\right). \tag{38}$$

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

$$B_y = \{\omega : g(X(\omega)) \leq y\} \tag{39}$$

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 3 (in yellow) for a compactly supported prototype function $g(x)$ (support in $[a, b]$) and a

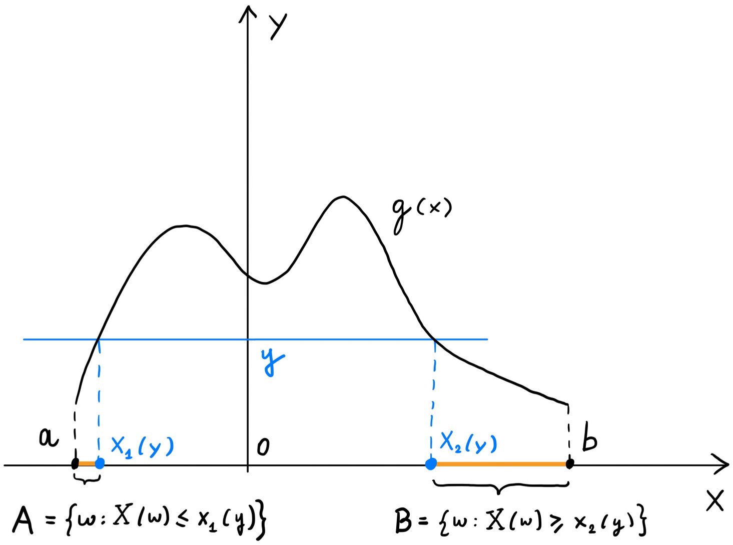

Figure 3: Sketch of the set B_y defined in equation (39) (yellow lines). The random variable X is compactly supported in $[a, b]$. The distribution function of the random variable $Y(\omega) = g(X(\omega))$ evaluated at y is the measure of the set $B_y = A \cup B$ (union of the two yellow lines), i.e., $F_Y(y) = F_X(x_1(y)) + 1 - F_X(x_2(y))$.

specific value of y. Clearly, the distribution function $F_Y(y)$ must be defined case-by-case. With reference to Figure 3 we have

$$F_Y(y) = F_X(x_1(y)) + 1 - F_X(x_2(y)), \tag{40}$$

where $x_1(y)$ and $x_2(y)$ are the branches of the inverse function $g^{-1}(y)$. The function (40) 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

$$p_Y(y) = \sum_{i=1}^{r} \frac{p_X(x_i(y))}{|g'(x_i(y))|}, \tag{41}$$

where $x_i(y)$ $(i = 1, \ldots, r)$ are the real roots of the equation $g(x) = y$.

Proof. We prove the theorem using Fourier transforms[8]. Let

$$\phi_Y(a) = \int_{-\infty}^{\infty} e^{iay} p_Y(y) dy = \int_{-\infty}^{\infty} e^{iag(x)} p_X(x) dx. \tag{42}$$

Taking the inverse Fourier transform yields

$$p_Y(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{ia(g(x)-y)} p_X(x) dx da. \tag{43}$$

Next, recall that

$$\delta(g(x) - y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ia(g(x)-y)} da. \tag{44}$$

[8] 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. (95)).

Substituting (44) into (43) yields (see also [5])

$$p_Y(y) = \int_{-\infty}^{\infty} \delta\left(g(x) - y\right) p_X(x) dx. \tag{45}$$

At this point we use the well-known identity[9]

$$\delta\left(g(x) - y\right) = \sum_{i=1}^{r} \frac{\delta\left(x - x_i(y)\right)}{\left|g'\left(x_i(y)\right)\right|}, \tag{46}$$

where $x_i(y)$ are the real roots of the $y = g(x)$ for each $y \in \mathbb{R}$. A substitution of (46) into (45) yields (41). This completes the proof. Alternatively, we can rely on conservation of probability mass which can be stated as

$$p_Y(y) dy = \sum_{i=1}^{r} p_X(x_i(y)) |dx_i|, \tag{47}$$

where $|dx_i|/dy = 1/|g'(x_i(y))|$. The absolute value in dx_i takes cares of the fact that $dy > 0$ may be associated with dx_i positive or negative (see, e.g., Figure 3).

□

Examples of PDF 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

  $$x_1(y) = \sqrt{y}, \qquad x_2(y) = -\sqrt{y} \qquad y \geq 0. \tag{48}$$

  By using Theorem 1 we immediately obtain

  $$p_Y(y) = \frac{1}{2\sqrt{y}} \left[p_X(\sqrt{y}) + p_X(-\sqrt{y})\right]. \tag{49}$$

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

  $$p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{50}$$

  then

  $$p_Y(y) = \frac{e^{-y/2}}{\sqrt{2\pi y}} \qquad (\chi^2\text{-distribution}). \tag{51}$$

  Similarly, if X is uniformly distributed in $[-1, 1]$ then[10]

  $$p_Y(y) = \begin{cases} \dfrac{1}{2\sqrt{y}} & \text{for all } 0 \leq y \leq 1 \\ 0 & \text{for all } y > 1 \text{ or } y < 0 \end{cases} \tag{52}$$

- 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 inverse) for all $y > 0$ as

  $$x = \frac{\log(y)}{t} \qquad y > 0. \tag{53}$$

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

  $$p_Y(y) = \frac{1}{ty} p_X\left(\frac{\log(y)}{t}\right) \qquad y > 0. \tag{54}$$

[9] The identity (46) if and only if $g'(x_i(y)) \neq 0$.

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

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
$$\begin{cases} \dfrac{dx}{dt} = f(x) \\ x(0) = X(\omega) \end{cases} \tag{55}$$
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
$$p(x,t) = \int_{-\infty}^{\infty} \delta\left(x - x(t,y)\right) p_X(y)dy. \tag{56}$$
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 (56) analytically. To this end, consider the system
$$\begin{cases} \dfrac{dx}{dt} = x^2 \\ x(0) = X(\omega) \end{cases} \tag{57}$$
We known that the analytical solution (flow map) is
$$x(t, X) = \frac{X(\omega)}{1 - tX(\omega)}. \tag{58}$$
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
$$g(x) = \frac{x}{1 - tx} = y \tag{59}$$
uniquely for each $t \geq 0$ ($x \leq 0$) as
$$x(1 + ty) = y \quad \Rightarrow \quad x(y) = \frac{y}{1 + ty}. \tag{60}$$
The first derivative of (59) with respect to x evaluated at the unique root $x(y) = y/(1 + ty)$ is
$$g'(x(y)) = \frac{1}{(1 - tx(y))^2} = (1 + ty)^2. \tag{61}$$
At this point we use Theorem 1 to conclude that the PDF of the solution to the ODE (57) at each fixed time t is
$$p(x,t) = \frac{1}{(1 + tx)^2} p_X\left(\frac{x}{1 + tx}\right). \tag{62}$$
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
$$-1 \leq \frac{x}{1 + tx} \leq 0 \quad \Rightarrow -\frac{1}{(1 + t)} \leq x \leq 0. \tag{63}$$
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,
$$\int_{-\infty}^{\infty} p(x,t)dx = \int_{-1/(1+t)}^{0} \frac{1}{(1 + tx)^2} \underbrace{1}_{p_X\left(\frac{x}{1+tx}\right)} dx = 1. \tag{64}$$

- Next, consider the linear decay problem

$$\begin{cases} \dfrac{dx}{dt} = \xi(\omega)x \\ x(0) = 1 \end{cases} \tag{65}$$

where $\xi(\omega)$ is a random variable with known probability density $p_\xi(x)$. The analytical solution to (65) is

$$x(t, \xi(\omega)) = e^{t\xi(\omega)}. \tag{66}$$

By using equation (54), we immediately conclude that the probability density of the solution to (65) is

$$p(x,t) = \frac{1}{tx} p_X\left(\frac{\log(x)}{t}\right) \qquad x > 0. \tag{67}$$

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

$$-2t \le \log(x) \le 0 \quad \Rightarrow \quad e^{-2t} \le x \le 1. \tag{68}$$

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

$$p(x, 0) = \delta(x-1) \quad \text{(deterministic initial condition).} \tag{69}$$

For $t > 0$ the PDF of the solution to (65) is[11]

$$p(x,t) = \frac{1}{2tx} \quad \text{for} \quad e^{-2t} \le x \le 1. \tag{71}$$

The Liouville equation. The PDF of the solution of the Cauchy problem (55) satisfies the following linear hyperbolic conservation law

$$\frac{\partial p(x,t)}{\partial t} + \frac{\partial}{\partial x}\left(f(x)p(x,t)\right) = 0, \qquad p(x,0) = p_X(x). \tag{72}$$

This equation is known as Liouville equation. It is straightforward to show by using the method of characteristics that (62) is the solution of Liouville equation (72) for $f(x) = x^2$, i.e., for the dynamical system (57). More generally, that the joint probability density function of the phase space variables of the n-dimensional autonomous nonlinear dynamical system

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}(\boldsymbol{x}), \qquad \boldsymbol{x}(0) = \boldsymbol{X}(\omega) \tag{73}$$

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

$$\frac{\partial p(\boldsymbol{x},t)}{\partial t} + \nabla \cdot \left(\boldsymbol{f}(\boldsymbol{x})p(\boldsymbol{x},t)\right) = 0. \tag{74}$$

To solve (74) 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 [2, 3] or physics-informed neural network techniques [8].

[11] Note that the PDF (71) integrates to one. In fact

$$\int_{-\infty}^{\infty} p(x,t)dx = \int_{e^{-2t}}^{1} \frac{1}{tx} \underbrace{\frac{1}{2}}_{p_X\left(\frac{\log(x)}{t}\right)} dx = 1. \tag{70}$$

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 given probability density $p_Y(y)$. With such mapping g available we can transform each sample of $X(\omega)$ to a sample of $Y(\omega)$, 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

$$F_Y(y) = \int_{-\infty}^{y} p_Y(x)dx \tag{75}$$

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

Proof. Suppose that F_Y is invertible[12]. Let us show that the random variable $F_Y^{-1}(X)$ has indeed cumulative distribution function $F_Y(y)$. By definition,

$$\begin{aligned} F_Y(y) &= P\left(\{\omega : Y(\omega) \leq y\}\right) \\ &= P\left(\{\omega : F_Y^{-1}(X(\omega)) \leq y\}\right) \\ &= P\left(\{\omega : X(\omega) \leq F_Y(y)\}\right) \qquad (F_Y \text{ invertible and nondecreasing}) \\ &= F_X(F_Y(y)) \\ &= F_Y(y). \end{aligned} \tag{76}$$

In fact, since $X(\omega)$ is uniform in $[0,1]$ we have $F_X(x) = x$ for all $x \in [0,1]$. This shows that the random variable $F_Y^{-1}(X)$, where X is uniform in $[0,1]$, has as distribution function $F_Y(y)$.

□

How do we use Lemma 1 to sample the PDF of Y, i.e., $p_Y(y)$? Very simple: we first generate samples of a uniform random variable X in $[0,1]$ and then we map each sample to a sample of p_Y by applying the nonlinear transformation

$$Y = F_Y^{-1}(X). \tag{77}$$

In practice, to compute $F_Y(y)$ and the inverse of the cumulative distribution function $F_Y^{-1}(y)$ we can proceed numerically, e.g., by using the cumulative trapezoidal rule and interpolation. In Figure 4 we show the results of this procedure applied to a PDF p_Y that is mixture between a Gaussian and a uniform PDF.

An alternative method for sampling arbitrary one-dimensional PDFs relies on first discretizing the PDF into bins and computing the corresponding bin probabilities as probabilities

$$\{p_1, \ldots, p_N\} \qquad \text{where} \qquad p_i = \int_{y_i-\Delta y/2}^{y_i+\Delta y/2} p_Y(z)dz \tag{78}$$

We can then sampling the states $\{y_1, \ldots, y_N\}$ with probabilities $\{p_1, \ldots, p_N\}$. The basic idea to do so is to divide the $[0,1]$ into N bins, with the length of each bin proportional to the probability mass p_k. Then,

[12] We know that the CDF F_Y is non-decreasing. Hence, it is always possible to invert it in appropriate regions of its domain. Wherever the function is not invertible the inverse does not exist. Recall that the CDF may not be even continuous (it is continuous from the right but not from the left). This about the CDF of a PDF mixture made of two uniform in $[0,1]$ and $[3,4]$.

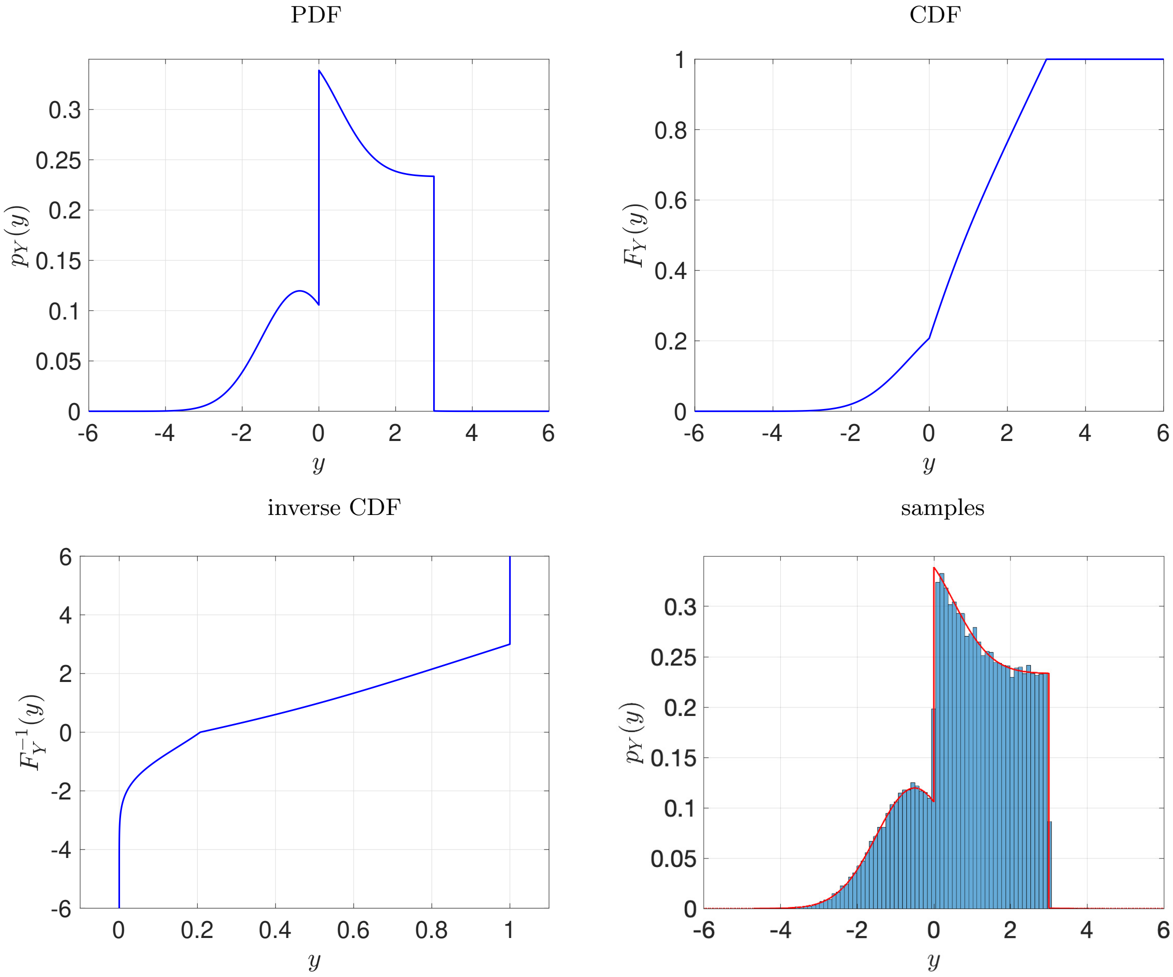

Figure 4: Sampling from arbitrary one-dimensional PDFs. In this example we sample from a given PDF $p_Y(y)$ (mixture between Gaussian and uniform PDFs) by first computing $F_Y(y)$ and $F_Y^{-1}(y)$ numerically, and then using $F_Y^{-1}(y)$ to map samples of a uniform random variable X in $[0,1]$ to samples of the PDF $p_Y(y)$ via the mapping (77).

we draw samples from a uniform distribution in $[0,1]$, and the bins that these values fall into are picked as results. This is also equivalent to rolling a loaded dice with N faces and probabilities given in (78).

Expectation, moments and cumulants

Let $(\Omega, \mathcal{B}, P)$ 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 [13]

$$\mathbb{E}\{g(X)\} = \int_{-\infty}^{\infty} g(x)dF_X(x) = \int_{-\infty}^{\infty} g(x)p_X(x)dx. \tag{80}$$

[13] We do not need to assume the existence of the PDF to define the expectation operator. In fact, a more general expression for (80) is

$$\mathbb{E}\{g(X(\omega))\} = \int_{\Omega} g(X(\omega))dP(\omega). \tag{79}$$

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

$$\mathbb{E}\{g(X)\} = \mathbb{E}\{Y\} = \int_{-\infty}^{\infty} y dF_Y(y) = \int_{-\infty}^{\infty} y p_Y(y) dy. \tag{81}$$

In particular, if we set $g(X) = X^k$ then $\mathbb{E}\left\{X^k\right\}$ are called *moments*[14] of the random variable X

$$\mathbb{E}\left\{X^k\right\} = \int_{-\infty}^{\infty} x^k dF_X(x) = \int_{-\infty}^{\infty} x^k p_X(x) dx. \tag{82}$$

The first few moments of a random variable X are

$$\mathbb{E}\{X\} = \int_{-\infty}^{\infty} x p_X(x) dx \qquad \text{(mean)}, \tag{83}$$

$$\mathbb{E}\left\{X^2\right\} = \int_{-\infty}^{\infty} x^2 p_X(x) dx \qquad \text{(second-order moment)}, \tag{84}$$

$$\mathbb{E}\left\{X^3\right\} = \int_{-\infty}^{\infty} x^3 p_X(x) dx \qquad \text{(third-order moment)}. \tag{85}$$

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

$$M(a) = \mathbb{E}\left\{e^{aX(\omega)}\right\} \tag{86}$$

In fact,

$$M(a) = M(0) + \underbrace{\frac{dM(0)}{da}}_{\mathbb{E}\{X\}} a + \frac{1}{2} \underbrace{\frac{d^2M(0)}{da^2}}_{\mathbb{E}\{X^2\}} a^2 + \cdots . \tag{87}$$

In general,

$$\mathbb{E}\left\{X^k\right\} = \frac{d^kM(0)}{da^k}. \tag{88}$$

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

$$\Psi(a) = \log(M(a)). \tag{89}$$

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

$$\Psi(a) = \Psi(0) + \underbrace{\frac{d\Psi(0)}{da}}_{\mathbb{E}\{X\}} a + \frac{1}{2} \underbrace{\frac{d^2\Psi(0)}{da^2}}_{\mathbb{E}\{X^2\}-\mathbb{E}\{X\}^2} a^2 + \cdots \tag{90}$$

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

$$\langle X \rangle_c = \mathbb{E}\{X\}, \tag{91}$$

$$\left\langle X^2 \right\rangle_c = \mathbb{E}\left\{X^2\right\} - \mathbb{E}\{X\}^2, \tag{92}$$

$$\left\langle X^3 \right\rangle_c = \mathbb{E}\left\{X^3\right\} - 3\mathbb{E}\{X\}\mathbb{E}\left\{X^2\right\} + 2\mathbb{E}\{X\}^3, \tag{93}$$

$$\cdots$$

[14] 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 quantity

$$\left\langle X^2\right\rangle_c = \mathbb{E}\left\{X^2\right\} - \mathbb{E}\left\{X\right\}^2, \tag{94}$$

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

$$\phi(a) = \mathbb{E}\left\{e^{iaX(\omega)}\right\} \tag{95}$$

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

$$\mathbb{E}\left\{X^k\right\} = \frac{1}{i^k}\frac{d^k\phi(0)}{da^k}. \tag{96}$$

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., [6])

$$\phi(a) = \exp\left[\sum_{j=1}^{\infty}\left\langle X^j\right\rangle_c \frac{(ia)^j}{j!}\right]. \tag{97}$$

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

$$\phi(a) = e^{i\mu a-\sigma^2 a^2/2}. \tag{98}$$

This expression can be derived by the taking the Fourier transform of (32), or by using (97). In fact, for Gaussian random variables we have that only the first two cumulants are non-zero, i.e.,

$$\left\langle X\right\rangle_c = \mathbb{E}\left\{X\right\} = \mu, \tag{99}$$

$$\left\langle X^2\right\rangle_c = \mathbb{E}\left\{X^2\right\} - \mathbb{E}\left\{X\right\}^2 = \sigma^2, \tag{100}$$

$$\left\langle X^k\right\rangle_c = 0 \quad \text{for all } k \geq 3. \tag{101}$$

Substituting these expressions into (97) yields (98).

References

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

[2] 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.

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

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

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

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

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

[8] 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.

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