Stochastics as Physics

With Applications to Geophysics and Engineering

Demetris Koutsoyiannis

Stochastics as Physics

With Applications to Geophysics and Engineering

Stochastics as Physics

With Applications to Geophysics and Engineering

Edition 0 (incomplete) Release 0.06 – April 2026.

Author: Demetris Koutsoyiannis, School of Civil Engineering, National Technical University of Athens

Please report problems and suggest corrections to dk@ntua.gr

National Technical University of Athens
Heroon Polytechneiou 9, 15780 Zographou
www.itia.gr

To the eternal memory of those I have loved and lost:
My mother, Aglaia, and my father, Nikolas,
my second mother (in every way that mattered), Gia,
my wife, Annouska,
my daughter-in-law, Katerina,
my sisters, Democratia, Elle, and Ioanna,
my brothers-in-law, Napoleon and Giorgos,
and my nephews, Nikos and Stathis.

May this book carry a lasting flicker of your light.

Contents

Fragment of Heraclitus*	**Original Greek**	**English translation**
DK 22 B30	*κόσμον τόνδε, τὸν αὐτὸν ἁπάντων, οὔτε τις θεῶν οὔτ' ἀνθρώπων ἐποίησεν, ἀλλ' ἦν ἀεὶ καὶ ἔστιν καὶ ἔσται πῦρ ἀείζωον, ἁπτόμενον μέτρα καὶ ἀποσβεννύμενον μέτρα.*	*This world, which is the same for all, no one of gods or men has made. But it always was, is, and will be: an ever-living Fire, with measures of it kindling, and measures quenching.*
DK 22 B90	*πυρὸς ἀνταμοιβὴ τὰ πάντα καὶ πυρὸς τὰ πάντα, ὅκωσπερ χρυσοῦ χρήματα καὶ χρημάτων χρυσός.*	*All things are an interchange for Fire, and Fire for all things, just like goods for gold and gold for goods.*
PC 439d, 440c	*Πάντα ῥεῖ.*	*Everything flows*
PC 401d	*τὰ ὄντα ἰέναι τε πάντα καὶ μένειν οὐδέν.*	*All things move and nothing remains still.*
PC 402a	*δὶς ἐς τὸν αὐτὸν ποταμὸν οὐκ ἂν ἐμβαίης.*	*You cannot step twice into the same river.*
DK 22 B12	*ποταμοῖσι τοῖσιν αὐτοῖσιν ἐμβαίνουσιν, ἕτερα καὶ ἕτερα ὕδατα ἐπιρρεῖ.*	*On those who enter the same rivers, ever different waters flow.*
DK 22 B52	*Αἰών παῖς ἐστι παίζων πεσσεύων. Παιδός ἡ βασιληίη.*	*Time is a child playing, throwing dice. To a child the ruling power belongs.*
DK 22 B8	*Τὸ ἀντίξουν συμφέρον καὶ ἐκ τῶν διαφερόντων καλλίστην ἁρμονίαν καὶ πάντα κατ' ἔριν γίνεσθαι*	*Opposition unites, the finest harmony springs from difference, and all comes about by strife*
DK 22 B67	*Ὁ θεὸς ἡμέρη εὐφρόνη, χειμὼν θέρος, πόλεμος εἰρήνη, κόρος λιμός [τάναντία ἅπαντα].*	*God is day and night, winter and summer, war and peace, surfeit and hunger [all the opposites].*
DK 22 B88	*ταὐτὸ δὲ ζῶν καὶ τεθνηκὸς καὶ ἐγρηγορὸς καὶ καθεῦδον καὶ νέον καὶ γηραίον· ταῦτα γὰρ μεταπεσόντα ἐκεῖνά ἐστι κἀκεῖνα πάλιν μεταπεσόντα ταῦτα.*	*And it is the same thing in us that is alive and dead, awake and asleep, young and old; for the former are shifted and become the latter, and the latter in turn are shifted and become the former.*

* DK 22: Standard Diels-Kranz numbering of Presocratic philosophers, chapter 22 (on Heraclitus of Ephesus), https://heraclitusfragments.com/
PC: Plato's Cratylus, https://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0171%3Atext%3DCrat.%3A

Preface

[to be written at a later stage]

Athens, ??? 2026

Demetris Koutsoyiannis

Acknowledgments

Athens, May 2026

Demetris Koutsoyiannis

Notational conventions

The book follows the *Guidelines for the use of units, symbols and equations in hydrology*[*]. In turn, these guidelines are based on (i) the Système International (SI) brochure[†]; (ii) the ISO 80000-2 Standard, *Mathematical Signs and Symbols to Be Used in the Natural Sciences and Technology*; and (iii) Unicode Technical Report #25, *Unicode Support for Mathematics*.[‡] We list some of the conventions here for the reader's convenience.

Physical dimensions and units

(a) All quantities are dimensionally consistent. In particular, arguments of functions such as exp( ) and ln( ) are dimensionless.
(b) We use s, min, h, and d for second, minute, hour and day respectively. We do not abbreviate week, month or year, which are non-SI units.[§]
(c) Multiplication of units is indicated by a space, e.g. N m, and division either by negative exponents (e.g. $\mathrm{m\,s^{-2}}$) or by use of the solidus (oblique line, e.g. $\mathrm{m/s^2}$); however repeated use of the solidus (e.g. m/s/s) is not permitted.
(d) Prefixes of units such as M (mega = 10^6) and μ (micro = 10^{-6}) have no space between (e.g. μs, MW). According to the SI, the prefix for kilo is lower case k (e.g. km—K is the symbol of the kelvin).
(e) For areas and volumes, we use $\mathrm{m^2}$ and $\mathrm{m^3}$; the hectare (ha) and the litre (L) are also allowed in SI. A million $\mathrm{m^2}$ is denoted as square kilometre ($1\ \mathrm{km^2} = 10^6\ \mathrm{m^2}$). A million $\mathrm{m^3}$ is denoted as cubic hectometre ($1\ \mathrm{hm^3} = 10^6\ \mathrm{m^3}$—not 1 $\mathrm{Mm^3}$ because $1\ \mathrm{Mm^3} = 10^{18}\ \mathrm{m^3}$; note that in SI any power to a unit applies also to the prefix); a billion $\mathrm{m^3}$ is denoted a cubic kilometre ($1\ \mathrm{km^3} = 10^9\ \mathrm{m^3}$).
(f) All units are typeset in upright (Roman) fonts, not italic or bold.
(g) Numerals are also typeset in upright fonts. The symbol for the decimal marker is the dot. To facilitate reading, numbers are divided in groups of three using a thin space (e.g. 12 345.6). (Note that neither dots nor commas are permitted as group separators). A space is used to separate the unit from the number (e.g. 10 m).

Symbols and equations

(a) We prefer single-letter variables (if necessary, with subscripts, e.g. E_{RMS}) over multi-letter ones. Single-letter variables or parameters and user-defined function symbols are italic (e.g. $x, Y, \beta, f(x)$). Multi-letter variables, if cannot be avoided, are typeset in upright, not italic (e.g. RMSE).
(b) Common, explicitly defined, functions are not italic, whether their symbols are single-letter (e.g. $\Gamma(x)$ for the gamma function, $\mathrm{B}(y, z)$ for the beta function) or multi-letter (e.g. $\ln x$, $\exp(x + y)$).
(c) Textual subscripts or superscripts are not italic (e.g. $x_{\max}$, $T_{\min}$ where 'max' and 'min' stand for maximum and minimum, respectively).
(d) Mathematical constants are upright (e.g. $\mathrm{e} = 2.718\ldots$, $\pi = 3.141\ldots$, $\mathrm{i}^2 = -1$). Also, mathematical operators are upright (e.g. $\mathrm{d}x$ in integrals and derivatives, $\Delta\gamma$ for the difference operator on γ).
(e) Vectors, matrices and vector functions are bold and, for single-letter variables, italic. In particular, vectors are usually denoted with lower case letters (e.g. $\boldsymbol{x}$, $\boldsymbol{\omega}$ as vectors; $\boldsymbol{f}(\boldsymbol{x})$ as a vector function of a vector variable) and matrices with upper case letters (e.g. $\boldsymbol{A}$ as matrix; $\boldsymbol{AB}$ as the product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A}^{\mathrm{T}}$ as the transpose of $\boldsymbol{A}$, $\det \boldsymbol{A}$ as the determinant of a square matrix $\boldsymbol{A}$).
(f) We use nested parentheses for grouping (e.g. $\ln(a\,(b + c))$ rather than $\ln[a\,(b + c)]$)
(g) To distinguish between stochastic variables from common variables we use the Dutch convention[**], i.e., we underline the stochastic variables. Further, we use the curly brackets for sets (e.g. $P\{\underline{x} \le x\}$ for a scalar x or $P\{\underline{\boldsymbol{x}} \le \boldsymbol{x}\}$ for a vector $\boldsymbol{x}$; note that the argument of probability (P) is a set, not a number).
(h) We use square brackets for expectations, variances and other operators on stochastic variables (e.g. $\mathrm{E}[\underline{x}]$, $\mathrm{var}[\underline{x}]$, $\mathrm{cov}[\underline{x}, \underline{z}]$; note that $\mathrm{E}[\underline{x}]$ is not a function of $\underline{x}$ and thus it should not be denoted as $\mathrm{E}(\underline{x})$.)
(i) Definitions by mathematical equations are denoted using the symbols '$:=$' and '$=:$' (e.g. to define c as the sum of a and b we write $c := a + b$ or $a + b =: c$).

---

[*] Prepared by D. Koutsoyiannis and H.H.G. Savenije, 2013, doi: 10.13140/RG.2.2.10775.21922

[†] Ninth edition, http://www.bipm.org/en/si/si_brochure/

[‡] http://www.unicode.org/reports/tr25

[§] We avoid 'a' for year, because in SI 'a' is the prefix atto, meaning 10^{-18}; also it is the symbol of an 'are', a non-SI unit whose multiple hectare is accepted in SI ($1\ \mathrm{a} = 100\ \mathrm{m^2}$; $1\ \mathrm{ha} = 100\ \mathrm{a} = 10^4\ \mathrm{m^2} = 1\ \mathrm{hm^2}$).

[**] Hemelrijk, J., 1966. Underlining random variables. *Statistica Neerlandica*, 20(1), pp.1-7.

Main use of single-letter symbols

a	coefficients of stochastic generators	α	time scale parameter in stochastic processes
A		A	as Latin A
b		β	background measure density
B	as a standard, the beta function B(,)	B	as Latin B
c	autocovariance	γ	climacogram; as a standard, the Euler's constant, $\gamma = 0.577216\ldots$
C		Γ	cumulative climacogram; as a standard, the gamma function $\Gamma(\ )$ and the incomplete gamma function $\Gamma_\alpha(\ )$.
d	as a standard, the differential operation d	δ	
D	time unit, discretization time step	Δ	
e	as a standard, $e = 2.71828\ldots$	ε	dimensionless location parameter in distributions
E	as a standard: the expectation, E[ ]	E	
f	probability density function	ζ	dimensionless shape parameter (lower-tail index) in distributions; as a standard, the Riemann zeta function $\zeta(\ )$
F	probability distribution function	Z	as Latin Z
g		η	time lag, dimensionless
G		H	as Latin H
h	time lag, dimensional	θ	angle (phase); also ombrian parameter
H	Hurst parameter (also, $H_p \coloneqq \sum_{i=1}^{p} 1/i$ and $H_p^{(a)} \coloneqq \sum_{i=1}^{p} 1/i^a$ are the pth harmonic numbers of orders 1 and a, respectively).	Θ	bias correction factor
i		ι	
I	identity matrix (in bold); as a standard, the indicator function I_A	I	as Latin I
j		κ	time scale, dimensionless (also cumulants)
J		K	as Latin K
k	time scale, dimensional	λ	state scale parameter in distributions
K	K-moment	Λ	Λ-coefficient (for K-moments)
l		μ	mean, moment
L	Length of observation period (also life span)	M	as Latin M
m	moment	ν	similar to n (size of sample or vector)
M	Mandelbrot parameter	N	as Latin N
n	size of sample or vector	ξ	dimensionless shape parameter in distributions (upper-tail index)
N	size of sample or vector	Ξ	Ξ-coefficient (for K-moments)
o	little-O notation	o	as Latin o
O	big-O notation	O	as Latin O
p	moment order	π	as a standard, $\pi = 3.14159\ldots$
P	probability	Π	set of partitions
q	moment order	ρ	standardized cross-climacogram
Q		P	as Latin P
r	correlation coefficient	σ	standard deviation
R		Σ	sum, σ-algebra
s	power spectrum	τ	time, dimensionless
S		T	as Latin T
t	time, dimensional	υ	structure function
T	return period (as a superscript, 'T': transpose)	Y	as Latin Y
u		φ	entropy production
U		Φ	Entropy
v	white noise process	χ	
V		X	as Latin X
w	frequency	ψ	climacospectrum; as a standard, the digamma function $\psi(x)$ or the polygamma function $\psi^{(n)}(x)$
W	as a standard, the Lambert W function $W_k(x)$	Ψ	odds function
x, y, z	stochastic variables and processes or time series	ω	frequency, dimensionless
X, Y, Z	cumulative stochastic processes or time series	Ω	Ground set

Chapter 1. A historical and philosophical introduction

1.1 From a clockwise universe to a stochastic* cosmos

In 1926, Albert Einstein, in a letter to Max Born, wrote what has become one of his most famous aphorisms:

> *Jedenfalls bin ich überzeugt, daß der nicht würfelt* (*I, at any rate, am convinced that He [God] does not throw dice*).

This reflects adherence to an earlier shaped belief of a clockwork universe and a philosophical view of *determinism*, widely accepted in modern science before and after Einstein till the present day. This whole idea was introduced in the 16th century, together with the development of mathematical concepts in natural philosophy (what today we call science), and was further processed in the works of the giants of modern science Johannes Kepler, Galileo Galilei and René Descartes (Figure 1.1). Determinism was perfected by the French mathematician and astronomer Pierre-Simon Laplace, and was reflected in the famous *Laplace's demon*, a hypothetical entity that, knowing the precise location and momentum of every atom in the universe at present, can deduce the future and the past using Newton's laws.

Laplace's demon is a manifestation of the deterministic perception of the world, in which the roots of uncertainty about the future are subjective—stemming from our ignorance of the precise present state or from inadequate models and methods. In this view, eliminating uncertainty is merely a matter of gathering better data and constructing better models. Essentially, the concept of the demon impoverishes the universe by dissolving the significance of time: it reduces four-dimensional space-time to something effectively three-dimensional, since a single moment suffices to determine—and thus fully encode—everything that has happened or will ever happen across the entire temporal expanse.

Figure 1.1 Founders of the deterministic worldview (from left to right): Johannes Kepler (1571-1630), Galileo Galilei (1564-1642), René Descartes (1596-1650) and Pierre-Simon Laplace (1749-1827). (Images publicly available by Wikipedia.)

* The meaning of the term stochastic is clarified in Digression 1.C.

Amazingly, however, the father of these very laws—Newton himself (Figure 1.2)—was acutely aware of the universe's fragility and did not embrace a vision of blind, self-sufficient mechanism. In his *Opticks* (Query 31) he wrote (spelling modernized):

> *For while comets move in very eccentric orbs in all manner of positions, blind fate could never make all the planets move one and the same way in orbs concentric, some inconsiderable irregularities excepted which may have arisen from the mutual actions of comets and planets on one another, and which will be apt to increase, till this system wants a reformation*.

This passage clearly reveals Newton's recognition of the solar system's complexity and instability over long timescales. He saw such fragility not as a flaw in divine design but as positive evidence for God's existence and active governance—rejecting Leibniz's thesis that a perfect Creator would necessarily fashion a flawless, self-sustaining world requiring no further intervention. In other words, Newton argued for the necessity of periodic divine reformation to preserve order in a fragile cosmos. This perspective essentially envisages an ever-evolving four-dimensional universe, rejecting Laplace's static, three-dimensional caricature in which time is redundant.

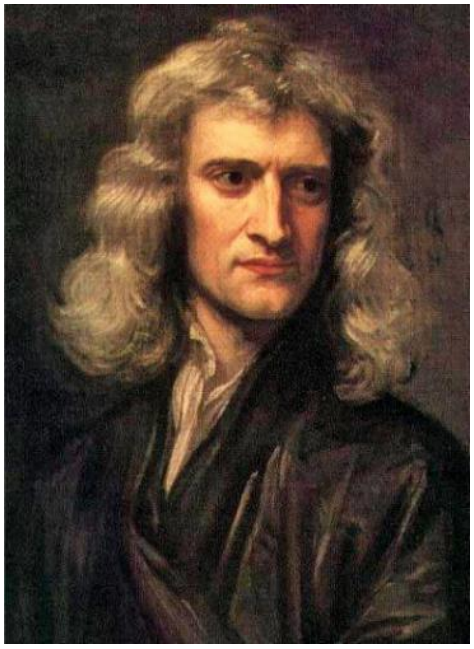

Figure 1.2 Isaac Newton (1643-1726). (Image publicly available by Wikipedia.)

However, it took centuries (namely up to the end of the 19th century) before the seemingly almighty determinism of the 17th century received strong blows in several fields of science, including the following.

- *Statistical physics* (cf. Boltzmann; Figure 1.3) used the probabilistic concept of entropy (which is nothing other than quantified uncertainty, defined within the probability theory; see below) to explain fundamental physical laws (most notably the Second Law of thermodynamics). This led to a new understanding of natural behaviours and to powerful predictions of macroscopic phenomena.

- The *dynamical systems* theory (cf. Poincare; Figure 1.3) has shown that uncertainty can emerge even from pure, simple and fully known deterministic (chaotic) dynamics, and cannot be eliminated.
- *Genetics* (cf. Mendel; Figure 1.3) and *evolutionary biology* have emphasized the importance of stochasticity (e.g., in gametes fusion, selection and mutation procedures, and environmental changes) as a driver of evolution.
- *Quantum theory* (cf. Heisenberg; Figure 1.3) has emphasized the intrinsic character of uncertainty and the necessity of probability in the description of Nature.
- Developments in *mathematical logic*, and particularly *Gödel's incompleteness theorem*, challenged the almightiness of deduction (inference by mathematical proof), which is the mathematical analogue of determinism. Ironically, Gödel (Figure 1.3) anticipated by one day (in 1930) David Hilbert who pronounced the opposite with his famous aphorism (also inscribed in his tombstone at Göttingen) "*Wir müssen wissen, wir werden wissen*" ("*We must know, we will know*").

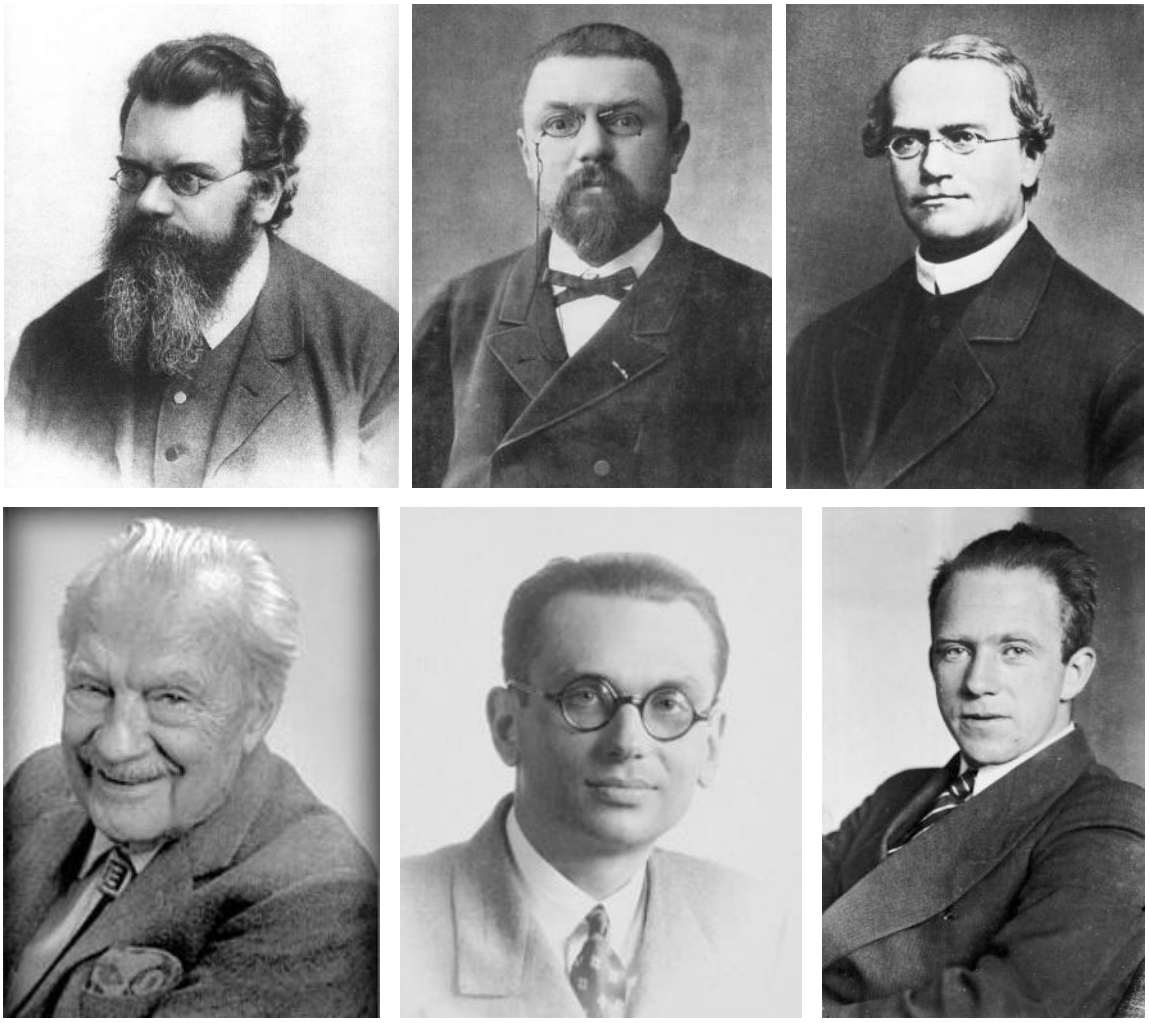

Figure 1.3 Giants of science who gave blows to scientific determinism (**from top left, clockwise**) Ludwig Boltzmann (1844–1906), Henri Poincaré (1854–1912), Gregor Mendel (1822–84); Werner Heisenberg (1901–1976), Kurt Gödel (1906–1978), Nicholas Metropolis (1915–1999). (Images publicly available by Wikipedia.)

- Developments in *numerical mathematics* (cf. Metropolis; Figure 1.3) highlighted the effectiveness of stochastic methods in solving even purely deterministic problems, such as *numerical integration* in high-dimensional spaces and *global optimization* of non-convex functions. Stochastic optimization techniques, e.g., evolutionary algorithms or simulated annealing, are in effect the only feasible solution in complex problems that involve many local optima.

However, roots in common sense and philosophy of a worldview that is not deterministic go far back in the past. For example, the Greek mythology included a goddess of chance or randomness, named *Tyche*, whom Romans later identified with their goddess *Fortuna*. Archaeological evidence has brought to light lots of ancient dice—emblems of randomness—such as the ancient Greek ones shown in Figure 1.4. Much older dice (up to 5000 years old) have been found in Asia (Iran, India).

Figure 1.4 Dice of the Greek antiquity (mostly of the early 6th century BC) from Greek archaeological sites, namely: (**left**) Kerameikos Ancient Cemetery Museum, Athens (photo by author); (**next two**) Archaeological Museum of Vravrona (photos by author); (**rightmost two**), Greek National Archaeological Museum, (www.namuseum.gr/object-month/2011/dec/dies_b.png, www.namuseum.gr/object-month/2011/apr/7515.png,). The dice are from terracotta and their size vary, reaching about 10 cm (**middle die**), except for the **rightmost**, which is from bronze (1.6 cm).

In philosophy, Heraclitus (Figure 1.5) was the first who highlighted randomness, and could thus be regarded the father of *indeterminism*, as well as of *dialectics*. Indeterminism is a philosophical belief contradictory to determinism, in which uncertainty is a structural element of Nature and, thus, cannot be eliminated. Interestingly, 2500 years before Einstein wrote what we quoted in the beginning of this section, Heraclitus made just the opposite claim in a masterly and poetic aphorism, yet much less known in the scientific community (Fragment DK 22 B52, see also the motto in the beginning of the book):

Αἰών παῖς ἐστι παίζων πεσσεύων (*Time is a child playing, throwing dice*)

More than a century after Heraclitus, Aristotle made scientific contributions that expand to all aspects of knowledge (cf. Koutsoyiannis and Mamassis, 2021), including physics (e.g., the principle of mass conservation; see section 0), geophysical sciences (in particular meteorology and hydrology), and epistemology. The Aristotelian logic offers a powerful instrument to distinguish sense from nonsense as well as deduction from induction, and the relative validity of the inference based on each of these two methods

(see Digression 4.A). Most relevant to our theme and of great importance in modern science, particularly in physics and stochastics, is the Aristotelian dipole *potentiality* (*δύναμις*, Latin *potentia*) vs. *actuality* (*ένέργεια*, Latin *actualitas*), formulated in his books *Physics*, *Metaphysics*, *Nicomachean Ethics* and *De Anima*. Practically, the idea behind the dipole is that several outcomes can be produced by a specified cause, while in deterministic thinking only one outcome is possible (albeit often difficult to predict which one).

Epicurus was another Greek philosopher who involved randomness in the explanation of Nature. While he kept Democritus' idea of atoms (ἄτομα) as constituents of matter, he rejected determinism and assumed that the motion of atoms is random.

The first to utilize the Aristotelian dipole *potentiality* vs. *actuality* in modern science, namely in quantum physics, was Heisenberg (1962):

> *The most important of these* [features of the interpretation by Bohr, Kramers and Slater] *was the introduction of the probability as a new kind of "objective" physical reality, the "potentia" of the ancients such as Aristotle; it is, to a certain extent, a transformation of the old "potentia" concept from a qualitative to a quantitative idea.*

This idea of Heisenberg was quoted by Popper (1982), who fully incorporated it into his philosophical system, further extending it to claim, for example, that:

> *Both classical physics and quantum physics are indeterministic.*

Popper is the main modern philosopher who, being fully aware of modern physics, theorized modern indeterminism also connecting it with the notion of probability, which he regarded as the extension (quantification) of the Aristotelian *potentia* (*δύναμις*).

Figure 1.5 Ancient and modern philosophers who founded or theorized the indeterministic worldview: (from left to right) Heraclitus (535–475 BC), Aristotle (384 – 322 BC), Epicurus (341–270 BC), and Karl Popper (1902–1994). (The source for the image of Heraclitus—a depiction in the back facet of a coin—is Visconti, 1817. All other images are publicly available by Wikipedia.)

More recently this Aristotelian dipole has been proposed by several scientists and philosophers, independently of Popper, as a simpler, more comprehensible and more effective interpretation of quantum physics (Jaeger, 2017, 2018; Kastner et al., 2018; Driessen, 2019; Sanders, 2018). In particular, Kastner et al. (2018), building on Heisenberg's (1962) idea, proposed an ontological dualism of actualities (*res extensa*) and potentialities (*res potentia*), with the latter not bounded by space–time constraints and being transformed to the former by an acausal process of *potentia* (*δύναμις*).

Apparently, the dilemma of determinism vs. indeterminism is not just an issue of philosophical belief. It affects our perception and orientation in scientific inquiry, as well as our decisions and actions. If the world were deterministic, decision making would be trivial and, by now, would be undertaken by computers and robots. Because the world is better viewed as indeterministic, decision making remains a human task, linked with responsibility.* This is vividly expressed by Julius Caesar's famous words when crossing the Rubicon in 49 BC, reportedly uttered in Greek:

Ἀνερρίφθω κύβος (*Let the die have been cast*).†

1.2 Logico-philosophical premises

Despite the blows given to scientific determinism since almost 150 years ago, it remains the main line of thought in the scientific community—in physics in particular. Its influence is so large that many think that *first principles* in physics are only deterministic—mostly mechanistic. To see that this is a wrong perception, it suffices to think of the Second Law in thermodynamics, perhaps the most important physical law, which relies on *entropy*—and, as we will see (section 2.3), entropy is a purely stochastic concept.

Other dominant canons that may misguide scientific inquiry are the reductionist approach and the adherence to equations. According to the former approach, the laws of complex physical systems can be inferred by synthesizing detailed representations of their elements. This cannot be true or useful: for example, while it is true that a book consists of several types of molecules, we cannot infer its content by examining the behaviour of the molecules. On the other hand, equations are useful if they correspond to reality, but they cannot handle all problems and they do not constitute the most powerful tool that mathematics offers to the study of the physical world.

This book, as evident from the considerations in section 1.1, does not follow a deterministic approach. While it respects deterministic equations of simple systems, it does not rely merely on them. Rather, it recognizes the fundamental character of uncertainty in Nature and uses stochastic approaches. While it makes use of important

---

* Acknowledging the expanding role of artificial intelligence (AI) in decision support and automation, true accountability—e.g., for violations carrying imprisonment or equivalent moral weight—still requires a responsible human actor, as no bot can bear punishment or ethical culpability in the way persons do.

† This is known from Plutarch's *Life of Caesar*, 32, and the exact quotation (from https://www.perseus.tufts.edu/hopper/text?doc=Plut.+Pomp.+60.2) is this: *Ἑλληνιστὶ πρὸς τοὺς παρόντας ἐκβοήσας, «Ἀνερρίφθω κύβος», διεβίβαζε τὸν στρατόν.* (*Declaring in Greek to those present in a loud voice, 'Let the die have been cast,' he led the army across.*). Later, the phrase rendered in Latin by Suetonius as *Iacta alea est* (*The die has been cast*). The phrase *ἀνερρίφθω κύβος* is traced back to the Greek playwright Menander (Μένανδρος, c. 342 – 290 BC) of the Athenian New Comedy. It appears in his play *Ἁρρηφόρος* (*Arrhephoros*, i.e. the bearer of ritual objects), or *Αὐλητρίς* (*Auletris*, i.e. the flute-girl), preserved only in fragments (http://www.poesialatina.it/_ns/greek/testi/Menander/Fragmenta.html) via Athenaeus (Deipnosophistae 13.559e). From the surviving excerpt:

Character A: *Οὐ γαμεῖς ἐὰν νοῦν ἔχῃς, τοῦτον καταλιπὼν τὸν βίον. Γεγάμηκα γὰρ αὐτός, διὰ τοῦτό σοι παραινῶ μὴ γαμεῖν.* (*You never marry, if you're smart, for the rest of your life. I've married myself, so that's why I advise you not to.*)

Character B: *Δεδογμένον τὸ πρᾶγμ', ἀνερρίφθω κύβος.* (*The matter has been decided; let the die have been cast.*)

physical laws expressed in the form of equations, it also highlights the variational mathematical principles and the implied extremization approach, which is more powerful and more natural than an equalization approach. Finally, instead of the reductionism of seeking detailed and inflationary representations of Nature, the book invokes the principle of parsimony and emphasizes the need of macroscopization (by removing details), inevitably accompanied by stochastics.

Many scientists contrast physics with statistics as if the two were incompatible. Such perception is rather incompatible with the important developments in science over the last 150 years, as summarized in section 1.1. Here we use a scientific field that encompasses probability and statistics, while being wider than the two (see section 1.5): the field of *stochastics*.

Digression 1.A: The simplicity of light trajectory (Fermat's principle)

Light trajectory can be viewed is several ways but here we will examine the simplest, which is relevant macroscopically. (A more detailed way in a microscopic setting is discussed in section 5.4). We use a simple example that can help us to see the ontological basis of the principle of parsimony. With reference to the left panel in Figure 1.6, we readily understand that light follows the simplest path from a candle A to the observer's eye B, which is a straight line (the red line) and not other more complex ones (e.g., the black lines ACB, ADB). If it followed those complex paths, the observer would not see the candle but would sense light coming from everywhere.

This observation leads us to formulate a law in this way: *Light follows the shortest possible path from A to B*. Let us call this formulation *attempt 1*. This is parsimonious law for a parsimonious natural behaviour. However, further investigation will show that this law is not correct: It is simpler than "as simple as possible".

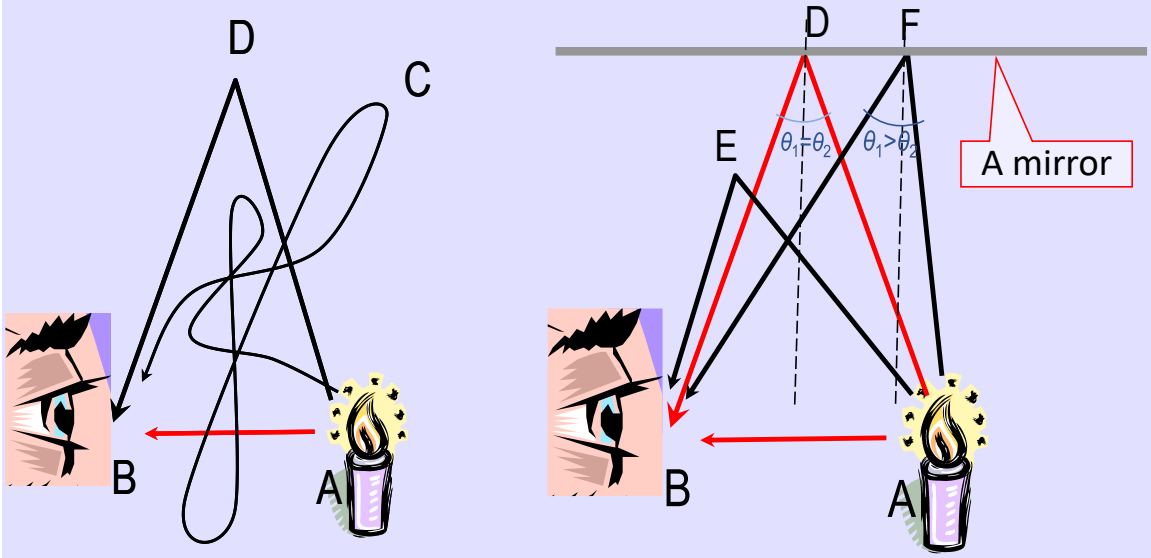

Figure 1.6 (**left**) The light from the candle A to an observer's eye B follows the simplest straight-line path AB (the red line) and not other more complex ones (e.g., the black lines ACB, ADB). (**right**) After putting a mirror at D, parallel to AB, the light follows the broken-line path ADB in addition to the straight-line path AB.

We can quantify and justify the law in a quick-and-dirty manner assuming that light can travel from A to B along a broken line with a break point F with coordinates (x, y). This is not restrictive: we can add a second, third, … break point (homework).

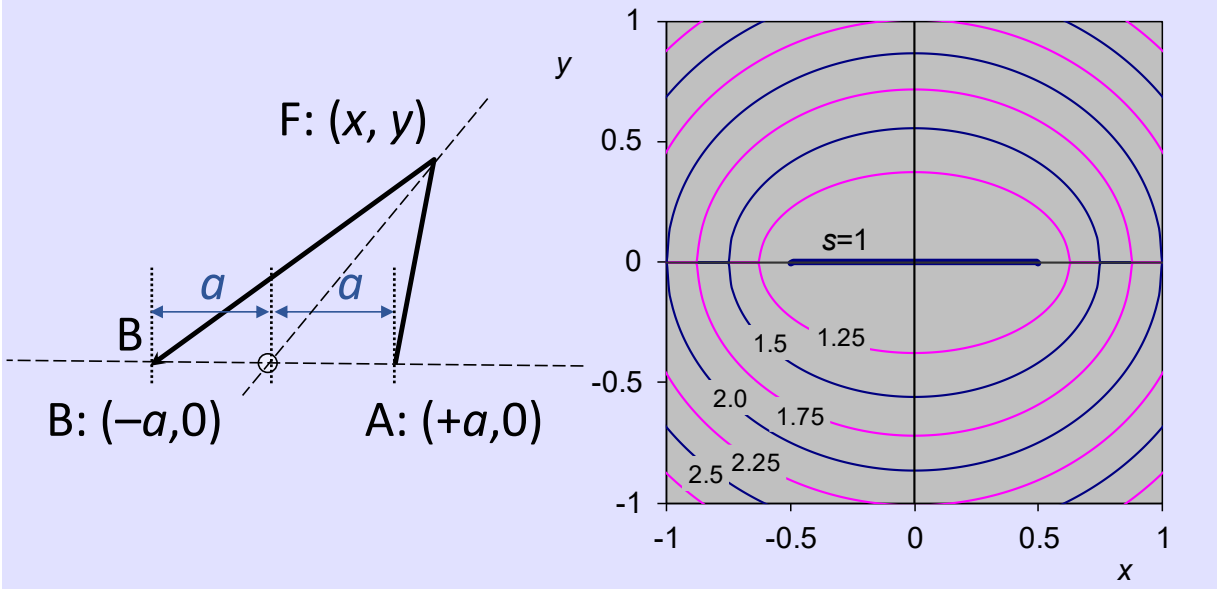

Figure 1.7 (**left**) Notation for a broken-line light path with one break point F with coordinates (x, y). (**right**) Iso-distant lines $s(x, y)$ for values (x, y) between –1 and 1 and for $a = 0.5$.

We observe that the mirror has imposed an inequality constraint to possible paths (by disallowing light to go through it) and thus generated a second minimum in the 'shortest path' problem—a local minimum. This allows us to reformulate the law in a *attempt 3*: *The paths followed by light have minimum length* (either global or local minimum). This is a parsimonious law, known as the principle of Heron of Alexandria (after the Greek physicist and engineer of 1st cent. BC or 1st cent. AD).

With reference to Figure 1.7 (left), the travel distance is

$$s(x, y) = \text{AF} + \text{FB} = \sqrt{(x - a)^2 + y^2} + \sqrt{(x + a)^2 + y^2}$$

Figure 1.7 (right) shows iso-distant lines, from which it becomes clear that any point F that lies in the straight-light segment AB yields minimal distance s. In no other point (x, y) does a maximum or minimum of s appear. We can confirm this by calculating the partial derivative $\partial s(x, y)/\partial y$ and equating it to zero; this readily results in $y = 0$.

Now, let us put a mirror parallel to AB as shown in the right panel in Figure 1.6. Now the light follows both red paths from A to B (AB, ADB)—but not other (the black) ones (e.g. AEB, AFB). The previous formulation of the law is no longer valid. Based on high-school knowledge (recalling the canon of equality of the angles of incidence and reflection), we could replace our *attempt 1* with the following *attempt 2*: *Light follows the shortest path, but when there is a mirror, it also follows a second path with a reflection by the mirror such that the angle of incidence equals the angle of reflection.* Apparently, this is a wordy law, not parsimonious, and reflects an equalizer thinking.

But again, the law as formulated in *attempt 3* is not good enough. To see this let us replace the flat mirror with a half-cylinder mirror, as in the right panel of Figure 1.8. Now the light follows several paths from A to B, including the red lines AB, ACB, ADB—but not the black ones, e.g. AEB, AFB. (There should be other feasible paths with two or more break points, but we ignore them for our illustration). Again, the mirror introduced an inequality constraint in the optimization: the point F should not be behind the mirror. Let us focus on the path ADB (Figure 1.8 right) and compare it with the corresponding path in the case of a flat mirror (Figure 1.8 left). They look identical, but there is a big difference: in the flat mirror, point D is a local minimum but in the half-cylindrical mirror it is a local maximum. This is clarified in the graphs of Figure 1.9. It is clearly seen that AB is the global minimum, ACB is a local minimum and ADB is a local maximum.

Hence, we should modify the law formulation in *attempt 3* to *attempt 4*, which reads: *The paths followed by light have extreme length* (either global or local minimum or maximum). Note that the points of local optima emerge on the mirror surface—the curve where the constraint is

binding. This shows that Nature is a skilful extremizer, as she finds all local minima and maxima (put many mirrors to see lots of paths materializing). Failure to observe this makes things difficult to explain, as indicated for instance in the debate by Gaertner (2003) and Schoemaker (2003).

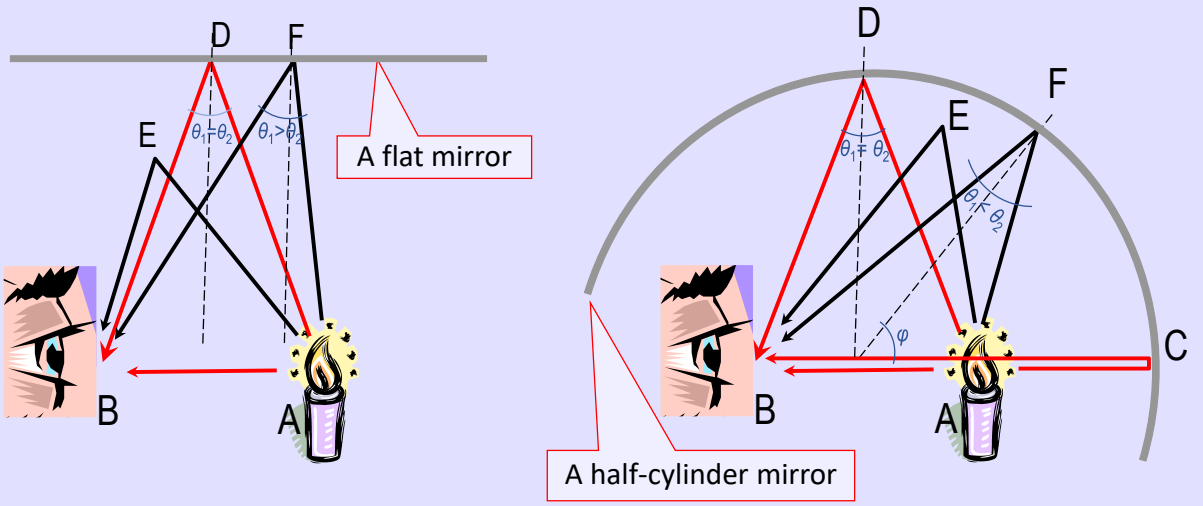

Figure 1.8 Trajectories of light from the candle A to an observer's eye B in the case of (**left**) a flat mirror and (**right**) a half-cylinder mirror.

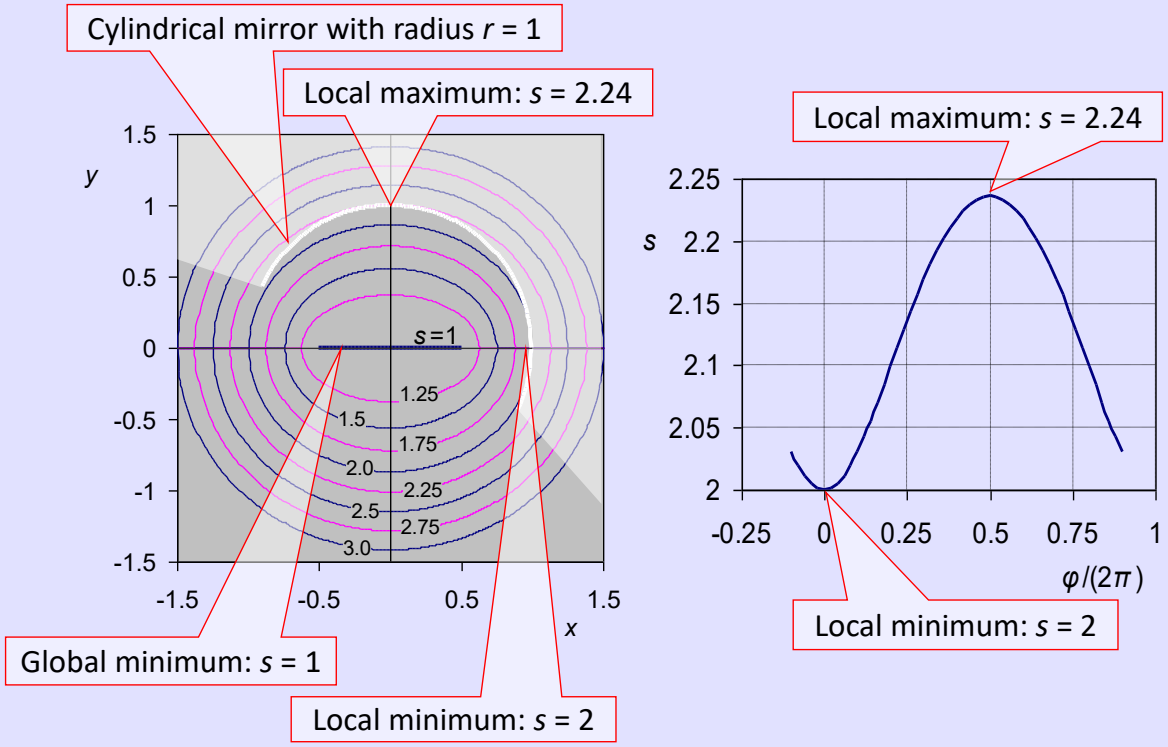

Figure 1.9 (**left**) Iso-distant lines $s(x, y)$ for the case of a half-cylinder mirror. (**right**) Variation of the length $s(x, y)$ of path AFB, as a function of the angle φ shown in Figure 1.8.

But again, the formulation in *attempt 4* does not make a perfect law. Specifically, refraction (Figure 1.10) makes clear that light does not always follow the shortest (straight line) path. This is related to the fact that the light speed in liquids is smaller than in air. Thus, the broken line ACB, rather than the straight-line AB, has the least travel time. Let us determine the point C so as to minimize the total travel time.

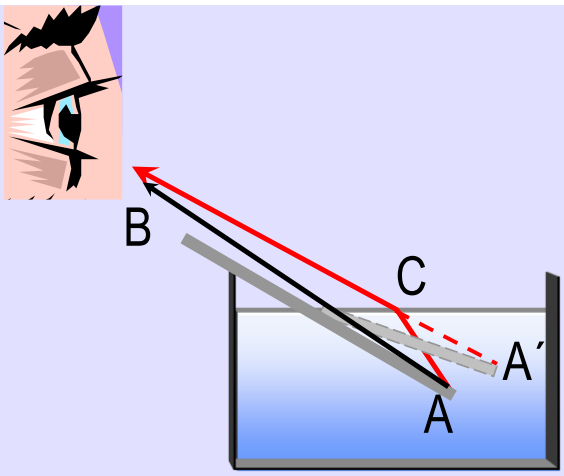

Figure 1.10 Sketch of refraction of a straight-line object partly submerged in water.

Assuming an x axis at the level of the water (so that $y_C = 0$) and denoting the light speed c_A and c_B in the water and air, respectively, the travel time is:

$$t_{ACB} = \frac{\sqrt{(x_A - x_C)^2 + y_A^2}}{c_A} + \frac{\sqrt{(x_C - x_B)^2 + y_B^2}}{c_A}$$

To extremize t_{ACB} we take its derivative with respect to x_C and equate it to zero. This yields:

$$\frac{1}{c_A} \frac{x_A - x_C}{\sqrt{(x_A - x_C)^2 + y_A^2}} = \frac{1}{c_A} \frac{x_C - x_B}{\sqrt{(x_C - x_B)^2 + y_B^2}}$$

We notice that the rightmost fractions in the two sides are the sines of the angles of incidence and refraction. Thus, extremization of the travel time results in the well-known standard result about refraction, expressed in terms of sines of angles.

However, if we kept that standard expression, we would have a non-parsimonious law. Instead, we will make our *attempt 5* to formulate the law as: *Light follows paths that have extreme travel time*. This is our final law, known as the Fermat's principle, but here we corrected it by substituting *extreme* for the standard expression *minimal*. With respect to Heron's principle, we have replaced length with travel time. This does not affect our previous results, in which the light speed was constant.

In conclusion, our example shows that Nature is indeed parsimonious (ontological parsimony). Our final law is also parsimonious (epistemological parsimony), reflecting the parsimony of Nature.

1.3 The principle of parsimony

The *principle of parsimony*, also known as *principle of simplicity*, *principle of economy*, or *Ockham's razor*, advises us to prefer the simplest theory among those that describe reality (e.g., fit the data) equally well. An example of a parsimonious natural law is: "dogs bark". Examples of non-parsimonious—yet not untrue—laws are: "black, white and spotted dogs bark"—or "dogs bark on Mondays, Wednesdays and Fridays".

Intuitively, the law "dogs bark" does not exclude that a particular dog is mute. We should not understand it as "there is no dog that does not bark". In other words, laws of complex systems (e.g. the biological system "dog") are necessarily probabilistic in nature: "dogs bark" means "any dog is very likely to bark". Failure to recognize the probabilistic

character of parsimony in complex systems may create confusion (see e.g. Courtney and Courtney, 2008, and the "all crows are black" example).

Like many aspects of modern scientific method, the principle of parsimony was introduced by Aristotle. Specifically, in his treatise *Posterior Analytics* (I, 25) he stated:

> *We may assume the superiority, other things being equal, of the demonstration which derives from fewer postulates or hypotheses or propositions.*[*]

and in his treatise *On the Heavens* (III, 4) he stated:

> *Obviously, it is much better to assume a finite number of principles, as few as possible yet sufficient to prove what has to be proved, like in what mathematicians demand.*[†]

The principle was reworked by Medieval philosophers: Robert Grosseteste (c. 1168-1253), Thomas Aquinas (c. 1225-1274), and William of Ockham (c. 1285-1347) who formulated it as:

> *Plurality is not to be posited without necessity*.

Ockham's formulation has assigned the principle its most popular name. A comprehensive analysis of the history and philosophy of parsimony, and of the scientific method has been offered by Gauch (2003).

Ockham put parsimony as an epistemological principle for choosing the best theory. However, earlier philosophers, from Aristotle to Grosseteste, had interpreted parsimony also as an ontological principle, thus expecting natural laws to be simple. In other words, Nature should be naturally parsimonious.

It is then natural to try to build parsimonious models for natural processes. Simple systems can be parsimoniously modelled by deterministic approaches. In complex systems parsimony should necessarily be combined with stochastic approaches. While recent mainstream research invested hopes in detailed approaches by building complicated deterministic models (cf. climate models), comparisons of complicated models with parsimonious ones indicate that the latter:

- facilitate insight and comprehension;
- improve accuracy, efficiency and predictive capacity; and
- require fewer data to achieve the same accuracy as the former.

In other words, parsimonious formulations and solutions to problems are more reasonable, rational, and easier to apply and monitor in practice (see more information in Gauch, 2003, and some examples in Koutsoyiannis, 2009).

From modern scientists' views on parsimony, it is useful to study the following formulation, attributed to Albert Einstein[‡]:

---

[*] *Ἔστω γὰρ αὕτη ἡ ἀπόδειξις βελτίων τῶν ἄλλων τῶν αὐτῶν ὑπαρχόντων, ἡ ἐξ ἐλαττόνων αἰτημάτων ἢ ὑποθέσεων ἢ προτάσεων* (Αναλυτικά Ύστερα, I, 25).

[†] *Φανερὸν ὅτι μακρῷ βέλτιον πεπερασμένας ποιεῖν τὰς ἀρχὰς, καὶ ταύτας ὡς ἐλαχίστας πάντων γε τῶν αὐτῶν μελλόντων δείκνυσθαι, καθάπερ ἀξιοῦσι καὶ οἱ ἐν τοῖς μαθήμασιν* (Περί Ουρανού, III, 4).

[‡] See: https://quoteinvestigator.com/2011/05/13/einstein-simple/.

Everything should be made as simple as possible, but not simpler.

It is relevant to ask: What does "as simple as possible" mean and how can it be quantified in mathematical terms? The traditional mathematical approach to physics is based on writing equations, which mostly express conservation laws. However, these govern just a few quantities:

- mass (scalar equation);
- linear momentum (vector equation);
- angular momentum (vector equation);
- energy (scalar equation);
- electric charge (scalar equation).

On the other hand, to find states or paths which are "as simple as possible", it seems more natural to formulate the problem in terms of using optimization rather than equations. Mathematically, *extremizing* is much more powerful than *equating*. Indeed, a system of equations

$$\boldsymbol{g}(\boldsymbol{s}) = \boldsymbol{0} \tag{1.1}$$

where $\boldsymbol{s}$ is a vector of variables and $\boldsymbol{g}$ is a vector function, can work if the number of equations equals the number of unknowns. However, a single (scalar) extremizing expression such as

$$\text{maximize } f(\boldsymbol{s}) \tag{1.2}$$

can work irrespective of the number of unknowns: it is equivalent to as many equations as needed. In other words, the mathematical representation of the phrase "as simple as possible" should be sought in extremization terms.

It is also relevant to think what the mathematical meaning of the phrase "but not simpler" is. One may think that this puts some constraints represented in terms of a vector function $\boldsymbol{h}(\boldsymbol{s})$ and some thresholds $\boldsymbol{h}_{\mathrm{L}}$ and $\boldsymbol{h}_{\mathrm{U}}$ for this function. Hence, a general line of thought would follow the following motif:

$$\begin{aligned} &\text{maximize } f(\boldsymbol{s}) \\ &\text{subject to } \boldsymbol{h}_{\mathrm{L}} \leq \boldsymbol{h}(\boldsymbol{s}) \leq \boldsymbol{h}_{\mathrm{U}} \end{aligned} \tag{1.3}$$

where the inequality constraints reduce to equalities in the case that $\boldsymbol{h}_{\mathrm{L}} = \boldsymbol{h}_{\mathrm{U}}$ (for some of the vector function coordinates).

All that is about the epistemological part, i.e., the models—the mathematical constructions—that we make. But as we discussed, parsimony may also be viewed as an ontological principle and hence the formalization in equation (1.3) represents the way that Nature works. Were Nature not parsimonious, it would be difficult to understand her, and life would be hard. In this respect, we may say that Nature per se is an *extremizer*—not an *equalizer*. This is illustrated by two examples, the simplicity of light trajectory in Digression 1.A, and the simplicity of the trajectory of a body in Digression 1.B. Despite the apparent diversity in the natural behaviours in the two cases, eventually they can be fully described by a single variational principle, the *principle of extremal action*.

Digression 1.B: The simplicity of trajectory of a body (principle of extremal action)

Our previous example is the easiest possible as light has constant velocity in a certain medium. There is no acceleration. However, if we have a body with mass and there is gravity, then we also have acceleration. Let us examine the motion of such a body, e.g., the tennis ball shown Figure 1.11.

To solve this problem solution we need to generalize Fermat's principle to the *principle of extremal action* (also known as the principle of *least* or *stationary action*), one of the most elegant and universal laws of classical dynamics. It states that Nature selects, among all the possible ways a system can evolve from one configuration to another, the path that makes a certain quantity called the *action* stationary—usually, but not exclusively, a minimum.

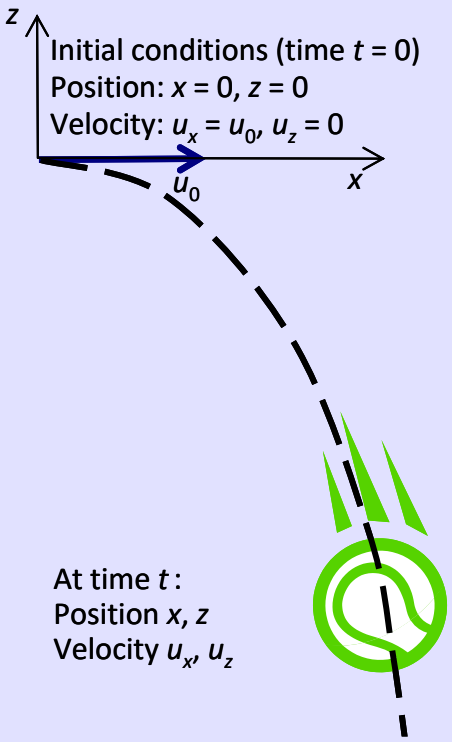

Figure 1.11 Sketch of the motion of a tennis ball hit at time $t = 0$ with a horizontal force so that it have velocity $u_x = u_0, u_z = 0$.

The conceptual roots of this idea trace back to Gottfried Wilhelm Leibniz (1646–1716), who laid important philosophical groundwork. Leibniz's metaphysics emphasized that the actual world, chosen by God from all possible worlds, achieves the maximum of perfection, variety, and order with the minimum "effort" or "resistance." His teleological "principle of the best" (or principle of optimality) suggested that natural processes follow paths of greatest efficiency or simplicity, minimizing unnecessary expenditure. Leibniz also introduced *vis viva* (mv^2, a precursor to kinetic energy; Hecht, 2016) and proposed that bodies follow optimal paths minimizing integrals involving kinetic quantities (e.g., time integrals of vis viva in some contexts). These ideas prefigured variational principles, though he did not publish a precise mathematical formulation for mechanics comparable to later versions. Some historical accounts credit him with early suggestions that optimal motion minimizes certain action-like quantities, and a 1751 controversy even claimed (disputedly) that Leibniz had anticipated the principle in a 1707 letter.[1]

A more precise and geometrically intuitive account appeared in the work of Pierre-Louis Moreau de Maupertuis (in 1744) and Leonhard Euler (also in 1744), building on similar foundations and on earlier optical ideas like Fermat's principle. Considering a particle travelling from point A to point B in space (fixed spatial endpoints), the action is the quantity

$$S_0 = \int m\, v \, \mathrm{d}s$$

where ds is an infinitesimal distance element along the path, v is the speed of the particle at that point, and m is the (constant) mass. Among all conceivable paths connecting these two points, Nature chooses the one that *minimizes* the quantity

This integral can be thought of as a *momentum-weighted path length* or a quantification of "accumulated effort." Like light in Fermat's principle, the particle prefers routes where it can move quickly over distance: high v contributes more per unit length, but to minimize the total integral, the path arranges itself to favour regions of high speed and minimize time spent in slow regions.

Under the assumption of *energy conservation*, the constant energy along the true path will be

$$E = T + V$$

with the two terms denoting kinetic and potential energy, respectively, where for the two-dimensional setting we examine,

$$T = \frac{1}{2}mu^2 = \frac{1}{2}m(u_x^2 + u_z^2), \qquad V = mgz$$

where m and g denote the mass of the body and the gravitation acceleration, respectively, while all other symbols are explained and Figure 1.11.

Since $\mathrm{d}s = v\,\mathrm{d}t$, by definition of speed, we have

$$S_0 = \int m\,v\,\mathrm{d}s = \int m\,v^2\,\mathrm{d}t = \int 2T\,\mathrm{d}t = \int (T+T)\mathrm{d}t$$

Substituting $T = E - V$ for the last term in the last integral we get

$$S_0 = \int (T + E - V)\mathrm{d}t = E\int \mathrm{d}t + \int (T - V)\mathrm{d}t = E\,\Delta t + \int (T - V)\mathrm{d}t$$

Considering paths that have equal total time $\Delta t = t_2 - t_1$, we can drop the term $E\,\Delta t$ and minimize the quantity

$$S = \int_{t_1}^{t_2} L\mathrm{d}t\,, \qquad L \coloneqq T - V$$

The difference of kinetic and potential energy is known as the *Lagrangian* and its integral is again the action, where for clarity we call S Hamilton's action, and S_0 Maupertuis action. The actual physical path is the one that extremizes the action:

$$\delta S = 0$$

for small virtual variations of the path that vanish at the fixed times t_1 and t_2. The parsimonious law in this case is the *principle of extremal action* (or principle of stationary action or Hamilton's principle): *From all possible motions between two points, the true motion has extremal action*.

Applying the calculus of variations the solution yields the Euler–Lagrange equation (for its derivation see e.g. Goldstein et al., 2002), a vector equation which for our two-dimensional system is

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial u_x}\right) - \frac{\partial L}{\partial x} = 0, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial u_z}\right) - \frac{\partial L}{\partial z} = 0$$

Applying this solution to our example we find a single (global) minimum (least action):

$$u_x = u_0\ (= \text{constant}), \qquad u_z = -gt$$

from which we obtain:

$$x = u_0 t, \qquad z = -\frac{gt^2}{2} = \frac{gx^2}{2u_0^2}$$

The latter equation denotes a parabola (going down). The solution gives not only the geometry (parabola) and direction (down) of the trajectory but the full description of the movement of the weight.

We note that when the velocity is constant, the Lagrangian is also constant and hence the action becomes proportional to the travel time. Thus, in the case of light, the principle of extremal action switches to Fermat's principle.

The principle of extremal action unifies philosophical teleology (Leibniz's optimization of the "best" world), geometric intuition (Maupertuis' least action over paths), and temporal generality (Hamilton's integral over time). The latter is more powerful because it does not require energy conservation or fixed spatial endpoints. Instead, in Maupertuis' formalization the conservation of energy is a required condition. Under this condition, the two converge on the same equations of motion.[2]

In its Hamiltonian form, the principle of extremal action not only yields the equations of motion but also, through Noether's theorem (see e.g. Goldstein et al., 2002), accounts for the fundamental conservation laws of classical physics whenever the action possesses corresponding continuous symmetries. Time-translation invariance of the Lagrangian (no explicit time dependence) implies conservation of energy. Spatial-translation invariance (no explicit position dependence) leads to conservation of linear momentum. Rotational invariance (isotropy of space, no preferred direction) produces conservation of angular momentum. These symmetries are not imposed ad hoc; they emerge naturally from the structure of the action. Thus, the variational principle elegantly unifies dynamics and conservation: the same global optimization that selects physical paths also safeguards the great invariants—energy, linear momentum, and angular momentum—revealing a profound harmony in the laws of Nature.

Thus, a single variational principle describes diverse phenomena in physics, including (but not limited to) classical mechanics and optics. A single global optimization principle over paths provides profound economy in classical and modern physics. But this works well only in simple systems.

_____

[1] Pierre-Louis Moreau de Maupertuis (1698–1759) – Biography. MacTutor History of Mathematics, https://mathshistory.st-andrews.ac.uk/Biographies/Maupertuis/.

[2] Principle of least action – Scholarpedia, http://www.scholarpedia.org/article/Principle_of_least_action#Relation_of_Hamilton_and_Maupertuis_Principles

1.4 From simple to complex systems

The illustrations of trajectories of simple/single objects that are studied in Digression 1.A and Digression 1.B are actually not within the scope of this book. We do not need stochastics to study them. Here we are interested in more complex systems. When we investigate a system of many "bodies", e.g. molecules of water in fluid form (liquid or gaseous phase; see Figure 1.12), we are not interested in the properties (position, momentum, angular momentum) of each particular particle.

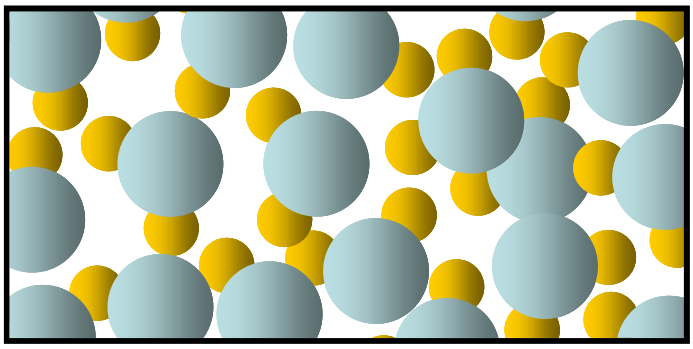

Figure 1.12 Many molecules of water in fluid form.

Even if we were interested, it would be infeasible (and extremely non-parsimonious) to know them. To see this, let us examine the total mass of a gas contained in a room of

volume 60 m^3. We assume standard conditions of temperature and pressure, so that the room contains $60/0.0224$ moles of air or $N = 6.022\times10^{23} \times 60/0.0224 = 1.61\times10^{27}$ molecules, where 6.022×10^{23} is the Avogadro number. Assuming that we wish to just register in a book the properties of each one of the molecules at a single time instance and that we need one line per molecule (9 numbers, i.e., 3 coordinates for each of the properties position, momentum, angular momentum), 60 lines per page and 1000 pages for one book, we can register 60 000 molecules in one book. Hence we will need $1.61 \times 10^{27} / 6 \times 10^4 = 2.69 \times 10^{22}$ books. A 1000-page book weighs about 1.5 kg, and thus the total weight of all books needed would be $2.69 \times 10^{22} \approx 4 \times 10^{22}$ kg. This is of the same order of magnitude of the Moon's mass (7.35×10^{22} kg). Assuming that we have a very fast printer with a capacity of 2 pages per second, we will need 2.69×10^{25} /2 s $\approx 4 \times 10^{17}$ years to print all books, which is about 30 million times the age of the universe ($\approx 1.5\times10^{10}$ years).

Therefore, only *macroscopic* (else known as *thermodynamic*) properties of the system are of interest and are feasible to study. Macroscopic properties are state variables such as pressure, internal energy, entropy, temperature, and characteristic constants such as specific heat and latent heat. Inevitably—albeit often not stated explicitly—macroscopic descriptions rely on probability and involve uncertainty. However, when the system components are very many and identical (as in Figure 1.12), due to the applicability of the probabilistic laws (e.g., law of large numbers), uncertainty at a microscopic level, macroscopically becomes near certainty.

In brief, when we move from simple to complex systems, parsimony demands replacement of microscopic with macroscopic properties and of deterministic with probabilistic descriptions. Now the main question is: *What does Nature extremize in complex systems?* Since we have omitted the system details, these will be replaced by some degrees of freedom. This will imply uncertainty. The quantified metric of both freedom and uncertainty is called *entropy* and will be formally defined in section 2.3. The concept of entropy is accompanied by the *principle of maximum entropy*, whose formulation will also be given in section 2.3.

Hence, in complex systems the quantity that gets extremized is the *entropy*—or it could be the *entropy production* when time is involved (Koutsoyiannis, 2011b). As will be clarified in section 2.3, entropy is a fully stochastic concept. Even its proper definition relies on *stochastics*. The general meaning of the latter term is discussed in Digression 1.C and formally clarified in Chapter 3.

If we consider a die's outcome, the principle of maximum entropy, or maximum uncertainty, yields the simple result that all outcomes are equiprobable, i.e. it yields a global maximum with equal probabilities of the different six outcomes: $P_1 = P_2 = \cdots = P_6 = 1/6$ (see derivation in Digression 2.L). Likewise, if we consider again the above example of the room of volume 60 m^3 and assume a partition into six slices, 10 m^3 each, again the probability that a molecule is found in one of the six slices, say the front sixth, is again $1/6$. The case that all $N = 1.6\times10^{27}$ molecules are in the front sixth and all other slices are empty is not impossible; it just has very low probability, which, assuming independence is $(1/6)^N = 10^{-1\,255\,000\,000\,000\,000\,000\,000\,000\,000}$. Here we note that, even though

we have not formally defined independence yet (this will be done in section 2.2 and 3.5 and Digression 3.B), we intuitively know that the probability of independent events occurring all together equals the product of the probabilities of the separate events—and we have used this simple rule to calculate the probability as $(1/6)^N$. Assuming that each nanosecond (10^{-9} s) the molecules are rearranged and that the age of the universe is 1.5×10^9 years $= 4.7 \times 10^{26}$ ns, we must wait $2 \times 10^{1\,254\,999\,999\,999\,999\,999\,999\,999\,974}$ times the age of the universe to see this unlikely event happen.

Furthermore, according to the principle of maximum entropy, the number of air molecules $\underline{M}$ in each of the six slices will be uniformly distributed, so that each slice has an expected number of molecules $\mathrm{E}[\underline{M}] = (1/6)^N = 2.7 \times 10^{26}$. The standard deviation of the number of molecules is easily found (see Chapter 2) to be $\left(N\,(1/6)(5/6)\right)^{0.5} = 1.5 \times 10^{13}$, which is only 5.6×10^{-14} times the mean $\mathrm{E}[\underline{M}]$. Such a small variation (of the order of 10^{-14}) can be neglected and, thus, the high uncertainty at the microscopic scale becomes almost certainty at a macroscopic level. This enables description of the conditions in the entire room with a few state variables, such as density, pressure, temperature. In other words, parsimony was re-established at the macroscopic level.

Stochastic descriptions of phenomena do not neglect deterministic laws. Rather they fully incorporate them into the stochastic dynamics, primarily as constraints. For instance, in entropy extremization the energy conservation is imposed as an equality constraint. And as discussed above, energy conservation itself can be derived from the principle of extremal action and symmetries of space-time. The principle of maximum entropy does not appear to be a generalization of the principle of extremal action. Rather these two powerful principles seem complementary to each other, enabling effective descriptions of both simple and complex systems:

- Simple systems are studied in deterministic terms, with the basic principle being the extremization of action.
- Complex systems are studied in stochastic terms with the basic principle being extremization of entropy. The principle of extremal action is nonetheless accounted for, primarily through its consequential conservation laws, thus providing the deterministic core of the entropy extremization.

Many believe that there are additional deterministic laws, simple or complex, which do not fit the above categories—for example the laws of thermodynamics or fluid dynamics (Navier-Stokes equations). However, this reflects a misconception of what constitutes a deterministic law. As we shall see in Chapter 6, the laws of thermodynamics are fundamentally stochastic, closely related to the principle of maximum entropy. The Navier-Stokes in their original formulation are deterministic, but are applicable as such only in laminar flow. When turbulence is present (i.e. in almost all real cases except trivial ones) the equations become stochastic. For example, in the Reynolds-averaged Navier–Stokes equations, apparent stresses arise from the fluctuating velocity field (generally known as the Reynolds stresses). These are inherently stochastic terms. Hence, the

atmosphere, for instance—whether in its equilibrium state or in flowing motion—can only be effectively studied in stochastic terms.

Digression 1.C: The meaning of stochastics

Literally, *stochastics* is a term of Greek origin, stemming from the adjective '*stochasticos*' (στοχαστικός), or better its feminine gender, '*stochastice*' (στοχαστική). It is generated from the verb '*stochazesthai*' (στοχάζεσθαι), which in turn comes from the noun '*stochos*' (στόχος), meaning the target.

Aristotle, in his treatise *Nicomachean Ethics* (written ~350 BC) uses the term *stochastice* in its original meaning, related to the target, which, according to him, is the *mean*: "*virtue, therefore, is a balance* ['mesotes']*, in the sense that it is able to hit* [as a target – 'stochos'] *the mean*"[1]. Furthermore, in his treatise *Rhetoric* he uses the term with a metaphorical meaning, which could be translated into English as *guessing* or *guesswork*: "*men have a sufficient natural instinct for what is true, and usually do arrive at the truth. Hence the man who makes a good guess at truth is likely to make a good guess at probabilities* [stochastically]."[2]

However, it was Plato who used the term with a meaning closer to the modern one, i.e., related to uncertainty. In his dialogue *Philebus* (written ~360 BC) he contrasts "*arithmetic and the sciences of measurement*" to *stochastics* and parallels the latter with music, which *"attains harmony by guesswork* [...] *so that the amount of uncertainty mixed up in it is great, and the amount of certainty small."* [3]

The contrast between stochasticity and precision is made clear later by Galenus using the example of a city's clock: "*When a city is being built, let the following problem be set before those who will inhabit it: they want to expertly know, not stochastically but precisely, on an everyday basis, how much time has passed, and how much is left before sunset.*"[4]

The connection of stochastics with prediction or forecast becomes evident in an excerpt from Basilius Caesariensis who contrasts a prophet to a '*stochastes*' (στοχαστής, a noun usually translated incorrectly into English as diviner): "*On the one hand, a prophet is he who foretells the future by revelation of the Spirit; on the other hand, a stochastes is he who infers the future by prudence, comparing similar states, and by the experience of forefathers.*"[5] It seems that this comment has influenced later scholars (e.g. Procopius) and perhaps determined the meaning of *stochastic* in modern Greek, which is imaginative, insightful, thoughtful, cogitative, contemplative, meditative.

The transplantation of *stochastics*, as an international scientific term, to the modern vocabulary is due to Jacob Bernoulli, evidently aware of the Greek language and literature, and in particular of the passage from Plato's Philebus mentioned above. In his famous book *Ars Conjectandi* (written in Latin in 1684-89 but published after his death; Bernoulli, 1713) he writes: "*To conjecture about something is to measure its probability. Therefore we define the art of conjecture, or stochastics, as the art of measuring the probabilities of things as exactly as possible, to the end that, in our judgments and actions, we may always choose or follow that which has been found to be better, more satisfactory, safer, or more carefully considered. On this alone turns all the wisdom of the philosopher and all the practical judgment of the statesman*."[6]

The term was revived by Bortkiewicz (1917; Russian economist and statistician of Polish ancestry) and also by Slutsky (1925, 1928a,b, 1929; Ukrainian/Russian/Soviet mathematical statistician and economist). It appears that the prevalence in USSR of the more sophisticated term *stochastic* (over the term *random*) must have been related to political and ideological reasons (incongruence of randomness with the *dialectical materialism*: models beyond strict determinism were considered with a priori suspicion; see Mazliak 2018).

But it was Kolmogorov (1931) who made the term popular and widespread, as he introduced the term *stochastic process*, also clarifying that *process* means *change of a certain system*. Additionally, he used the term *stationary* to describe a probability density function that is unchanged in time (while at the same time the system state changes). Soon after, Kolmogorov (1933) introduced the modern and consistent definition of probability in an axiomatic manner, based on the *measure theory* (see section 2.1).

[1] «*μεσότης τις ἄρα ἐστὶν ἡ ἀρετή, **στοχαστική** γε οὖσα τοῦ μέσου*» (Aristot. Nic. Eth. 1106b, translation into English adapted from that by H. Rackham. Cambridge, MA, Harvard University Press; London, William Heinemann Ltd. 1934). The notion of 'mesotes' (μεσότης), loosely translated as balance, middle, mean between a respective 'too much' and 'too little', is a key concept in Aristotle's ethical philosophy and thus to hit it as a target is important for him.

[2] «*οἱ ἄνθρωποι πρὸς τὸ ἀληθὲς πεφύκασιν ἱκανῶς καὶ τὰ πλείω τυγχάνουσι τῆς ἀληθείας: διὸ πρὸς τὰ ἔνδοξα **στοχαστικῶς** ἔχειν τοῦ ὁμοίως ἔχοντος καὶ πρὸς τὴν ἀλήθειάν ἐστιν*» (Aristot. Rh. 1.1, translation into English by W. Rhys Roberts, http://classics.mit.edu/Aristotle/rhetoric.1.i.html).

[3] The complete passage is: ΣΩΚΡΑΤΗΣ: «*οἷον πασῶν που τεχνῶν ἄν τις ἀριθμητικὴν χωρίζῃ καὶ μετρητικὴν καὶ στατικήν, ὡς ἔπος εἰπεῖν φαῦλον τὸ καταλειπόμενον ἑκάστης ἂν γίγνοιτο.* […] *τὸ γοῦν μετὰ ταῦτ᾽ εἰκάζειν λείποιτ᾽ ἄν καὶ τὰς αἰσθήσεις καταμελετᾶν ἐμπειρίᾳ καί τινι τριβῇ, ταῖς τῆς **στοχαστικῆς** προσχρωμένους δυνάμεσιν ἃς πολλοὶ τέχνας ἐπονομάζουσι, μελέτῃ καὶ πόνῳ τὴν ῥώμην ἀπειργασμένας.* […] *οὐκοῦν μεστὴ μέν που μουσικὴ πρῶτον, τὸ σύμφωνον ἁρμόττουσα οὐ μέτρῳ ἀλλὰ μελέτης **στοχασμῷ**, καὶ σύμπασα αὐτῆς αὐλητική, τὸ μέτρον ἑκάστης χορδῆς τῷ **στοχάζεσθαι** φερομένης θηρεύουσα, ὥστε πολὺ μεμειγμένον ἔχειν τὸ μὴ σαφές, σμικρὸν δὲ τὸ βέβαιον.*»

(SOCRATES: "*For example, if arithmetic and the sciences of measurement and weighing were taken away from all arts, what was left of any of them would be, so to speak, pretty worthless.* […] *All that would be left for us would be to conjecture and to drill the perceptions by practice and experience, with the additional use of the powers of guessing, which are commonly called arts and acquire their efficacy by practice and toil.* […] *Take music first; it is full of this; it attains harmony by guesswork based on practice, not by measurement; and flute music throughout tries to find the pitch of each note as it is produced by guess, so that the amount of uncertainty mixed up in it is great, and the amount of certainty small*" (Plat. Phileb. 55e, translation by Harold N. Fowler; Cambridge, MA, Harvard University Press; 1925.)

[4] «*πόλεως κτιζομένης προκείσθω τοῖς οἰκήσουσιν αὐτὴν ἐπίστασθαι βούλεσθαι, μὴ **στοχαστικῶς** ἀλλ᾽ ἀκριβῶς, ἑφ᾽ ἑκάστης ἡμέρας, ὁπόσον τε παρελήλυθεν ἤδη τοῦ χρόνου τοῦ κατ᾽ αὐτήν, ὁπόσον θ᾽ ὑπόλοιπόν ἐστιν ἄχρι δύσεως ἡλίου.*» (Γαληνοῦ Περὶ Διαγνώσεως καὶ Θεραπείας τῶν ἐν τῇ ἑκάστου Ψυχῇ Ἁμαρτημάτων — De Dignotione et Curatione cujusque Animi Peccatorum, 80, http://www.poesialatina.it/_ns/greek/testi/Galenus/De_animi_cuiuslibet_peccatorum_dignotione_et_curatione.html).

[5] «*Οὐκοῦν Προφήτης μέν ἐστιν, ὁ κατὰ ἀποκάλυψιν τοῦ Πνεύματος προαγορεύων τὸ μέλλον· **στοχαστὴς** δὲ, ὁ διὰ σύνεσιν ἐκ τῆς τοῦ ὁμοίου παραθέσεως, διὰ τὴν πεῖραν τῶν προλαβόντων, τὸ μέλλον συντεκμαιρόμενος.*» (Basilius, Ερμηνεία εις τον προφήτην Ησαΐαν —Enarratio in prophetam Isaiam, 3.102.26).

[6] "*Conjicere rem aliquam est metiri illius probabilitatem: ideoque Ars Conjectandi sive **Stochastice** nobis definitur ars metiendi quàm fieri potest exactissimè probabilitates rerum, eo fine, ut in judiciis & actionibus nostris semper eligere vel sequi possimus id, quod melius, satius, tutius aut consultius fuerit deprehensum; in quo solo omnis Philosophi sapientia & Politici prudential versatur*" (Bernoulli, 1713).

1.5 Stochastics as a merger and generalization of probability and statistics

A consistent theory for complex systems should necessarily be based on probability—but in an enhanced setting. The mathematical tool to reconcile the complexity of natural systems with parsimony is stochastics. The meaning of the latter term is the following:

$$\textit{Stochastics} = \textit{Probability theory} + \textit{Stochastic processes} + \textit{Statistics}$$

These three pillars of stochastics are detailed in Chapters 2, 3 and 4, respectively.

The probability theory provides the theoretical basis for:

- moving from a microscopic to a macroscopic view of phenomena by mapping sets of diverse elements and events of complex systems to single numbers (a probability or an expected value);
- making deduction when there is uncertainty;
- making induction.

Stochastic processes and Monte Carlo simulations provide the means for:

- probabilistic prediction;
- uncertainty estimation;
- design and management of complex systems.

Statistics provides the empirical basis for:

- summarizing data;
- making inference from data;
- supporting decision making.

It is worth highlighting that, while stochastics provides the way to make reasonable induction from data (observations), it also enables powerful deduction (without data) in particular cases dominated by uncertainty, as will be shown in Chapter 5 and Chapter 6.

Classical probability and statistics are based on the prototype of independence and repeatability (the die-throw prototype). This prototype works well for systems with many *identical* particles for which *independence* can be assumed (this is the case e.g. for ideal gases). However, more complex natural (real-world) systems evolving in time may behave differently from the classical prototype (e.g. turbulent flows, hydrometeorological and climatic processes). In such cases, stochastic models admitting dependence in time/space/state are necessary. Typical stochastic models (particularly the multivariate ones) are often not parsimonious themselves. Therefore, a more advanced stochastic approach is necessary to make models more consistent with (a) the observed natural behaviours, and (b) the principle of parsimony. Such approach will be discussed in Chapter 3.

Digression 1.D Practical difference between dependence and independence

Using observational data of river discharge, we have concluded that the probability of the event that the mean daily discharge at a certain location of a river exceeds 500 m^3/s is small, equal to 10^{-3}. Practically, this means that this event happens on average once every 1000 days or once every 2.74 years. What is the probability that this event occurs for five consecutive days?

Even though we have not yet defined what independence formally is (this will be done in section 2.2 and 3.5, and in Digression 3.B), we intuitively know that the probability of independent events occurring all together equals the product of the probabilities of the separate events. Thus, under independence, the probability sought is simply $(10^{-3})^5 = 10^{-15}$. This is an extremely low probability: it means that we have to wait *on the average* 10^{15} days or 2.74 trillion years, or about 200 times the age of the universe, to see this event happen. However, such events (successive occurrences of extreme events for multiday periods) have been observed in several historical samples. This indicates that the independence assumption is not a justified assumption and yields erroneous results. Thus, we should avoid such an assumption if our target is to estimate probabilities for periods longer than the reference period. Methodologies admitting dependence, i.e., based on the theory of stochastic processes, are more appropriate for such problems and will result in probabilities much greater than 10^{-15}; these will be described in subsequent chapters.

Now let us assume that for four successive days our extreme event has already occurred, i.e., that the mean daily river discharge was higher than 500 m^3/s in all four days. What is the probability that this event will also occur in the fifth day?

Many people, based on an unrefined intuition, may answer that the occurrence of the event already for four days will decrease the probability of another consecutive occurrence, and would

be inclined to give an answer in between 10^{-3} and 10^{-15}. This is totally incorrect. If we assumed independence, then the correct answer would be exactly 10^{-3}; the past does not influence the future. If we assume positive dependence, which is a more correct assumption for natural processes, then the probability sought becomes higher, not lower, than 10^{-3}; it becomes more likely that a flood day will be followed by another flood day.

Similar things happen if we move from the daily scale of the above example to the annual scale, or to even larger scales. For example, if several warm winters have occurred in a series, then the probability that the next winter would also be warm is increased—not decreased. Ignorance of this simple truth may have severe consequences for those who aspire to predict the future and those who believe their predictions. A didactic historical example is the failed prediction of Hitler's meteorologist Franz Baur about the 1941-42 winter in Russia, which marked the Battle of Moscow. Quoting a fascinating paper by Neumann and Flohn (1987)[1]:

> *Baur was requested by the headquarters of the German Air Force to distribute his long-range forecasts to about 25 military offices. A forecast for winter 1941-42 was issued by him, probably at the end of October 1941, based on regional climatology and (supposed) sun-spot-climate relationships. The prediction called for a normal or a mild winter. Baur's main justification for this rested with the assertion that never in climatic history did more than two severe winters occur in a row. Since both of the preceding two winters, 1939-40 and 1940-41, were severe in Europe, he did not expect that the forthcoming winter would also be severe.*

However, that winter, in which the first major Soviet counteroffensive of the war was launched, turned out to be one of the coldest in record:

> *The cold outbreak of early December, coming after a cool to cold October and November* [...] *gravely hit the German armies that were not appropriately clothed (Hitler expected to break the resistance of the USSR before the coming of winter) and which were not equipped with armaments, tanks, and motorized vehicles that could properly function even in a "normal" winter in the northern parts of the USSR, let alone in a winter as rigorous as that of 1941-42.*
> *On or about 8 December, K. Diesing, chief of the CWG and scientific adviser to the chief of the Weather Service of the Air Force (General Spang), asked Flohn to listen in on a second earphone to a telephone call to Baur. In the call, Diesing cited to Baur the reports of very low temperatures in the East and asked him if he maintains his seasonal forecast in face of the reports. Baur's response was "the observations must be wrong".*

Those who interact with deterministic modellers of today may recognize in the last phrase in quotation marks a pretty modern attitude.

[1] A more detailed account about Baur's infamous prognosis, which is now public, can be found in Wiuff (2023).

1.6 Change, randomness and predictability

It is trivial to say that physics and geosciences deal with change. Actually, change has been studied very early, at the birth time of science and philosophy. Most of the Greek philosophers had something to say about it. Heraclitus summarized the dominance of change in a few famous aphorisms, in which he used simple notions to describe change: the flow and the river. Most famous is an aphorism no longer than two words, magically put together:

> *Everything flows* (better known in its original Greek version, "*Πάντα ῥεῖ*")*.

* Quoted in Plato's *Cratylus*, 439d, 440c.

Alternative expressions of the same idea are listed in the motto at the beginning of the book.

Also, Aristotle (particularly in his *Meteorologica*) not only understood the scale and extent of change:

> *all changes in course of time*.*

but he also tried to find invariant properties within change, something that is quite important in modern science. Namely, he understood and aptly expressed the conservation of mass within the hydrological cycle (cf. Koutsoyiannis and Mamassis, 2021):

> *Thus,* [the sea] *will never dry up; for* [the water] *that has gone up beforehand will return to it.*†

> *Even if the same amount does not come back every year or in a given place, yet in a certain period all quantity that has been abstracted is returned.*‡

The hierarchical chart of Figure 1.13 tries to classify change with respect to its predictability. In simple systems (left part of the graph) the change is regular. The regular change can be aperiodic such as the motion of a weight examined in Digression 1.B, or periodic, such as in the motion of Earth with daily and annual period. Whatever it is, using equations of dynamical systems, regular change is predictable.

But this type of change is rather trivial. More interesting are the more complex systems at long time horizons (right part of the graph), where change is unpredictable in deterministic terms, or random. Pure randomness, like in classical statistics (e.g., dice throws), where different variables are identically distributed and independent, is sometimes a useful model, but in other cases it is inadequate. A structured randomness, characterized by dependence in time and emergence of patterns, should be assumed instead. As will be seen in Chapter 3, the structured randomness, typically characterized by Hurst-Kolmogorov dynamics and maximization of entropy production, is enhanced randomness, expressing enhanced unpredictability of enhanced multi-scale change.

It should be clarified that the concept of randomness is used here in a sense quite different from the most common one. According to the prevailing dichotomous view, natural processes consist of two distinct, usually additive, components: a deterministic part (signal) and a random part (noise). Over large time scales, the randomness averages out and does not produce net change. Thus, under this view, only an exceptional external forcing can produce sustained long-term change.

In contrast, the perspective adopted here (better explained in Koutsoyiannis, 2010, 2013, and Dimitriadis et al., 2016) regards randomness as none other than unpredictability. In macroscopic systems—even those governed by fully deterministic

---

* *μεταβάλλει τῷ χρόνῳ πάντα*; *Meteorologica*, I.14, 353a 16.

† *ὥστε οὐδέποτε ξηρανεῖται· πάλιν γὰρ ἐκεῖνο φθήσεται καταβὰν είς τὴν αὐτὴν τὸ προανελθόν*; *ibid.*, 356b 26.

‡ *κἂν μὴ κατ' ἐνιαυτὸν ἀποδιδῷ καὶ καθ' ἑκάστην ὁμοίως χώραν, ἀλλ' ἔν γέ τισιν τεταγμένοις χρόνοις ἀποδίδωσι πᾶν τὸ ληφθέν*; *ibid.*, II.2, 355a 26.

dynamics without any random excitation—unpredictability (i.e., randomness) emerges intrinsically. Depending on the time horizon and time scale of the prediction, there can be either predictability or unpredictability, i.e. randomness, which is not mere noise but the voice of Nature itself (cf. rain falling or water flowing). With the exception of very simple systems—mostly idealized or imaginary—all natural systems, from dice to planets, behave in this way. The specific time length at which the transition from predictability to unpredictability occurs can vary widely: from about a tenth of a second for dice, to tens of millions of years for the solar system. At long time horizons (longer than this characteristic time length) all is random—and far from static.

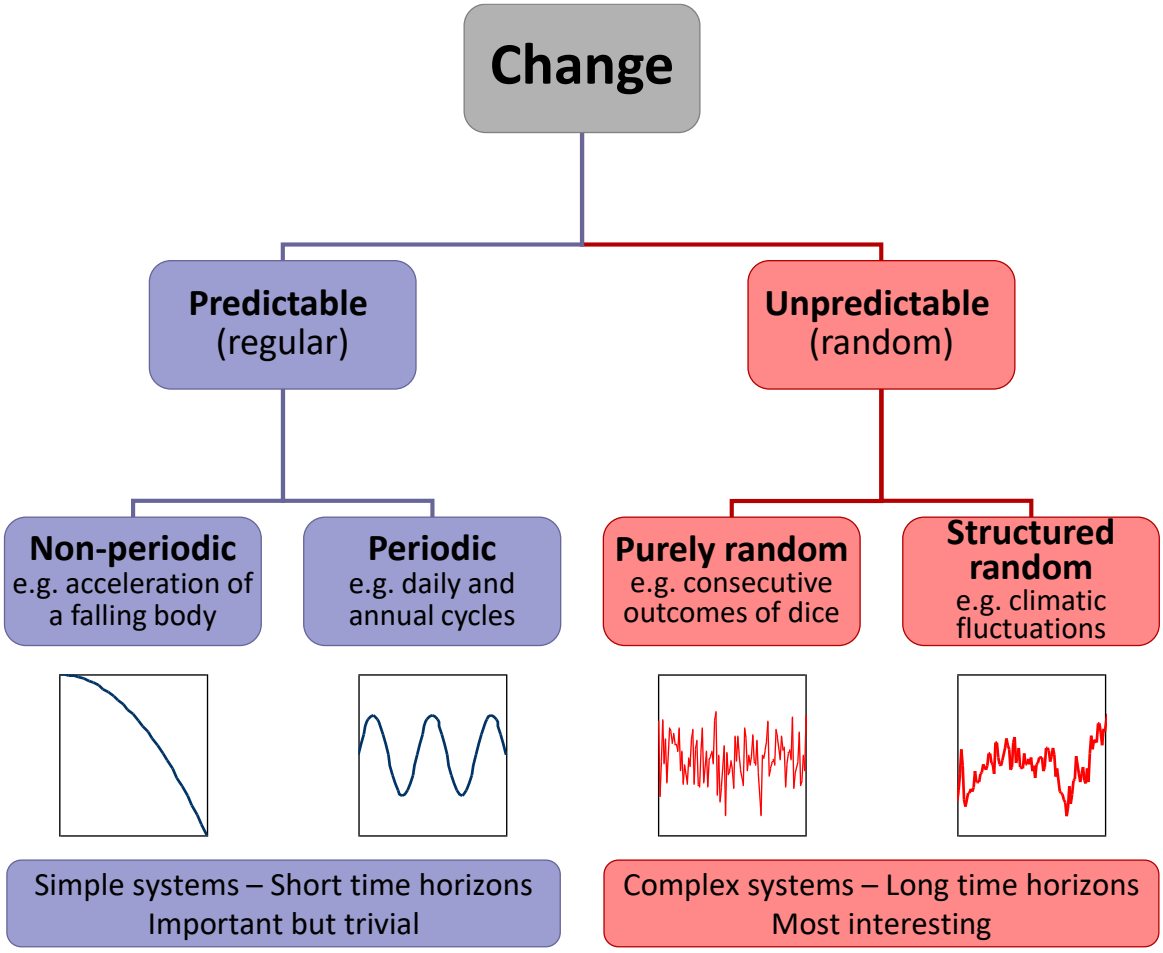

Figure 1.13 Classification of change. (Source: Koutsoyiannis 2013b.)

Chapter 2. Basic concepts of probability

2.1 Definition of probability

For the proper understanding and use of probability, it is very important to insist on the definitions and clarification of its fundamental concepts. Such concepts may differ from other, more familiar, arithmetic and mathematical concepts, and this may cause confusion or even collapse of our cognitive construction, if we do not base it on solid foundations. For instance, in our everyday use of mathematics, we expect all quantities to be expressed by numbers and the relationship between two quantities to be expressed by a function, which to a numerical input quantity associates (maps) another numerical quantity, a unique output. Probability is also a mapping, but instead of a number, the input quantity is an event, which can be represented mathematically by a set. Probability is then a quantified likelihood that the specific event will occur. This type of representation was proposed by Kolmogorov (1933). There are other probability systems different from Kolmogorov's axiomatic system, according to which the input is not a set. Thus, in Jaynes (2003)* the input of the mapping is a logical proposition and probability is a quantification of the plausibility of the proposition. The two systems are conceptually different, but the differences lie mainly on interpretation rather than on the mathematical results. Here we follow Kolmogorov's system, which we complement with an important addition, a new axiomatic definition of entropy (section 2.3).

Kolmogorov was an outstanding member of the Moscow School of Mathematics, which gave importance to definitions and to clarity, following the Aristotelian tradition of *sapheneia* (Digression 2.A). His approach to probability theory is based on the notion of *measure*, which maps *sets* onto *numbers*. The theory of measure is a very important development in mathematics and was founded by Émile Borel (first σ-additive measure) and Henri Lebesgue (Lebesgue measure and integral) in the turn of the 19th to 20th century, with Giuseppe Vitali (non-measurable sets) following soon after. In the next two decades, the subject was advanced by the Moscow School of Mathematics, Dmitry Egorov (Egorov's theorem), Nikolai Luzin (measurable functions, analytic sets, descriptive set theory) and Mikhail Suslin (analytic sets, Suslin's problem), while at the same period it was explored by Constantin Carathéodory (outer-measure approach). Kolmogorov understood that probability is simply a normalized measure.

The objects of probability theory, the *events*, to which probability is assigned, are thought of as sets. For instance, the outcome of a roulette spin, i.e. the pocket of the wheel into which the ball falls is one of 37. (In a European roulette wheel, pockets are numbered 0 to 36 and coloured black or red except 0 which is green). Thus, all sets {0}, {1}, … {36} are events (also called elementary events). But they are not the only ones. All possible subsets of Ω, including the empty set $\emptyset$, are events. The set $\Omega \coloneqq \{0, 1, \ldots, 36\}$ is an event too. Since any possible outcome is contained in Ω, the event Ω occurs in any case and is called the *certain event*. The sets ODD ≔ {1, 3, 5, …, 35}, EVEN ≔ {2, 4, 6, …, 36}, RED ≔ {1, 3, 5, 7, 9, 12, 14, 16, 18, 19, 21, 23, 25, 27, 30, 32, 34, 36}, and BLACK ≔ Ω – RED – {0}

* Jaynes's book cited here was published posthumously (he died in 1998).

are also events (and also bettable). While events are represented as sets, in probability theory there are certain differences from set theory in terminology and interpretation, which are shown in Table 2.1.

Table 2.1 Terminology correspondence in set theory and probability theory (adapted from Kolmogorov, 1933).

Set theory	Events
$A = \emptyset$	Event A is impossible
$A = \Omega$	Event A is certain
$AB = \emptyset$ (or $A \cap B = \emptyset$; disjoint sets)	Events A and B are incompatible (mutually exclusive)
$AB \cdots N = \emptyset$	Events A, B, …, N are incompatible
$X \coloneqq AB \cdots N$	Event X is defined as the simultaneous occurrence of A, B, …, N
$X \coloneqq A + B + \cdots + N$ (or $X \coloneqq A \cup B \cup \cdots \cup N$)	Event X is defined as the occurrence of at least one of the events A, B, …, N
$X \coloneqq A - B$	Event X is defined as the occurrence of A and, at the same time, the non-occurrence of B
$\overline{A} \coloneqq \Omega - A$ (the complement of A)	The opposite event $\overline{A}$ consisting of the non-occurrence of A
$B \subseteq A$ (B is a subset of A)	From the occurrence of event B follows the inevitable occurrence of event A

According to Kolmogorov's (1933) axiomatization, probability theory is based on three fundamental concepts and four axioms. The concepts form the triplet (Ω, Σ, P), called *probability space*, where:

1. Ω is a non-empty set, which Kolmogorov calls the *basic set*, whose elements ω are called *elementary events* (also known as *outcomes* or *states*). Sometimes Ω is also called *sample space*, *ground set* or *certain event*. Here we use the last two terms.
2. Σ is a set known as *σ-algebra* (or *σ-field*) whose elements E are subsets of Ω, known as *events*. Ω and $\emptyset$ are both members of Σ, and, in addition, (a) if E is in Σ then the complement $\Omega - E$ is in Σ; (b) the union of countably many sets in Σ is also in Σ.
3. P is a function called *probability* that maps events (i.e., sets) to real numbers, assigning to each event E (member of Σ) a number between 0 and 1.

The four axioms, which define the properties of P, are the following:

I. *Non-negativity*: For any event A, $P(A) \geq 0$.
II. *Normalization*: $P(\Omega) = 1$.
III. *Additivity*: For any incompatible events A and B (i.e., $AB = \emptyset$), $P(A + B) = P(A) + P(B)$.
IV. *Continuity at zero*: If $A_1 \supseteq A_2 \supseteq \ldots \supseteq A_n \supseteq \ldots$ is a decreasing sequence of events, with $A_1 A_2 \cdots A_n \cdots = \emptyset$, then $\lim_{n\to\infty} P(A_n) = 0$.

We note that in the case that Σ is finite, axiom IV follows from axioms I-III; however, for infinite ground sets it should be put forward as an independent axiom. We note that

Kolmogorov's numbering of axioms is different as he also put the basic notions 1-3 as axioms. Modern texts merge axiom III and IV forming an axiom of additivity of infinite events, but here we prefer the original formulation as it is more intuitive and useful (see an example later in section 2.3)

Digression 2.A: What is sapheneia?

It is stunning that before Kolmogorov, the concept of probability was in wide use for almost three centuries, since its introduction by Jacob Bernoulli, without a proper definition. Earlier definitions were problematic (e.g. affected by circular logic). For this reason, they are not referred to here, but the interested reader can find them in any probability book.

One may notice the modern world's recent disrespect for clarity in science, which also affects definition. This disrespect is "theorized" in the following statement by Mandelbrot (1999, p. 14):

> *Let me argue that this situation* [absence of a definition] *ought not create concern and steal time from useful work. Entire fields of mathematics thrive for centuries with a clear but evolving selfimage, and nothing resembling a definition.*

Perhaps the reason why modern science prefers a pace of fuzziness over that of clarity is its strengthening links to politics and finance. Fuzziness indeed better serves contemporary politics. On the other hand, fuzziness per se has been theorized by the modern fuzzy set theory, which however is one of the several modern reinventions of probability.

Probability and stochastics try to replace fuzziness with rigour in fields where uncertainty dominates. Therefore, it needs a rigorous definition per se, and this has been provided by Kolmogorov. The Moscow School of Mathematics, and in particular its founders Dimitri Egorov and Nikolai Luzin (the latter being Kolmogorov's mentor) had a different approach, opposite to Mandelbrot's. This is vividly expressed by the following Luzin's note, quoted by Graham (2011):

> *Each definition is a piece of secret ripped from Nature by the human spirit. I insist on this: any complicated thing, being illumined by definitions, being laid out in them, being broken up into pieces, will be separated into pieces completely transparent even to a child, excluding foggy and dark parts that our intuition whispers to us while acting; only by separating into logical pieces can we move further, towards new successes due to definition.*

In fact, Luzin's approach was formed much earlier, in the first steps of the development of science. Aristotle promoted *sapheneia* (σαφήνεια[1]), which includes clarity and is also related to the accurate accounting of the phenomena and the attainment of accurate scientific knowledge (Lesher, 2010). Aristotle clearly linked sapheneia with truth:

> *We must always endeavor, from statements that are true but not clearly* [ού σαφῶς] *expressed, to arrive at a result that is both true and clear* [σαφῶς] (Aristotle, Eudemian Ethics 1220a).[2]

The importance Aristotle gave to sapheneia can be seen in the way he equated untrained soldiers to those who do not practice it:

> *These thinkers* [...] *seem to have grasped* [...] *the causes* [...] *only vaguely and indefinitely* [ἀμυδρῶς καὶ ούθὲν σαφῶς]. *They are like untrained soldiers in a battle, who rush about and often strike good blows, but without science; in the same way these thinkers do not seem to understand their own statements, since it is clear that upon the whole they seldom or never apply them* (Aristotle, Metaphysics 985a).[3]

The introduction of terminology, i.e., of sophisticated terms (which either do not exist in the colloquial language or exist with a loose meaning) and their definitions, is another reflection of the sapheneia desideratum. Note that, in Greek, the names *term* and *definition* have common origin (ὅρος and ὀρισμός, respectively), and Aristotle sometimes used the two interchangeably, perhaps reflecting the fact that a term without a definition is not a proper term. He emphasized the need to name scientific concepts:

> *Now most of these* [concepts] *have no names, and we must try* [...] *to invent names ourselves for the sake of clarity* [σαφήνεια] *and ease to follow* (Aristotle, Nicomachean Ethics, 985a).[4]

Furthermore, Aristotle gave credit to Socrates for the introduction of definitions and emphasized that the need for them is linked to the use of abstract theoretical concepts rather than of sensible things:

> *Socrates, disregarding the physical universe and confining his study to moral questions, sought in this sphere for the universal and was the first to concentrate upon definitions* [ὁρισμῶν]. [Plato] *followed him and assumed that the problem of definition is concerned not with any sensible thing but with entities of another kind; for the reason that there can be no general definition* [ὅρος] *of sensible things which are always changing* (Aristotle, Metaphysics 1.987b).[5]

The importance of names, especially in mathematics, has been emphasized by Graham (2011), who asserted that naming plays an essential role because mathematical objects that have not yet been named are difficult to work with. For mathematicians naming is the path to gaining control over the objects they conceive. In their book *Naming Infinity*, Graham and Kantor (2009) gave a detailed account of how the naming of abstract concepts contributed to the development of the Moscow School of Mathematics and the founding of descriptive set theory, which gave birth to the modern definition of probability and the development of stochastics.

---

[1] Greek words related to the noun *σαφήνεια* (sapheneia) are the adjective *σαφής*/*σαφές* (saphes), the adverb *σαφῶς* (saphos) and the verb *σαφηνίζειν* (saphenizein).

[2] *Ἀεὶ διὰ τῶν ἀληθῶς μὲν λεγομένων οὐ σαφῶς δὲ πειρᾶσθαι λαβεῖν καὶ τὸ ἀληθῶς καὶ σαφῶς*. (Ἀριστοτέλης, Ἠθικά Ευδήμια, 1220a).

[3] *Οὗτοι μὲν οὖν* [...] *ἠμμένοι φαίνονται,* [...] *ἀμυδρῶς μέντοι καὶ οὐθὲν σαφῶς ἀλλ᾽ οἷον ἐν ταῖς μάχαις οἱ ἀγύμναστοι ποιοῦσιν: καὶ γὰρ ἐκεῖνοι περιφερόμενοι τύπτουσι πολλάκις καλὰς πληγάς, ἀλλ᾽ οὔτε ἐκεῖνοι ἀπὸ ἐπιστήμης οὔτε οὗτοι ἐοίκασιν εἰδέναι ὅ τι λέγουσιν: σχεδὸν γὰρ οὐθὲν χρώμενοι φαίνονται τούτοις ἀλλ᾽ ἢ κατὰ μικρόν* (Ἀριστοτέλης, Μετά τα Φυσικά, 985a).

4 *Εἰσὶ μὲν οὖν καὶ τούτων τὰ πλείω ἀνώνυμα, πειρατέον δ᾽* [...] *αὐτοὺς ὀνοματοποιεῖν σαφηνείας ἕνεκα καὶ τοῦ εὐπαρακολουθήτου* (Ἀριστοτέλης, Ἠθικὰ Νικομάχεια, 1108a).

[5] *Σωκράτους δὲ περὶ μὲν τὰ ἠθικὰ πραγματευομένου περὶ δὲ τῆς ὅλης φύσεως οὐθέν, ἐν μέντοι τούτοις τὸ καθόλου ζητοῦντος καὶ περὶ ὁρισμῶν ἐπιστήσαντος πρώτου τὴν διάνοιαν,* [Πλάτων] *ἐκεῖνον ἀποδεξάμενος διὰ τὸ τοιοῦτον ὑπέλαβεν ὡς περὶ ἑτέρων τοῦτο γιγνόμενον καὶ οὐ τῶν αἰσθητῶν: ἀδύνατον γὰρ εἶναι τὸν κοινὸν ὅρον τῶν αἰσθητῶν τινός, ἀεί γε μεταβαλλόντων* (Αριστοτέλης, Μετά τα Φυσικά, 1.987b).

Digression 2.B: An elementary illustration of probability

For clarification of the basic concepts of probability theory, we give the following example of hydroclimatic interest. Specifically, we study: (a) the occurrence of rainfall at a particular site and a specific time of the year, and (b) the rainfall depth at that site and time.

In (a) we are interested in the mathematical description of the possibilities that a certain day in the specified site and time is wet or dry. These are the outcomes or states of our problem, so the ground set is:

$$\Omega = \{\text{wet}, \text{dry}\}$$

The σ-algebra Σ contains all possible events, i.e.:

$$\Sigma = \{\emptyset, \{\text{wet}\}, \{\text{dry}\}, \Omega\}$$

To fully define probability on Σ it suffices to define the probability of one of the two states, say $P\{\text{wet}\}$. In fact, this is not easy. Usually it is done by induction, and it needs a set of observations to be available and concepts of the *statistics* theory (see Chapter 4) to be applied. For the time being let us arbitrarily assume that $P\{\text{wet}\} = 0.2$. The remaining probabilities are obtained by applying the axioms. Clearly, $P(\Omega) = 1$ and $P(\emptyset) = 0$. Since wet and dry are incompatible, $P\{\text{wet}\} + P\{\text{dry}\} = P(\{\text{wet}\} + \{\text{dry}\}) = P(\Omega) = 1$, so $P\{\text{dry}\} = 0.8$.

In (b), the state is described by the rainfall depth which can be zero or positive. Therefore, the ground set is the set $\mathbb{R}^+ \cup \{0\}$. We will see how we can assign probabilities in this case in Digression 2.J.

2.2 Conditional probability, independent and dependent events

By definition (Kolmogorov, 1933), *conditional probability of the event A given B* (i.e. under the condition that the event B has occurred) is the quotient:

$$\frac{P(AB)}{P(B)} =: P(A|B) \tag{2.1}$$

Obviously, if $P(B) = 0$, this conditional probability cannot be defined (except in limiting cases by applying the l'Hôpital's rule). It follows that:

$$P(AB) = P(A|B)P(B) = P(B|A)P(A) \tag{2.2}$$

From this it follows that:

$$P(B|A) = P(B)\frac{P(A|B)}{P(A)} \tag{2.3}$$

Equation (2.3) is known as the *Bayes theorem*.

If it happens that $P(A|B) = P(A)$, i.e., the probability of A does not depend on whether or not B has occurred, then the events A and B are called (*stochastically*) *independent*. In this case from equation (2.1) it follows that:

$$P(AB) = P(A)P(B) \tag{2.4}$$

Otherwise, A and B are called (*stochastically*) *dependent*.

The definition can be extended to many events. Thus, the events $A_1, A_2, \ldots$ are *independent* (or *mutually independent*) if for any finite set of distinct indices $i_1, i_2, \ldots, i_n$:

$$P\left(A_{i_1}A_{i_2}\ldots A_{i_n}\right) = P\left(A_{i_1}\right)P(A_{i_2})\ldots P(A_{i_n}) \tag{2.5}$$

Thus, handling probabilities of independent events is easy. However, this is a special case because often macroscopic natural events are dependent. In handling dependent events the notion of conditional probability is vital.

It is easy to show that the generalization of (2.5) for dependent events takes the forms:

$$P(A_n \ldots A_1) = P(A_n|A_{n-1}\ldots A_1)\cdots P(A_2|A_1)P(A_1) \tag{2.6}$$

$$P(A_n \ldots A_1|B) = P(A_n|A_{n-1}\ldots A_1B)\cdots P(A_2|A_1B)P(A_1|B) \tag{2.7}$$

which are known as the *chain rules*. It can also be proved (homework) that if A and B are mutually exclusive, then

$$P(A + B|C) = P(A|C) + P(B|C) \tag{2.8}$$

$$P(C|A + B) = \frac{P(C|A)P(A) + P(C|B)P(B)}{P(A) + P(B)} \tag{2.9}$$

and if $A + B = \Omega$, so that $P(A + B) = 1$, then

$$P(C) = P(C|A)P(A) + P(C|B)P(B) \tag{2.10}$$

Digression 2.C: An example of the dependence of probability on information

We assume that at a certain place on Earth (say, in a city in the United Kingdom) and a certain period of the year, a dry and a wet day are equiprobable and that on different days the states (wet or dry) are independent. What is the probability that two consecutive days are wet in the following cases? (a) Unconditionally. (b) If we know that the first day is wet. (c) If we know that the second day is wet. (d) If we know that one of the two days is wet.

We assume that the probability spaces in the two days are identical. We denote Ω_A and Ω_B the ground sets for the first and second day,respectively. The ground set for the two consecutive days is the Cartesian product $\Omega_A \times \Omega_B$. Further, we denote $A \coloneqq$ {first day wet}, $\overline{A} \coloneqq$ {first day dry}, $B \coloneqq$ {second day wet}, $\overline{B} \coloneqq$ {second day dry}. Then, $\Omega_A = \{A, \overline{A}\}$ and $\Omega_B = \{B, \overline{B}\}$. The ground set $\Omega_A \times \Omega_B$ can be written as $\{AB, A\overline{B}, \overline{A}B, \overline{AB}\}$.

(a) We want to find $P(AB)$. Obviously, given the independence assumption, $P(AB) = P(A)P(B) = (1/2)^2 = 1/4$. Because of equiprobability and independence, each of the four events has probability $1/4$.

(b) Now the probability sought is $P(AB|A)$. Using the chain rule in equation (2.7) we find $P(AB|A) = P(A|AB)P(B|A) = 1 \times 1/2 = 1/2$.

(c) Like in (b), we find $P(AB|B) = 1/2$.

(d) The condition that one of the two days is wet corresponds to the composite even $AB + A\overline{B} + \overline{A}B$. Thus, the probability sought is

$$P\left(AB\middle|AB + A\overline{B} + \overline{A}B\right) = \frac{P\left(AB\left(AB + A\overline{B} + \overline{A}B\right)\right)}{P\left(AB + A\overline{B} + \overline{A}B\right)} = \frac{P(AB)}{P\left(AB + A\overline{B} + \overline{A}B\right)} = \frac{1/4}{3/4} = \frac{1}{3}$$

where we have used the definition of conditional probability and the fact that $AB, A\overline{B}, \overline{A}B$ are mutually exclusive.

To connect the example to the real world, let us assume that a friend travelled to this city for a specified couple of days. If we do not have any information except the specific dates, then to the event that she used her umbrella on both days we will assign probability $1/4$. If we knew that she went there without an umbrella and she was forced to buy one, then to the same event we might assign a probability of $1/3$. If, during the first day of her trip, we saw (e.g. in her social media posts) a photo showing her in the city holding an umbrella, then we would change the probability to $1/2$. In other words, the information we have for a problem may introduce dependencies in events that are initially assumed independent. More generally, the probability is not an invariant quantity, characteristic of physical reality in absolute terms, but a quantity that depends on our knowledge or information about the examined phenomenon. It may seem paradoxical that the probability depends on information, but it is not. The rules according to which we are assigning probabilities are objective and theoretically consistent. Yet it may not always be direct to assign probabilities and also the assigned values may depend on the way the information was obtained (see relevant discussion for the particular problem examined here in Bar-Hillel and Falk,1982). We may additionally recall that even in classical deterministic physics we are dealing with similar situations. For instance, the location and velocity of a moving particle are not absolute objective quantities. If we change the coordinate system, the numerical values of the coordinates and the velocity will also change (see also Digression 6.A).

Digression 2.D: An example of dependent events

The independence assumption in the problem in Digression 2.C is often a poor representation of physical reality. To construct a slightly more realistic model, let us assume that the probability of today being wet (B) or dry ($\overline{B}$) depends on the previous day's state (A or $\overline{A}$). It is reasonable to assume that the following inequalities hold:

$$P(B|A) > P(B) = 0.5, \qquad P\left(\overline{B}\middle|\overline{A}\right) > P\left(\overline{B}\right) = 0.5$$

Now, the problem becomes more complicated than before. Let us arbitrarily assume that $P(B|A) = 0.6$. Then the probability that both days are wet is $P(AB) = P(B|A)P(A) = 0.6 \times 0.5 = 0.3 > 1/4$. For the sake of completeness, we also calculate the probabilities of the other combinations. From (2.10), we get $P(B) = P(B|A)P(A) + P\left(B\middle|\overline{A}\right)P\left(\overline{A}\right)$, from which we find:

$$\begin{bmatrix} P(B|A) & P\left(B\middle|\overline{A}\right) \\ P\left(\overline{B}\middle|A\right) & P\left(\overline{B}\middle|\overline{A}\right) \end{bmatrix} \begin{bmatrix} P(A) \\ P\left(\overline{A}\right) \end{bmatrix} = \begin{bmatrix} P(A) \\ P\left(\overline{A}\right) \end{bmatrix}, \quad \begin{bmatrix} P(A|B) & P\left(B\middle|\overline{B}\right) \\ P\left(\overline{B}\middle|A\right) & P\left(\overline{B}\middle|\overline{A}\right) \end{bmatrix} \begin{bmatrix} P(B) \\ P\left(\overline{B}\right) \end{bmatrix} = \begin{bmatrix} P(B) \\ P\left(\overline{B}\right) \end{bmatrix}, \quad \begin{bmatrix} P(B) \\ P\left(\overline{B}\right) \end{bmatrix} = \begin{bmatrix} P(A) \\ P\left(\overline{A}\right) \end{bmatrix}$$

where for convenience we have used matrix/vector representation. Thus,

$$P(B|\bar{A}) = \frac{P(B) - P(B|A)P(A)}{P(A)} = \frac{0.5 - 0.6 \times 0.5}{0.5} = 0.4$$

Hence, $P\left(\overline{A}B\right) = P\left(B\middle|\overline{A}\right)P\left(\overline{A}\right) = 0.4 \times 0.5 = 0.2 < 1/4$. Because of symmetry $P\left(\overline{AB}\right) = 0.3$ and $P\left(A\overline{B}\right) = 0.2$. Thus, the dependence resulted in higher probabilities that the consecutive events are similar and smaller probabilities that they are dissimilar. This corresponds to a typical natural behaviour (see also Chapter 3).

2.3 Completion of the probability system: Entropy and the principle of maximum entropy

While Kolmogorov's system to define probability is a major step to rigorous science, it does not help to determine probability in real-world problems, as it is too general. Typically, assignment of probabilities is done by induction, usage of data and statistical tools (see Chapter 4). Yet deduction is possible in stochastics and the best tool for it is the *principle of maximum entropy*. This principle can assign probabilities to events, even if we know nothing about them, and modify the initial assignments in cases where information becomes available.

The definition of *entropy* requires reference to *partitions* of the ground set $\Omega = \{\omega_1, \omega_2, \dots \omega_m\}$, where m may be infinite. A partition is a set of nonempty mutually exclusive sets (events) whose union equals Ω. Symbolically, $\mathbb{A} = \{A_1, A_2, \dots, A_n\}$ (or $\mathbb{A} = [A_i]$), with $A_1 + A_2 + \cdots + A_n = \Omega$, $A_i A_j = \emptyset, i, j = 1 \dots, n$. The sets (events) $A_1, A_2, \dots, A_n$ are the elements of the partition. Here are some basic properties of partitions:

- The coarsest partition has one element, $\mathbb{A} = \{\Omega\}$
- The finest partition is composed of elementary events, $\mathbb{A} = \left\{\{\omega_1\}, \{\omega_2\}, \dots, \{\omega_m\}\right\} =: \mathbb{V}$.
- A partition with two elements, $\mathbb{A} = \{A, \overline{A}\}$ is called a bipartition.
- A refinement of a partition $\mathbb{A}$ is a partition $\mathbb{B}$ such that each element B_j of $\mathbb{B}$ is a subset of some element A_i of $\mathbb{A}$.
- A common refinement of two partitions is a refinement of both.
- Given two ground sets Ω, Ω', and partitions thereof $\mathbb{A} = [A_i], \mathbb{A}' = [A'_j]$, their product partition, denoted as $\mathbb{A} \otimes \mathbb{A}'$, is a partition of the cartesian product $\Omega \times \Omega'$, consisting of all events $A_i \times A'_j, A_i \in \mathbb{A}, A'_j \in \mathbb{A}'$ for all i, j. Since the occurrence of

the event $A_i \times A'_j$ is equivalent to the occurrence of both A_i and A'_j, we can simplify the notation $A_i \times A'_j$ by omitting '×' and writing $A_i A'_j$.

- For a finite Ω with n elements, the number of elements (cardinality) of Σ, the set of subsets of Ω, is 2^n and that of Π, the set of all partitions, is much greater (for $n > 4$), known as Bell number, B_n, and given by the recursive formula

$$B_0 = 1, \qquad B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k \tag{2.11}$$

- For Ω equivalent to the set of natural numbers (i.e. for a discrete Ω, or a continuous one with interval partitions) the cardinality of either Σ or Π is infinite, $2^{\aleph_0}$, where $\aleph_0$ (aleph-zero) is the cardinality of the set of the natural numbers. If all partitions of an Ω composed of real numbers are considered, then the cardinality is $2^{\left(2^{\aleph_0}\right)}$.

While probability is a function $P(A)$ that maps an event (i.e. subset of the ground set Ω) to a real number in the interval [0,1], entropy, which we will denote by the Greek letter Φ,* maps a partition $\mathbb{A}$ of the ground set to a real number. If the ground set has finite or countably many elements, that number is nonnegative and bounded from above. Hence, entropy has a maximum value, whose determination can be used to assign probabilities. This is the idea behind the principle of maximum entropy. For a non-countable ground set, entropy can diverge to infinity, but again there are two formal ways to make it finite. The first is to adhere to a countable number of partition elements, so that entropy remains nonnegative with a finite maximum value. The second is to use the notion of *relative entropy*, which is described below. The latter concept can take on negative values but has a finite upper bound, a maximum (relative) entropy.

By incorporating entropy into probability theory, we extend Kolmogorov's triplet (Ω, Σ, P) to the pentad $(\Omega, \Sigma, \Pi, P, \Phi)$, where the meanings of the five symbols are:

Ω: the ground set;
Σ: the σ-algebra that contains subsets of Ω;
Π: the set of all partitions of Ω, where each element of any partition $\mathbb{A}$ (where $\mathbb{A}$ is an element of Π) should be an element of Σ;
P: the probability function, $P: \Sigma \to [0,1]$;
Φ: the entropy function, $\Phi: \Pi \to \mathbb{R}^+$ (or $\mathbb{R}$ for relative entropy).

This pentad enables deductive probability assignments via maximum entropy, as detailed in section 2.11. Probability measures the degree of certainty about a single event A, i.e. a subset of Ω, so that a value of $P(A)$ close to zero or one suggests a nearly impossible or a nearly certain event, respectively. On the other hand, entropy refers to a collection of

* In classical thermodynamics, entropy is denoted by S (the original symbol used by Clausius—see Digression 2.E), while probability texts use the symbol H. Here Φ was preferred as a unifying symbol for information and thermodynamic entropy, under the interpretation that the two are essentially the same thing. One of the reasons for this preference is historical: for long time, entropy used to be denoted by Φ (Perry, 1903; Swinburne, 1904; Ewing, 1920), and this is still echoed in the term tephigram (T-Φ-gram) used in meteorology.

events covering the entire ground set Ω, instead of a single event. If $\Omega = \{A, \overline{A}\}$ with $P(A)$ close to zero or one (equivalently, $P(\overline{A})$ close to one or zero), then the entropy Φ is close to zero in both cases.

Below we will introduce and define entropy in a manner different from that common in literature, even though the final formula does not differ from the commonly used. The reasons for not following the literature are explained in Digression 2.E, along with the historical context. Specifically, here we incorporate the principle of maximum entropy into the definition. Put it differently, we first posit that principle and then we set the required postulates for an entropy definition that materializes the principle. We interpret entropy as quantification of uncertainty and express the principle of maximum entropy as follows:

> **Principle of maximum entropy:** Uncertainty will not be lower than its maximum possible value without a reason.

Such a reason, if exists, would be expressed mathematically as a constraint, but the definition of entropy should be independent of such constraints. Our definition will be based on partitions $\mathbb{A}$ of the ground set Ω. It will also use the notion of independence and the intuitive assumption that independence between two events A, A' results in maximum uncertainty as the occurrence of A does not contain any information about the occurrence of A'. Two elements A_i, A_j of a partition $\mathbb{A}$ are mutually exclusive, so $P(A_iA_j) = 0$. Therefore, in order for independence to be meaningful, we need more than one partition. As the most parsimonious case, here we consider two independent partitions. With these ideas, we now form the following postulates which can define entropy as a function of the probabilities $P(A_i)$ of the events A_i.

> **Postulate 1, continuity**: Given a partition $\mathbb{A} = \{A_1, \dots, A_n\}$ of the ground set Ω, its entropy $\Phi(\mathbb{A})$ is a twice continuously differentiable function $\Phi_P(\,)$ of the probabilities of the events A_i that form the partition. Namely, $\Phi(\mathbb{A}) \coloneqq \Phi_P\big(P(A_1), \dots, P(A_n)\big)$, i.e., $\Phi_P: \mathbb{R}^n \to \mathbb{R}^+, \Phi_P \in C^2$.
>
> **Postulate 2, zero at certainty**: The entropy of the coarsest partition is zero, $\Phi(\{\Omega\}) = 0$.
>
> **Postulate 3, non-interaction**: There is no interaction between any two elements of the partition: $\partial^2\Phi(\mathbb{A})/\partial P(A_i)\partial P\big(A_j\big) \equiv 0, i \neq j$.
>
> **Postulate 4, maximization at independence**: If Ω, Ω' are two identical ground sets, and $\mathbb{A}, \mathbb{A}'$ are partitions thereof, then the entropy of the product partition $\mathrm{A} \otimes \mathbb{A}'$ is maximized when each pair of events $A_i \in \mathbb{A}, A'_j \in \mathbb{A}'$ are independent, i.e. $P\big(A_iA'_j\big) = P(A_i)P\big(A'_j\big)$.

In addition to these four postulates, we state the following useful property, which is a direct consequence of the fact that a partition is a set and so the order of its members is indifferent.

> **Property 5, symmetry**: $\Phi(\mathbb{A})$ is permutation-invariant: if $\pi(i)$ is any permutation of $i = 1,2, \dots, n$, then $\Phi(\mathbb{A}) = \Phi_P\big(P(A_1), \dots, P(A_n)\big) = \Phi_P\left(P\big(A_{\pi(1)}\big), \dots, P\big(A_{\pi(n)}\big)\right)$.

Based on the four postulates we specify the form of Φ_P, completing the entropy definition, as

$$\Phi(\mathbb{A}) \coloneqq -\sum_{i=1}^{n} P(A_i) \ln P(A_i) \tag{2.12}$$

We readily observe that, as $P(A_i)$ is a dimensionless quantity, so will also be entropy $\Phi(\mathbb{A})$. As the material presented in this section is new, not contained in literature, we chose to present it formally, by means of theorems and corollaries. The result in equation (2.12) is obtained by means of two theorems (1 and 2) and a corollary (2.1) listed below. Additional theorems and corollaries, also listed below, give important properties of entropy. The proofs of the theorems are given in Appendix 2-I, while those of the corollaries are direct.

Theorem 1: There exists a univariate function $\varphi(P)$ such that

$$\Phi(\mathbb{A}) = \sum_{i=1}^{n} \varphi\big(P(A_i)\big) \tag{2.13}$$

Theorem 2: The function $\varphi(P)$ in equation (2.13) is uniquely determined as

$$\varphi(P) = -k\, P \ln P \tag{2.14}$$

where $k > 0$ is a constant.

Corollary 2.1: The entropy is determined by equation (2.12) uniquely up to a constant multiplier.

Corollary 2.2: $\varphi(x) \geq 0$ (with $\varphi(x) = 0$ when $x = 0$ or $x = 1$).

Corollary 2.3: $\Phi(\mathbb{A}) = \Phi_P\big(P(A_1), \dots, P(A_n)\big)$ is a concave function of $P(A_1), \dots, P(A_n)$.

Corollary 2.4: For independent $\mathbb{A}, \mathbb{A}'$, the entropy of the product partition is

$$\Phi(\mathbb{A} \otimes \mathbb{A}') = \Phi(\mathbb{A}) + \Phi(\mathbb{A}') \tag{2.15}$$

Corollary 2.5: For any ground sets Ω, Ω' and partitions thereof $\mathbb{A}, \mathbb{A}'$,

$$\Phi(\mathbb{A} \otimes \mathbb{A}') \leq \Phi(\mathbb{A}) + \Phi(\mathbb{A}') \tag{2.16}$$

with equality holding when $\mathbb{A}, \mathbb{A}'$ are independent.

Note: Corollary 2.4 is often called *additivity* (or *additivity property* or *additivity principle*) and verbally expressed as "the total entropy of two or more independent systems is equal to the sum of their individual entropies". According to corollary 2.5, that sum is the maximum that a product partition can reach.

Theorem 3: If $\mathbb{B}$ is a refinement of $\mathbb{A}$, then

$$\Phi(\mathbb{B}) \geq \Phi(\mathbb{A}) \tag{2.17}$$

Corollary 3.1: From all partitions of Ω, the finest partition, $V = \big\{\{\omega_1\}, \{\omega_2\}, \dots, \{\omega_m\}\big\}$ has the largest entropy.

Theorem 4: Without any constraint on $P(A_i)$, the maximum entropy is obtained for equal probabilities $P(A_i) = 1/n$ and is $\Phi(\mathbb{A}) = \ln n =: \Phi_{\mathrm{E}}(n)$ (with the subscript 'E' standing for equiprobability).

Note: Theorem 4 expresses what is known as the *principle of insufficient reason* (attributed to Bernoulli and Laplace). This principle is clearly a direct result of the principle of maximum entropy, so there is no need to posit it as an independent logical principle.

Corollary 4.1: Unconstrained maximization of the entropy of the finest partition $\mathbb{V}$ results in entropy

$$\Phi(\mathbb{V}) = \Phi_{\mathrm{E}}(n) = \ln m \geq \ln n \tag{2.18}$$

Note: If the ground set has infinite elements, then unconstrained maximization of the entropy will result in zero probabilities and infinite entropy. This is meaningless and hence in such cases constraints are necessary to derive meaningful results.

Corollary 4.2: $\Phi(\mathbb{V}) = \ln m$ is a monotonic increasing function of m.

Corollary 4.3: If $\Omega^N = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_N$, where Ω_i are identical, each containing m equiprobable events and the events are independent among different Ω_i, then

(a) the probability of each event in the finest partition is $1/m^N$,

(b) the entropy is $\Phi(\mathbb{V}^N) = N \ln m$, and

(c) the quantity $\Phi(\mathbb{V}^N)/N$ is constant.

Note: Case (c) of corollary 4.3 is known as *extensivity*; more generally, extensivity is defined via the $\lim_{N\to\infty} \Phi(\mathbb{V}^N)/N$, which must be finite and nonzero in order for a system to be characterized as extensive (Tsallis, 2022).

Extension to degenerate partitions: We consider the case $A = \Omega, \overline{A} = \emptyset$, which does not constitute a genuine bipartition but a degenerate one. Yet, when considering an infinite Ω, the degenerate bipartition could be thought of as a limit of a sequence of genuine bipartitions $\{A_1, \overline{A}_1\}, \ldots, \{A_i, \overline{A}_i\}, \ldots$, with $A_1 \supseteq \cdots \supseteq A_i \supseteq \cdots$ and $A_1 \ldots A_i \ldots = \emptyset$. Combining the probability's property of continuity at zero and the continuity of entropy (Postulate 1), we infer that the entropy of the degenerate partition will also exist and be given by the same formula, $\Phi(\mathbb{A}) = \varphi(1) + \varphi(0) = \varphi(0)$. Since $\varphi(0) = \lim_{x\to 0}(-k\, x \ln x) = 0$, we get $\Phi(\mathbb{A}) = 0$. This can be extended to degenerate partitions of any dimension, so that $\Phi(\mathbb{A}) = 0$ whenever an element is $\{\Omega\}$ (and all other elements are $\{\emptyset\}$). This can also be written as $\Phi_P(1,0,\ldots,0) = 0$.

Relative entropy: A useful concept, particularly for continuous variables, is the *relative entropy*, which is defined in terms of a background measure B, either normalized (so that $\sum_{i=1}^{n} B(A_i) = 1$) or not. If not, the most typical (but not exclusive) case for continuous variables is the Lebesgue measure, which for an interval $[a, b]$ (or (a, b)) equals the length $b - a$. For any background measure B, the relative entropy is defined as:

$$\Phi(\mathbb{A}||\mathrm{B}) \coloneqq -\sum_{i=1}^{n} P(A_i) \ln\left(\frac{P(A_i)}{B(A_i)}\right) \tag{2.19}$$

It is noted that other names (e.g. Kullback–Leibler divergence) have also been in use for the same concept, while Shore and Johnson (1980) used the term *cross-entropy* for the same concept but with changed sign. Basic properties of relative entropy are given by the following theorem and corollaries:

Theorem 5: The maximum possible relative entropy is $\Phi(\mathbb{A}||\mathrm{B}) = \ln \sum_{i=1}^{n} B(A_i)$.
Corollary 5.1: For constant $B(A_i) = B$, the maximum possible relative entropy is $\Phi(\mathbb{A}||\mathrm{B}) = \ln(nB)$.
Corollary 5.2: If $B(\,)$ is a normalized measure, so that $\sum_{i=1}^{n} B(A_i) = 1$, the maximum possible relative entropy is zero, $\Phi(\mathbb{A}||\mathrm{B}) = 0$.

Conditional entropy: The *event-conditional entropy* of a partition $\mathbb{A} \coloneqq \{A_1, \dots, A_n\}$, conditional on the occurrence of an event B is by definition:

$$\Phi(\mathbb{A}|B) \coloneqq -\sum_{i=1}^{n} P(A_i|B) \ln P(A_i|B) \tag{2.20}$$

Given the partitions $\mathbb{A} \coloneqq \{A_1, \dots, A_n\}$ and $\mathbb{B} \coloneqq \{B_1, \dots, B_{n_B}\}$ we define the partition-conditional entropy of $\mathbb{A}$ given $\mathbb{B}$ as the weighted sum:

$$\Phi(\mathbb{A}|\mathbb{B}) \coloneqq \sum_{j=1}^{n_B} P(B_j)\, \Phi(\mathbb{A}|B_j) \tag{2.21}$$

Theorem 6: If partitions $\mathbb{A}$ and $\mathbb{B}$ are independent, then

$$\Phi(\mathbb{A}|\mathbb{B}) = \Phi(\mathbb{A}), \qquad \Phi(\mathbb{B}|\mathbb{A}) = \Phi(\mathbb{B}) \tag{2.22}$$

Theorem 7: For any partitions $\mathbb{A}$ and $\mathbb{B}$

$$\Phi(\mathbb{A} \otimes \mathbb{B}) = \Phi(\mathbb{A}) + \Phi(\mathbb{B}|\mathbb{A}) = \Phi(\mathbb{B}) + \Phi(\mathbb{A}|\mathbb{B}) \tag{2.23}$$

Corollary 7.1 (resulting from inequality (2.16) and Theorem 7):

$$\Phi(\mathbb{A}|\mathbb{B}) \leq \Phi(\mathbb{A}) \tag{2.24}$$

Theorem 8: If $\mathbb{A}$ is a refinement of $\mathbb{B}$ then

$$\Phi(\mathbb{B}|\mathbb{A}) = 0 \tag{2.25}$$

Corollary 8.1: If $\mathbb{A}$ is a refinement of $\mathbb{B}$ then

$$\Phi(\mathbb{A}) = \Phi(\mathbb{B}) + \Phi(\mathbb{A}|\mathbb{B}) \tag{2.26}$$

Corollary 8.2: If the partition $\mathbb{A} \coloneqq \{A_1, \dots, A_n\}$ is a refinement of the bipartition $\mathbb{B} \coloneqq \{B, \overline{B}\}$, the following relationship holds true:

$$\Phi(\mathbb{A}) = \Phi(\mathbb{B}) + P(B)\Phi(\mathbb{A}|B) + P(\overline{B})\Phi(\mathbb{A}|\overline{B}) \tag{2.27}$$

Corollary 8.3: For any partition element A_j of $\mathbb{A}$, the entropies of $\mathbb{A}$ and $\mathbb{A}_j \coloneqq \{A_j, \overline{A_j}\}$ are related by

$$\Phi(\mathbb{A}) = \Phi\left(\mathbb{A}_j\right) + P\left(\overline{A}_j\right)\Phi\left(\mathbb{A}|\overline{A}_j\right) \tag{2.28}$$

Note: Equation (2.28) is the Shannon's grouping rule, posited by him as a postulate to define entropy (see Digression 2.E). In our framework this is just a corollary.

Mutual information: The *mutual information* of two partitions $\mathbb{A}$ and $\mathbb{B}$ is by definition:

$$I(\mathbb{A}, \mathbb{B}) = \Phi(\mathbb{A}) + \Phi(\mathbb{B}) - \Phi(\mathbb{A} \otimes \mathbb{B}) \tag{2.29}$$

and has the properties

$$I(\mathbb{A}, \mathbb{B}) = \Phi(\mathbb{A}) - \Phi(\mathbb{A}|\mathbb{B}) = \Phi(\mathbb{B}) - \Phi(\mathbb{B}|\mathbb{A}), \qquad I(\mathbb{A}, \mathbb{B}) \geq 0 \tag{2.30}$$

Connection of the entropy definition with the principle of maximum entropy: As already mentioned, it is intuitively understood that Postulate 4 reflects the principle of maximum entropy: Independence between two events A, A' results in maximum uncertainty as the occurrence of A does not contain any information about the occurrence of A'. Postulate 3 is another manifestation of the same: If for some reason we get to know the probability of one of the partition elements, say A_i, then maximum uncertainty would mean that the probabilities of any other event A_j, $i \neq j$ would remain unknown and this is expressed by the condition $\partial^2\Phi(\mathbb{A})/\partial P(A_i)\partial P\left(A_j\right) \equiv 0$. Had this second derivative been nonzero, the knowledge of $P(A_i)$ would provide some information on $P\left(A_j\right)$ and therefore would decrease uncertainty—without a reason. As a final note, the fact that entropy is a quantity that gets maximized implies concavity for its functional expression—a fact formally expressed in Corollary 2.3. In turn, this implies that the second derivative for each variable would necessarily be negative, while the joint second derivatives cannot be positive. The joint second derivatives being zero is the extreme and most intuitive case, fully consistent with Postulate 3.

Digression 2.E: Historical overview of entropy definition and the need for a new one

Entropy is etymologized from the ancient Greek word *έντροπία* (from the verb *έντρέπειν*, to turn into, to turn about) but was introduced as a scientific term by Rudolf Clausius only in 1865, although the concept appears also in his earlier works (as described in Clausius, 1872). The rationale for introducing the term is explained in his own words (Clausius, 1867, p. 358, which indicate that he was not aware of the existence of the word *έντροπία* in ancient Greek):

> *We might call S the transformational content of the body* […]. *But as I hold it to be better to borrow terms for important magnitudes from the ancient languages, so that they may be adopted unchanged in all modern languages, I propose to call the magnitude S the entropy of the body, from the Greek word τροπή, transformation. I have intentionally formed the word entropy so as to be as similar as possible to the word energy; for the two magnitudes to be denoted by these words are so nearly allied in their physical meanings, that a certain similarity in designation appears to be desirable.*

In addition to its semantic content, this quotation contains a very important insight: the recognition that entropy is related to transformation and change and the contrast between entropy and energy, where the latter is a quantity that is conserved in all changes. This meaning has been more clearly expressed in Clausius' famous aphorism (Clausius, 1865):

> *Die Energie der Welt ist konstant. Die Entropie der Welt strebt einem Maximum zu.*

(The energy of the world is constant. The entropy of the world strives to a maximum).

In other words, entropy and its ability to increase (as contrasted to energy and other quantities that are conserved) is the driving force of change. This property of entropy has seldom been acknowledged (Hill and Holman, 1986; Atkins, 2003, 2007). Instead, in common perception entropy epitomizes all "bad things", as if it were disconnected from change, or as if change can only have negative consequences, always leading to deterioration (Koutsoyiannis and Sargentis, 2021; see also Digression 2.G).

Mathematically, thermodynamic entropy, S, is defined in the same Clausius' texts through the equation $\delta S = \delta Q/T$, where Q and T denote heat and temperature. The definition, however, applies to a reversible process only. The fact that in an irreversible process $\delta S > \delta Q/T$ makes the definition imperfect and affected by circular reasoning, as, in turn, a reversible process is one in which the equation holds (see details in section 6.11).

A decade later, Ludwig Boltzmann (1872, 1877; see also Swendsen, 2006) gave entropy a statistical content as he linked it to probabilities of statistical mechanical system states, thus explaining the Second Law of thermodynamics as the tendency of the system to run toward more probable states, which have higher entropy. In particular, Boltzmann (1872; English translation Boltzmann, 2003) did not literally write the entropy formula in its expression (2.12) but reached at a very similar one. Denoting $f(x, y, z, \xi_1, \eta_1, \dots w_r)\mathrm{d}x\ \mathrm{d}y\ \mathrm{d}z\ \mathrm{d}\xi_1 \dots \mathrm{d}w_r$ the number of molecules in a volume element $\mathrm{d}x\ \mathrm{d}y\ \mathrm{d}z$ at (x, y, z) at time t and at states described by the remaining variables, in his equation (78) he used (without naming it) the quantity

$$E = \iint \dots \int f \ln f\ \mathrm{d}x\ \mathrm{d}y\ \mathrm{d}z\ \mathrm{d}\xi_1 \dots \mathrm{d}w_r \tag{2.31}$$

Then he showed that this quantity cannot increase. If we change sign (i.e. substitute $-\ln f$ for $\ln f$) we have the entropy in its modern definition, which cannot decrease.

The probabilistic concept of entropy was advanced later in thermodynamics by Gibbs (1902), while Planck (1906, 1914) used a quantification of entropy identical to the modern one. The next important step was made by Shannon (1948) who used a definition essentially similar to Planck's to describe the information content, which he also called entropy, at von Neumann's suggestion (Robertson, 1993; Brissaud, 2005; Koutsoyiannis, 2011b). According to the latter definition, entropy is a probabilistic concept, a measure (as Shannon calls it) of information or, equivalently, of uncertainty. In the same year, in his famous book *Cybernetics*,[1] Wiener (1948a) used the same definition for information, albeit with the opposite sign (p. 62) because he regarded information as the negative of entropy (p. 11).

A few years later, von Neumann (1956) obtained virtually the same definition of entropy as Shannon, though in a slightly different way. Notably, as von Neumann, in addition to being a mathematician and computer scientist, was also a physicist, engineer and polymath, he clearly understood the connection of the probabilistic definition of entropy with its pre-existing physical content. Specifically, he wrote:

> *An important observation about this definition is that it bears close resemblance to the statistical definition of the entropy of a thermodynamical system.* […] *Pursuing this, one can construct a mathematical theory of the communication of information patterned after statistical mechanics.*

He also cited an earlier work in physics by Szilard (1929), who had implied the same definition of entropy in a thermodynamic system. Perhaps he was not aware that mathematical expressions similar to Shannon's and Szilard's had already appeared in a thermodynamic context in Boltzmann (1872, 1896/1898), Gibbs (1902) and especially Planck (1906, 1914).

The last fundamental contribution to the entropy concept was made a year later by Jaynes (1957), who introduced the *principle of maximum entropy*, which he formulated in a somewhat different way from that in section 2.3, namely:

> *in making inferences on the basis of partial information we must use that probability distribution which has maximum entropy subject to whatever is known.*

Since then, this principle has been used for *logical inference* as well as for *modelling physical systems*. The tendency of entropy to become maximal (Second Law of thermodynamics), which is the driving force of natural change, can result from this principle (see Chapter 6). Also, the same principle equips the entropy concept with a powerful tool for logical inference.

Turning to the mathematical details in the history of the entropy definition, we first observe that Shannon derived the entropy formula (2.12) by positing three postulates (or properties, as he calls them). Here we quote relevant passage noting that he uses the symbol H for entropy:

> *If there is such a measure, say $H(p_1, p_2, \dots, p_n)$, it is reasonable to require of it the following properties:*
>
> 1. *H should be continuous in the p_i.*
> 2. *If all the p_i are equal, $p_i = 1/n$, then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.*
> 3. *If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H. The meaning of this is illustrated in Fig. 6. At the left we have three possibilities $p_1 = 1/2, p_2 = 1/3, p_3 = 1/6$. On the right we first choose between two possibilities, each with probability 1/2, and if the second occurs make another choice with probabilities 2/3,1/3. The final results have the same probabilities as before. We require, in this special case, that $H(1/2,1/3,1/6) = H(1/2,1/2) + (1/2)H(2/3,1/3)$.The coefficient 1/2 is because this second choice only occurs half the time.*

We observe in this passage that the term "measure" was incorrectly used by Shannon (and also by all others except Khinchin, as detailed below). For, strictly speaking, entropy is not a measure, because it does not satisfy the properties defining a measure. Specifically, a measure should satisfy the axioms of probability, as stated in section 2.1, except for the normalization one, as probability is none other than a normalized measure. With respect to additivity (Kolmogorov's axiom III), this cannot be applied to entropy, simply because the union of two partitions is a set that is not a partition per se. Even if we interpret it loosely, we will have difficulty in identifying disjoint sets in a "mixture" of partitions so as to apply the axiom of additivity.

What is more, Shannon did not write down the mathematical form of his property 3 ("weighed sum" or grouping of events in consecutive steps)—not even in the proof of his theorem 2 (which he gives in his Appendix 2). This was done later, implicitly by Khinchin and explicitly by Jaynes as will be seen below. Quantification of his postulate 3 was also given in Robertson (1993, p. 3) and Uffink (1995; theorem 1), and is related to refinement of partitions to which the probabilities P_j refer.

Khinchin (1957) initially introduced entropy by its mathematical form, equivalent to the entropy formula (2.12), and proceeded to demonstrate a "Uniqueness Theorem" about it. In this, he used a different variant of Shannon's postulates as follows:

> 1. *For given n and for $\sum_{k=1}^{n} p_k = 1$ the function $H(p_1, p_2, \dots, p_n)$ takes its largest value for $p_k = 1/n$ $(k = 1, 2, ., n)$.*
> 2. *$H(AB) = H(A) + H_A(B)$.*
>
> *We add to these two properties a third, which obviously must be satisfied by any reasonable definition of entropy. Since the schemes*
>
> $$\begin{pmatrix} A_1 & A_2 & \dots & A_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix} \text{ and } \begin{pmatrix} A_1 & A_2 & \dots & A_n & A_{n+1} \\ p_1 & p_2 & \dots & p_n & 0 \end{pmatrix}$$
>
> *are obviously not substantively different, we must have*
>
> 3. *$H(p_1, p_2, \dots, p_n, 0) = H(p_1, p_2, \dots, p_n)$ (Adding the impossible event or any number of impossible events to a scheme does not change its entropy.)*

He then went on to prove his theorem that gives the entropy expression as in equation (2.12), also assuming that "*for any n this function is continuous with respect to all its arguments*". We note that in his postulate 2 he used the notion of conditional entropy, $H_A(B)$. This does not seem appropriate if the aim is to define entropy for the first time. (In our framework conditional entropy has been defined after entropy.) What he called a *scheme* looks similar to a partition, but

includes impossible events (empty sets), which deviates from our partition-based approach by allowing $P = 0$ explicitly—potentially leading to inconsistencies. In his definition of a (finite) scheme, he stated that it is a "*complete system of events* $A_1, A_2, \dots, A_n$ [...] *such that one and only one of them must occur at each trial*", along with their probabilities $p_1, p_2, \dots, p_n$. There may be imperfection here as the impossible event never occurs. Furthermore, in his proof he essentially used Shannon's "weighting sum" technique, which in essence is reflected in his property 2.

Jaynes, (2003, p. 347) reformulated Shannon's postulates as follows:

(1) We assume that some numerical measure $H_n(p_1, p_2, \dots, p_n)$ *exists; i.e., that it is possible to set up some kind of association between 'amount of uncertainty' and real numbers.*

(2) We assume a continuity property: H_n *is a continuous function of the* p_i*. Otherwise, an arbitrarily small change in the probability distribution would lead to a big change in the amount of uncertainty.*

(3) We require that this measure should correspond qualitatively to common sense in that, when there are many possibilities, we are more uncertain than when there are few. This condition takes the form that in the case that the p_i *are all equal, the quantity* $h(n) = H_n(1/n, 1/n, \dots, 1/n)$ *is a monotonic increasing function of n. This establishes the 'sense of direction'.*

(4) We require that the measure H_n *be consistent in the same sense as before; i.e., if there is more than one way of working out its value, we must get the same answer for every possible way.*

Like Shannon, Jaynes used the term "measure". His postulate (4) looks too verbal, but later on he quantifies it, essentially mathematizing Shannon's graphical depiction, in the following manner (equation (11.7) in his text, slightly modified here):

$$H_3(p_1, p_2, p_3) = H_2(p_1, p_2 + p_3) + (p_2 + p_3)H_2\left(\frac{p_2}{p_2 + p_3}, \frac{p_3}{p_2 + p_3}\right) \tag{2.32}$$

It can be seen that this equation is a special case of equation (2.28) in our Corollary 8.3. Jaynes (2003, p. 351) was fully aware that his last postulate and its mathematical expression are not satisfactory. Specifically, he wrote:

> *Although the above demonstration appears satisfactory mathematically, it is not yet in completely satisfactory form conceptually. The functional equation (11.7) does not seem quite so intuitively compelling as our previous ones did. In this case, the trouble is probably that we have not yet learned how to verbalize the argument leading to (11.7) in a fully convincing manner. Perhaps this will inspire others to try their hand at improving the verbiage that we used just before writing (11.7).*

Perhaps a better idea would be to abandon the postulate altogether and replace it with a more satisfactory one, as hopefully was done in section 2.3 above.

Furthermore, Jaynes continued presenting the "Wallis derivation" which is not a rigorous one—it is rather algorithmic that uses the Stirling approximation. He also cited a different approach by Shore and Johnson (1980), which is more interesting, but perhaps too complicated.

Another approach worth mentioning is that by Papoulis (1991, p. 533). Like others, Papoulis used the incorrect term "measure". Like our approach, Papoulis's was based on partitions of the ground set. He claimed that he used Shannon's postulates, whom he cited, after rephrasing them as follows:

1. $H(\mathbb{A})$ *is a continuous function of* $p_i = P(A_i)$.
2. *If* $p_1 = \cdots p_N = 1/N$*, then* $H(\mathbb{A})$ *is an increasing function of* N.
3. *If a new partition* $\mathbb{B}$ *is formed by subdividing one of the sets of* $\mathbb{A}$*, then* $H(\mathbb{B}) \geq H(\mathbb{A})$.

He continued stating:

> *It can be shown that the sum* [as in our equation (2.12)] *satisfies these postulates and it is unique within a constant factor. The proof of this assertion is not difficult but we choose not to reproduce it. We propose, instead, to introduce* [the entropic expression as in equation (2.12)] *as the definition of entropy and to develop axiomatically all its properties within the framework of probability.*

While it was correct to introduce the entropic expression from the outset as a definition of entropy, his previous statement is clearly incorrect. There is not a unique function that satisfies his, rather too loose, postulates. As a counterexample, we consider the function:

$$\Phi(\mathbb{A}) = 1 - \sum_{i=1}^{N} p_i^2 \tag{2.33}$$

This is a continuous function, satisfying his postulate 1. At the case of equiprobability, we have $\Phi_{\mathrm{E}}(N) = 1 - 1/N$, which is an increasing function of N with maximum value 1, thus satisfying his postulate 2. Furthermore, its second derivatives are $\partial^2\Phi(\mathbb{A})/\partial p_i^2 = -2$, $\partial^2\Phi(\mathbb{A})/\partial p_i\,\partial p_j = 0$ $(i \neq j)$, thus assuring concavity of $\Phi(\mathbb{A})$, which in turn ensures that his postulate 3 is satisfied.

As a second counterexample, we consider another function which violates even our non-interaction postulate, yet it fully satisfies Papoulis' postulates. This is:

$$\Phi(\mathbb{A}) = 1 - \sum_{i=1}^{N}\sum_{j=i}^{N} p_i p_j \tag{2.34}$$

Again, this is a continuous function, satisfying his postulate 1. At the case of equiprobability, we have $\Phi_{\mathrm{E}}(N) = (1 - 1/N)/2$, which is an increasing function of N with maximum value $1/2$, thus satisfying his postulate 2. Furthermore, its second derivatives are $\partial^2\Phi(\mathbb{A})/\partial p_i^2 = -2$, $\partial^2\Phi(\mathbb{A})/\partial p_i\,\partial p_j = -1$ $(i \neq j)$, thus assuring concavity of $\Phi(\mathbb{A})$, which in turn ensures that his postulate 3 is satisfied.

We will finally give two additional counterexamples, which are well known in literature as generalized entropies and both satisfy Papoulis' postulates. The first is the so-called *Rényi entropy* (after Rényi, 1961):

$$H_q^{\mathrm{R}}(p_1, \ldots, p_n) = \frac{1}{1-q}\ln\sum_{i=1}^{N} p_i^q \tag{2.35}$$

The second is the so-called *Tsallis entropy* (after Tsallis, 1988, who however was preceded by Havrda and Charvát, 1967):

$$H_q^{\mathrm{T}}(p_1, \ldots, p_n) = \frac{1}{q-1}\left(1 - \sum_{i=1}^{n} p_i^q\right) \tag{2.36}$$

Both have been introduced heuristically and contain the standard entropy (equation (2.12)) as a special case when $q \to 1$. Notice that equation (2.36) yields (2.33) when $q = 2$. Rényi and Tsallis entropies have been used in several applications but here we avoid their use preferring standard entropy, sometimes using a non-Lebesgue background measure to produce effects similar to those resulting from these generalized entropic forms but using the standard form.

The problems of the earlier approaches, as listed above, make room for a better introduction of entropy, which was attempted here.

---

[1] Interestingly, Wiener formed the celebrated term *Cybernetics* from the Greek word *κυβερνήτης*, meaning steersman, pilot, skipper, or governor, albeit incorrectly spelling it in his book (p. 11) as *χυβερνήτης*.

Digression 2.F: Entropy in stochastics vs. entropy in thermodynamics

More than one and a half century after the introduction of entropy, its meaning is still debated and a diversity of opinion among experts is encountered (Swendsen, 2011). In particular, despite having the same name, probabilistic (or information) entropy and thermodynamic entropy are still regarded by many as two distinct notions having in common only the name. The classical definition of thermodynamic entropy (as in Digression 2.E) gives no hint of similarity to probabilistic entropy. The fact that the latter is a dimensionless quantity and the former has units

(J/K) has been regarded as an argument that the two are dissimilar. Even Jaynes (2003), the founder of the maximum entropy principle, states:

> *We must warn at the outset that the major occupational disease of this field is a persistent failure to distinguish between the information entropy, which is a property of any probability distribution, and the experimental entropy of thermodynamics, which is instead a property of a thermodynamic state as defined, for example by such observed quantities as pressure, volume, temperature, magnetization, of some physical system. They should never have been called by the same name; the experimental entropy makes no reference to any probability distribution, and the information entropy makes no reference to thermodynamics. Many textbooks and research papers are flawed fatally by the author's failure to distinguish between these entirely different things, and in consequence proving nonsense theorems.*

Here one may notice that there is no "experimental entropy" nor "observational entropy" as others call it in the context of thermodynamics. There is no instrument to measure entropy. Entropy is an abstract quantity, inferred by measurement of other variables, such as temperature and pressure. Furthermore, the units of thermodynamic entropy are only an historical accident, related to the arbitrary introduction of temperature scales (Atkins, 2007). In a recent book, Ben-Naim (2008) has attempted to replace the concept of entropy altogether with the concept of information. Such a replacement is unnecessary, even meaningless (and opposite to von Neumann's suggestion to Shannon) if we accept that the two concepts are identical. As has recently been shown (Koutsoyiannis, 2013a, 2014a) and as will be discussed in detail in Chapter 6, the thermodynamic entropy can easily be produced by formal probability theory without using strange notions (like the indistinguishability of particles). The logical basis of the latter study includes the following points:

- The classical definition of thermodynamic entropy is not necessary. It can be abandoned and replaced by the probabilistic definition.
- Defined in this way, entropy is the fundamental thermodynamic quantity which supports the definition of all other derived ones. For example, the temperature is defined as the inverse of the partial derivative of entropy with respect to the internal energy (see section 6.8).
- Entropy retains its dimensionless character even in thermodynamics, thus rendering the unit of kelvin an energy unit.
- Entropy retains its probabilistic interpretation as quantified uncertainty, leaving aside the traditional but obscure 'disorder' interpretation (see Digression 2.G).
- Entropy is intrinsically related to the principle of maximum entropy, which expresses its tendency to reach its maximum permitted value given the available information about the system. The latter is incorporated into maximization in the form of constraints.
- The tendency of entropy to reach a maximum is the driving force of natural change. This tendency can be regarded as both a physical (ontological) principle obeyed by natural systems, and as a logical (epistemological) principle applicable when making inferences about natural systems.

Examples of deductive reasoning used in deriving thermodynamic laws from the formal probabilistic principle of maximum entropy have been provided in Koutsoyiannis (2014a) and Koutsoyiannis and Tsakalias (2025). These are further discussed in Chapter 6, where, notably, by maximizing entropy, i.e. uncertainty, at the microscopic level, we obtain physical laws that are virtually certain at the macroscopic level.

Digression 2.G: On different interpretations of entropy

In the public perception, entropy is a negative notion, typically identified with disorganization, disorder, decadence, decay, deterioration etc. (Koutsoyiannis and Sargentis, 2021). This misleading perception has its roots in the scientific community, albeit not with the founders of the concept (except one, as we shall see below). Boltzmann did not identify entropy with disorder,

even though he used 'disorder' in a footnote appearing in two of his papers (Boltzmann, 1897, 1901) speaking about the

> *agreement of the concept of entropy with the mathematical expression of the probability or disorder of a motion.*

Clearly, he referred to the irregular motion of molecules in the kinetic theory of gases, for which his expression makes perfect sense. Boltzmann also used the notion of disorder with the same meaning, in his Lectures on Gas Theory (Boltzmann,1896/1898). On the other hand, Gibbs (1902), Shannon (1948) and von Neumann (1956) did not use the terms disorder or disorganization at all.

One of the earliest uses of the term disorder is in a paper by Darrow (1944), in which he stated:

> *The purpose of this article has been to establish a connection between the subtle and difficult notion of entropy and the more familiar concept of disorder. Entropy is a measure of disorder, or more succinctly yet, entropy is disorder: that is what a physicist would like to say.*

Epistemologically, it is interesting that a physicist preferred the "more familiar" but fuzzy concept of disorder over the "subtle and difficult", yet well-defined at his time, concept of entropy.

However, it appears that Wiener (1948b) was the most influential scientist to support the disorder interpretation. In his keynote speech at the New York Academy of Sciences he declared that:

> *Information measures order and entropy measures disorder.*

Additionally, in his influential book *Cybernetics* (Wiener, 1948a, p. 11), he stated that

> *the entropy of a system is a measure of its degree of disorganization*

wherein he replaced the term "disorder" with "disorganization", as in this book he extensively used the former term for mental illness.

Even in the 21st century, the disorder interpretation is dominant. For example, Chaitin (2002) stated:

> *Entropy measures the degree of disorder, chaos, randomness, in a physical system. A crystal has low entropy, and a gas (say, at room temperature) has high entropy.*

More recently, Bailey (2009) claimed:

> *As a preliminary definition, entropy can be described as the degree of disorder or uncertainty in a system. If the degree of disorder is too great (entropy is high), then the system lacks sustainability. If entropy is low, sustainability is easier. If entropy is increasing, future sustainability is threatened.*

It is relevant to remark that in the latter quotations disorder was used as equivalent to uncertainty or randomness—where the latter two terms are in essence identical (Koutsoyiannis, 2010). Furthermore, the claim that a high-entropy system lacks sustainability is puzzling, given that the highest entropy occurs when a system is in the most probable and hence the most stable state (cf. Moore, 2002).

Interestingly, Atkins (2003) also explained entropy as disorder. Additionally, he noted:

> *That the world is getting worse, that it is sinking purposelessly into corruption, the corruption of the quality of energy, is the single great idea embodied in the Second Law of thermodynamics.*

Inevitably, the notion of entropy is hard to grasp, the main reason being that our education is based on the deterministic paradigm and produces a mindset reluctant to incorporate stochastic concepts. The determinist mindset regards order as a friendly concept. Thus, whatever is defined as the opposite appears in a negative light.

However, the notions of order and disorder are less appropriate and less rigorous as scientific terms, and more appropriate for describing mental states (as in Wiener's use described above; cf.

personality disorder, stress disorder, bipolar disorder, mental disorder), and even more so in describing socio-political states. The latter is manifest in the frequent use of expressions such as "world order", "new order", "new world order", "global order", etc., in political texts (Koutsoyiannis and Sargentis, 2021).

In one of the earliest critiques of the disorder interpretation of entropy, Wright (1970) made a plea for moderation in the use of "intuitive qualitative ideas concerning disorder". More recently, with a more absolute tone, Leff (2012) stated:

> *The too commonly used disorder metaphor for entropy is roundly rejected.*

Furthermore, in an even more recent article, Styer (2019) stated:

> *we cannot stop people from using the word "entropy" to mean "disorder" or "destruction" or "moral decay." But we can warn our students that this is not the meaning of the word "entropy" in physics.*

Styer attributed an excessive contribution to the misconception of entropy as disorder to the autobiographical book "The Education of Henry Adams" (Adams, 1918). As he asserted, that book proved to be enormously influential, as it won the 1919 Pulitzer Prize for biography, and in April 1999 was named by the Modern Library the 20th century's best nonfiction book in English. As quoted by Styer, Adams disliked chaos and anarchy, and stated:

> *The kinetic theory of gas is an assertion of ultimate chaos. In plain words, Chaos was the law of nature; Order was the dream of man.*

This is a very strong statement, contrasting Nature with man and also implying that there is only one type of order that dreamt up by man—a rather naïve idea.

Those viewing entropy as disorder have difficulty understanding the concept of life. In early 20th century, the Swiss physicist C.-E. Guye (1922) asked the question: How is it possible to understand life, when the whole world is ruled by such a law as the second principle of thermodynamics, which points toward death and annihilation? He was followed by many other scientists who were puzzled by the existence of life. As insightfully discussed by Brillouin (1949), scientists of the era wondered if there was a "life principle", a new and unknown principle that would explain life as an entity countering the second law of thermodynamics. A year later, Brillouin (1950) coined the term *negentropy* as an abbreviation of *negative entropy*. In this, he used information theoretical concepts to express the idea that every observation in a laboratory requires the degradation of energy, and is made at the expense of a certain amount of negentropy, taken away from the surroundings.

The term *negative entropy* had earlier been used by Schrödinger (1944) in his famous book "What is life?". Specifically, he argued that "What an organism feeds upon is negative entropy" without providing another "life principle" additional to the Second Law that would drive life and evolution.

There is no general agreement about the meaning of *negative entropy* or *negentropy*. Some (e.g., Lago-Fernández and Corbacho, 2009) use them as technical terms referring to the difference between the entropy of any variable and that of a variable with normal distribution, with the same mean and variance (distance to normality). However, others, in a rather metaphysical context and assuming a non-statistical definition of negentropy (e.g., Larouche, 1993) saw a negentropic principle governing life, the biosphere, the economy, etc., because these convert things that have less order into things with more order.

Today it makes sense to ask: Has this question been answered yet? Or is it even relevant, one hundred years after? Perhaps it is relevant to quote here Atkins (2003), who, as we have seen, explained entropy as disorder. Yet he neatly remarked:

> *The ceaseless decline in the quality of energy expressed by the Second Law is a spring that has driven the emergence of all the components of the current biosphere.* […] *The spring of change is aimless, purposeless corruption, yet the consequences of interconnected change are the amazingly delightful and intricate efflorescences of matter we call grass, slugs, and people.*

Apparently, if we abandon the disorder interpretation of entropy, we could also stop seeking a negentropic "life principle", which was never found and probably will never be. For, if we see entropy as uncertainty, we also understand that life is fully consistent with entropy maximization. Human-invented steam engines (and other similar machines) increase entropy all the time, and are fully compatible with the Second Law, yet they produce useful work. Likewise, the biosphere increases entropy, yet it produces interesting patterns, much more admirable than steam engines. Life generates new options and increases uncertainty (Sargentis et al., 2020; Koutsoyiannis and Sargentis, 2021). Compare Earth with a lifeless planet: Where is uncertainty greater? On which of the two planets would a newspaper have more events to report every day?

However, if entropy is not disorder, what is its consistent interpretation? This question is not as difficult to answer as the above discussion seems to imply. According to its standard definition (section 2.3), entropy is uncertainty quantified. Hence, maximum entropy means the maximum uncertainty that is allowed in natural processes, given the constraints implied by natural laws (or human interventions). It should be stressed that, with this general definition, entropy and its maximization do not apply only to physics—in particular to thermodynamics—but also to any natural (or even uncontrolled artificial) process in which there is uncertainty necessitating a (macroscopic) probabilistic description. This application is not meant as an "analogy" with physics. Rather, it is a formal application of the general definition of entropy, which relies on stochastics.

Unsurprisingly, if "disorder" is regarded by many as a "bad thing", the same could be said for uncertainty. The expressions "uncertainty monster" and "monster of uncertainty" appear in about 250 scholarly articles registered in Google Scholar (samples are van der Sluijs, 2005, and Curry and Webster, 2011, to mention a couple of the most cited with the word "monster" appearing in their title). However, if uncertainty is a monster, it is thanks to this monster that life is liveable and fascinating. Uncertainty is not an enemy of science or of life. Rather, it is the mother of creativity and evolution. Without uncertainty, life would be a "universal boredom" (to borrow a phrase by Saridis, 2004, and reverse its connotation), and concepts such as hope, will (particularly, free will), freedom, expectation, optimism, etc., would hardly make sense. A technocratic system wherein an elite of super-experts who, using super-models, could predict the future without uncertainty, would also take full control of the society (Koutsoyiannis et al., 2008). Fortunately, this will never happen because entropy, i.e., uncertainty, is a structural property of Nature and Life. Hence, in our view, uncertainty is neither disorder nor a "bad thing". How could the most important law of physics (the Second Law) be a "bad thing"?

In a deterministic worldview there is no uncertainty, and therefore no point in speaking about entropy. If there is no uncertainty, each outcome can be accurately predicted, and hence there are no options. In contrast, in an indeterministic world, there is a plurality of options. This corresponds to the Aristotelian idea of *δύναμις* (Latin: *potentia*—English: *potency* or *potentiality*). The existence of options entails there is freedom. Thus:

$$\textit{entropy} \leftrightarrow \textit{uncertainty} \leftrightarrow \textit{plurality of options} \leftrightarrow \textit{freedom}$$

To conclude this Digression: When discussing entropy, we should always bear in mind that entropy per se is a probabilistic concept based fundamentally on a macroscopic view of phenomena, rather than focusing on individuals or small subsets, and hence temporal or spatial scale is important to consider. Analysing a particular die-throw, we may say that it was irregular, uncertain, unpredictable, chaotic, or random. However, macroscopization, by removing the details, may also remove irregularity. For example, the application of the principle of maximum entropy to the outcomes of a die-throw results in equal probabilities (1/6) for each outcome (see Digression 2.L), and the average outcome after many throws tends to be 3.5. This is a regular macroscopic result. In precisely the same manner, the maximum uncertainty in a particular water molecule's state (in terms of position, kinetic state and phase), on a macroscopic scale results in the Clausius–Clapeyron law (see sections 6.20 and 6.21). Again, we have perfect regularity, as the accuracy of this law is so high that most people believe that it was a deterministic law.

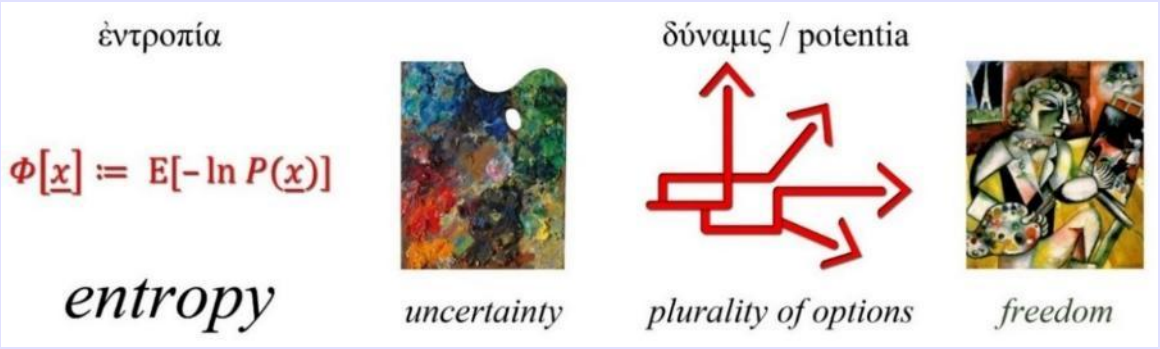

Figure 2.1 An attempt at an artistic representation of the notion of entropy from Koutsoyiannis and Sargentis (2021). The entropic formula shown on the left is justified in section 2.10; uncertainty is depicted by Marc Chagall's Palette (adapted from https://www.metmuseum.org/art/collection/search/493731) and freedom by Marc Chagall's *Self-Portrait with Seven Fingers* (https://en.wikipedia.org/wiki/Self-Portrait_with_Seven_Fingers); δύναμις (Greek) or *potentia* (Latin) is the Aristotelian idea of potency or potentiality.

2.4 The concept of a stochastic variable

A *stochastic variable* or *random variable*[*] is a function that maps outcomes to numbers, i.e. enumerates the ground set Ω. More formally, according to Kolmogorov's (1933) definition, a real single-valued function $x(\omega)$, defined on the ground set Ω, is called a *random variable* if for each choice of a real number a, the set $\{x(\omega) < a\}$ for all ω for which the inequality $x(\omega) < a$ holds true, belongs to Σ. With the concept of the stochastic variable, we can conveniently express events using basic mathematics. In most cases enumeration is done almost automatically. For instance, a stochastic variable that takes values 1 to 6 is intuitively assumed when we deal with a die throw experiment. If the phenomenon we study is related to the physical world and the quantity being studied is represented as a real number, then this real number (e.g. a) has some dimension (e.g. length) and hence a physical unit (e.g. m) associated with it. It is convenient to extend the notion of the stochastic variable to also include the same unit, so that $\{x(\omega) < a\}$ become meaningful.

We must be attentive to the fact that a stochastic variable is not a number but a function. Intuitively, we could think of a stochastic variable as an object that represents *simultaneously* all possible outcomes and only them. The following analogy may help us to develop intuition about stochastic variables. Let us consider the equation $x^3(x-1)^2 = 0$. This has five roots, three of them being $x = 0$ and two being $x = 1$. What do we mean when we say "a root of this equation"? Probably we mean both $x = 0$ and $x = 1$ and also we have in mind that there is no symmetry between the two; rather we would give a weight $3/5$ on the former and $2/5$ on the latter. Similar is the situation for a stochastic variable which takes on the values 0 and 1 with probabilities $3/5$ and $2/5$, respectively. In other words, the function $\mathrm{NR}(x)$, which yields the relative multicity of roots of the above fifth-order equation at the point x, can readily be compared to the function $P(x)$, which assigns values of probabilities in the probabilistic example.

[*] The two terms *stochastic variable* and *random variable* have identical meaning. Here we prefer the former, even though the latter is more common.

While formally a stochastic variable is a function $x(\omega)$, we usually omit the reference to its argument ω and keep the symbol x. However, in this case we need to distinguish it symbolically from a common variable. The best notation devised for this, and used here, is the so-called Dutch convention (see Hemelrijk, 1966, who mentions that it was introduced by D. Van Dantzig in 1947, i.e., later than Kolmogorov's foundation of probability). According to it, stochastic variables are underlined, i.e. $\underline{x}$. In this case the inequality $\{x(\omega) < a\}$ used for the formal definition of the stochastic variable is written as $\{\underline{x} < a\}$. Accordingly, $\{\underline{x} < a\}$ denotes an event (a subset of Ω), and therefore it has a probability, $P(\{\underline{x} < a\})$. For simplicity, in the latter notation we drop the parenthesis and we write $P\{\underline{x} < a\}$. Some texts drop the curly brackets instead of the parentheses, but this practice misrepresents the important point that the argument of probability is a set. The notation is further explained in Digression 2.H, along with its importance.

From a practical point of view, compared to a common variable, a stochastic variable is a more abstract mathematical entity which we use when a quantity of interest is something uncertain, unpredictable, unknown; this is the meaning of *stochastic* and *random* (cf. Koutsoyiannis, 2010; Dimitriadis et al., 2016). While a common variable takes on one value at a time, a stochastic variable can be thought of as taking on all of its possible values at once, but not necessarily in a uniform manner. Therefore, a probability distribution function, to be defined in section 2.5, should always be associated with a stochastic variable. A stochastic variable becomes identical to a common variable only if it can take on only one value.

When an observation of a quantity that is modelled as a stochastic variable is made, then this observation is usually a common variable. For example, we model a die throw with a stochastic variable $\underline{x}$ with possible values 1 to 6. After a specific throw of the die and before we observe the outcome, we still have the same uncertainty as described by stochastic variable $\underline{x}$. When we observe the outcome, it becomes a common variable x (e.g. $x = 5$). The particular value is called a *realization* of $\underline{x}$ and is denoted by the non-underlined symbol x. This happens when our observation is exact. Sometimes the observation is contaminated by error—our observations are not always exact (particularly those of real valued variables). Then we can use another stochastic variable to describe the uncertain outcome. For example, if an observer has presbyopia combined with astigmatism (like the author) he may not be sure whether the outcome was 5 or 4 and he could model it as a stochastic variable $\underline{z}$ with possible outcomes 4 and 5.

Considering a certain (deterministic) function $y = g(x)$, mapping the common variable x to the common variable y (e.g. $y = g(x) = x^2$), we can extend its meaning to apply to stochastic variables, i.e., $\underline{y} = g(\underline{x})$ (e.g. $\underline{y} = g(\underline{x}) = \underline{x}^2$). As implied by the notation, when the function's argument $\underline{x}$ is a stochastic variable, the result $\underline{y}$ is also a stochastic variable (formally, it is the composite function $y(\omega) = g(x(\omega))$. In other words, functions of stochastic variables are stochastic variables.

Digression 2.H: The importance of notation

The following simple example shows that the common practice of not distinguishing the notation of common and stochastic variables is bad practice. Let $\underline{x}$ and $\underline{y}$ represent the outcomes of each of two dice. What is the probability of the following cases?

$$\text{(a) } \left\{\underline{x} < \underline{y}\right\}, \qquad \text{(b) } \{\underline{x} < y\}, \qquad \text{(c) } \left\{x < \underline{y}\right\}, \qquad \text{(d) } \{x < y\}.$$

(a) First, we clarify that the event $\{\underline{x} < \underline{y}\}$ includes all elementary events (x, y) in which $x < y$.* There are $6^2 = 36$ different possible combinations of outcomes of $\underline{x}$ and $\underline{y}$. In six of them $\underline{x} = \underline{y}$. Due to symmetry, in half of the remaining 30, $\underline{x} < \underline{y}$. Thus:

$$P\left\{\underline{x} < \underline{y}\right\} = \frac{15}{36} = \frac{5}{12}$$

(b) Now y is a number, not a stochastic variable. For convenience we assume that y is an integer, even though it can also be assumed to be real. If $y > 6$ then obviously the event $\left\{\underline{x} < y\right\}$ is certain. If $y = 6$ then the probability of $\left\{\underline{x} < y\right\}$ is $5/6$. Continuing like this we conclude that:

$$P\{\underline{x} < y\} = \max\left(0, \min\left(1, \frac{y-1}{6}\right)\right)$$

(c) Thinking as in (b) and noting that x is a number, an assumed integer, and $\underline{y}$ a stochastic variable we find that:

$$P\left\{x < \underline{y}\right\} = \max\left(0, \min\left(1, 1 - \frac{x}{6}\right)\right)$$

(d) As both x and y are numbers, the expression $\{x < y\}$ does not denote an event and therefore, strictly there is no probability associated with this expression. Loosely we may say that $P\{x < y\} = 1$ if $x < y$ and 0 otherwise.

Obviously, if we did not distinguish y from $\underline{y}$, we would not even be aware of the fact that $P\left\{\underline{x} < \underline{y}\right\}$ is a number while $P\left\{x < \underline{y}\right\}$ is a function of x.

Many texts (research articles and probability theory books) make the notational distinction between stochastic and common variables, but they use upper case letters for stochastic variables and lower case ones for common variables. This practice may also be inadequate. If in our context we used another quantity denoted by the Greek letter χ (and actually χ is quite common in statistical texts—cf. the chi and chi-squared distributions), how would we distinguish the stochastic variables corresponding to x and χ? (In both cases the upper case letter is X, while in our convention $\underline{x}$ and $\underline{\chi}$ are distinguishable.) Furthermore, this would be too restrictive in our use of mathematical symbols. For example, the symbol H typically used to denote the Hurst parameter would be an incorrect notation if we adopted the upper- vs. lower-case notation. Another convention was used by Papoulis (1990, 1991), who denoted stochastic variables in bold letters. However, the typical use of bold letters is to denote vectors. Therefore, the Dutch convention of underlining the stochastic variables is the most convenient, clearest and safest.

———

* More generally, here we interpret an event $\left\{\underline{x} < \underline{y}\right\}$ as equivalent to the event $\{\underline{z} < 0\}$, where $\underline{z} \coloneqq \underline{x} - \underline{y}$. For a more formal discourse on stochastic inequalities and ordering see Shaked and Shanthikumar (2007).

2.5 Distribution function, tail function, odds function and quantile

According to Kolmogorov's (1933) foundation* of probability theory, the function of the real variable x,

$$F(x) \coloneqq P\{\underline{x} \leq x\} \tag{2.37}$$

where $\underline{x}$ is a stochastic variable, is called the *distribution function*. We notice that the stochastic variable with which this function is associated is not an argument of the function. Even though we use the same letter for both $\underline{x}$ and x, the two are fundamentally different. For example, in a die throw, the stochastic variable $\underline{x}$ represents the whole numbers 1 to 6 and the common variable x takes on any real value from –∞ to +∞. (The domain of $F(x)$ is not identical to the range of the stochastic variable $\underline{x}$; rather it is always the set of real numbers.†) If there is risk of confusion (e.g., if we study a problem with many stochastic variables), the stochastic variable should also appear in the notation of the distribution function. Usually, it is denoted as a subscript: $F_{\underline{x}}(x)$.

Typically, $F(x)$ has a mathematical expression depending on some parameters. It is a non-decreasing function of x obeying the relationship:

$$0 \;=\; F(-\infty) \leq\; F(x) \leq\; F(+\infty) \;=\; 1 \tag{2.38}$$

As it is a non-decreasing function, in the English literature $F(x)$ is also known as *cumulative distribution function*, but here we adhere to Kolmogorov's (1933) original terminology, which did not contain the adjective *cumulative*. In practical applications the distribution function is also known as *non-exceedance probability*. Likewise, the non-increasing function:

$$\overline{F}(x) \coloneqq P\{\underline{x} > x\} = 1 - F(x) \tag{2.39}$$

i.e., the complement of $F(x)$ from 1, is called here the *distribution function complement*. It is also known as *tail function*, *survival function*, or *survivor function*, and represents *exceedance probability.*

The distribution function is always continuous on the right. However, if the ground set Ω is finite or countable, $F(x)$ is discontinuous on the left at all points x_i that correspond to outcomes ω_i, and it is constant between them (staircase-like). Such a stochastic variable is called *discrete*. If $F(x)$ is a continuous function, then the stochastic variable is called *continuous*. A *mixed* case is also common. Here, the distribution function has some discontinuities on the left, but is not staircase-like. These are better explained in Digression 2.J.

A useful derived function is the so-called *odds function*:

---

* We note that Kolmogorov used '<' in his definition but modern literature uses '≤' as in (2.37).

† A generalization for a complex stochastic variable $\underline{z} \coloneqq \underline{x} + i\underline{y}$ is possible and useful in some cases (see section 5.4). In this case, $\underline{z}$ does not have a distribution function because the inequality $\underline{x} + i\underline{y} \leq x + iy$ does not have a meaning. Instead, we use the joint distribution function of $\underline{x}$ and $\underline{y}$ (see section 2.13). Nonetheless, other types of statistical characterization of $\underline{z}$ (e.g. expectation) are meaningful.

$$\Psi(x) := \frac{F(x)}{\overline{F}(x)} = \frac{F(x)}{1 - F(x)} \tag{2.40}$$

Like $F(x)$, $\Psi(x)$ is a nondecreasing function. For continuous stochastic variables, the inverse function $F^{-1}(\,)$ of $F(\,)$ exists. Consequently, the equation $u = F(x)$ has a unique solution for x, called *u-quantile* of the variable $\underline{x}$, that is:

$$x_u = F^{-1}(u) \tag{2.41}$$

Digression 2.I: The concept of return period

In several applications to geophysics and in engineering applications the concept of return period is widely used. This is closely related to probability and it offers some advantages in terms of more intuitive probability plots and engineering design procedures. The return period T of an event, which has probability P to occur in a time interval D, is related to P and D by the almost obvious relationship:

$$P = \frac{D}{T} \tag{2.42}$$

Apparently, T has dimensions of time. If the event of interest is the exceedance of a threshold value x (e.g., a value whose exceedance results in some risk as in a heatwave or a flood), i.e. the event $\{\underline{x} > x\}$ occurring within a time interval D, then

$$T(x) = \frac{D}{\overline{F}(x)} = \frac{D}{1 - F(x)} \tag{2.43}$$

Conversely, if the event of interest is the non-exceedance of a threshold value x (e.g., a value whose non-exceedance results in some risk, as in a cold wave or a drought), i.e. the event $\{\underline{x} \le x\}$ occurring within a time interval D, then

$$\overline{T}(x) = \frac{D}{F(x)} \tag{2.44}$$

The concept offers an easy means of empirical estimation of T provided that there is a sample of n observations taken at equidistant times D. Even though we have not formally introduced yet the concepts of sample and estimation (this will be made in Chapter 4) we provide here the following equations for easy estimate of T of the sample value $x_{(i:n)}$, which is the ith smallest of the n values. For positively symmetric distributions:

$$\frac{\hat{T}\left(x_{(i:n)}\right)}{D} = \frac{n + \mathrm{e}^{1-\gamma} - 1}{n - i + \mathrm{e}^{-\gamma}} = \frac{n + 0.526}{n - i + 0.561} \tag{2.45}$$

and for symmetric distributions

$$\frac{\hat{T}\left(x_{(i:n)}\right)}{D} = \frac{n + 2\mathrm{e}^{1-\gamma} - 1}{n - i + \mathrm{e}^{-\gamma}} = \frac{n + 0.123}{n - i + 0.561} \tag{2.46}$$

where $\gamma = 0.5772$ is the Euler constant and the caret (^) denotes an estimate. This gives an unbiased estimate of the logarithm of T, as well as for the distribution quantiles of distributions of exponential type, and was developed in Koutsoyiannis (2025). We will give more details in section 4.12.

2.6 Probability mass and density function

In discrete stochastic variables, the probability of each event:

$$P_j \equiv P(x_j) \coloneqq P\{\underline{x} = x_j\} = F(x_j) - F(x_{j-1}), \qquad j = 1, \dots, J \tag{2.47}$$

where J is the number of possible outcomes (which can be infinite), is the *probability mass function*. It is easy then to see that the step (discontinuity) of the distribution function $F(x)$ at point x_j equals P_j.

In continuous variables there are no discontinuities and hence any particular value x has zero probability of occurring. However, we can still tell which of two outcomes is more probable and by how much, by examining the ratio of the two probabilities. As this is a $0/0$ expression, bearing in mind l'Hôpital's rule, we need to examine the ratio of derivatives of probabilities.

The derivative of the distribution function is called the *probability density function* (PDF) or simply *density*:

$$f(x) \coloneqq \frac{\mathrm{d}F(x)}{\mathrm{d}x} \tag{2.48}$$

and its basic properties are:

$$f(x) \geq 0, \qquad \int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1 \tag{2.49}$$

Obviously, the probability density function does not represent a probability and, therefore, it can take on values higher than 1. Its relationship with probability is described by the following equation:

$$f(x) = \lim_{\Delta x \to 0} \frac{P\{x \leq \underline{x} \leq x + \Delta x\}}{\Delta x} \tag{2.50}$$

The distribution function can be calculated from the density function, i.e.:

$$F(x) = \int_{-\infty}^{x} f(y)\mathrm{d}y \tag{2.51}$$

In discrete stochastic variables, the density is a sequence of Dirac δ functions (see definition of δ in equation (3.52)), while in mixed distributions Dirac δ functions appear at the points of discontinuity. This text mostly deals with continuous variables.

Some of the most common distributions of discrete and continuous variables are shown in Table 2.2. Additional continuous distributions are shown in Table 2.3, along with their moments, while the derivation of these and other distributions in terms of the principle of maximum entropy is discussed in section 2.11 (see also Table 2.4 and Table 2.5).

As already discussed (section 2.4), the one-to-one mathematical transformation of $\underline{x}$, $\underline{y} = g(\underline{x})$ defines a new stochastic variable $\underline{y}$. If the function $g(x)$ is invertible, then the event $\{\underline{y} \leq y\}$ is identical to the event $\{\underline{x} \leq g^{-1}(y)\}$ where g^{-1} is the inverse function of g. Consequently, the distribution functions of $\underline{x}$ and $\underline{y}$ are related by:

$$F_{\underline{y}}(y) = P\{\underline{y} \leq y\} = P\{\underline{x} \leq g^{-1}(y)\} = F_{\underline{x}}(g^{-1}(y)) \tag{2.52}$$

In the case that the variables are continuous and the function g differentiable, it can be shown that the density functions of $\underline{x}$ and $\underline{y}$ are related by:

$$f_{\underline{y}}(y) = \frac{f_{\underline{x}}\left(g^{-1}(y)\right)}{\left|g'\left(g^{-1}(y)\right)\right|} \tag{2.53}$$

where g' is the derivative of g.

Table 2.2 Some of the simplest and most common distributions.

Name (and parameters)	Probability mass function or probability density function	Probability distribution function
Discrete variable $\underline{x}$ with values $x_j \equiv j$		
Discrete uniform, $j = 1, \ldots, J$	$P(x_j) = \frac{1}{J}$	$F(x) = \max(0, \min(\lfloor x \rfloor / J, 1))$
Geometric $j = 0,1, \ldots (\mu > 0)$	$P(x_j) = \frac{1}{1+\mu}\left(\frac{\mu}{1+\mu}\right)^j$	$F(x) = \max\left(0, 1 - \left(\frac{\mu}{1+\mu}\right)^{\lfloor x \rfloor + 1}\right)$
Poisson $j = 0,1, \ldots (\mu > 0)$	$P(x_j) = e^{-\mu}\frac{\mu^j}{j!}$	$F(x) = e^{-\mu}\sum_{j=0}^{\lfloor x \rfloor}\frac{\mu^j}{j!} = \frac{\Gamma_{\lfloor x+1 \rfloor}(\mu)}{\lfloor x \rfloor!}$
Continuous variable $\underline{x}$		
Uniform in $[0, J]$	$f(x) = \begin{cases} 1/J, & \text{for } 0 \le x \le J \\ 0, & \text{otherwise} \end{cases}$	$F(x) = \max(0, \min(x/J, 1))$
Exponential $(\mu > 0)$	$f(x) = \begin{cases} \mathrm{e}^{-x/\mu}/\mu, & x \ge 0 \\ 0, & x < 0 \end{cases}$	$F(x) = \begin{cases} 1 - \mathrm{e}^{-x/\mu}, & \text{for } x \ge 0 \\ 0, & \text{for } x < 0 \end{cases}$
Normal $(\mu \in \mathbb{R}, \sigma > 0)$	$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$F(x) = \frac{1}{2}\mathrm{erfc}\left(-\frac{x-\mu}{\sqrt{2}\sigma}\right)$

Note: $\lfloor x \rfloor$ denotes the floor of the number x (the greatest integer less than or equal to x).

Digression 2.J: Illustration of distribution function by an example

Here we continue the example in Digression 2.B to illustrate the notion of the stochastic variable and its distribution function. Again we study: (a) the occurrence of rainfall at a particular site and a specific time of the year, and (b) the rainfall depth at that site and time.

(a) The ground set is $\Omega = \{\text{wet, dry}\}$ and we define a stochastic variable $\underline{x}$ based on the rule

$$x(\text{dry}) = 0, \qquad x(\text{wet}) = 1$$

We can now easily determine the distribution function of $\underline{x}$. For any $x < 0$,

$$F(x) = P\{\underline{x} \le x\} = 0$$

(because $\underline{x}$ cannot take negative values). For $0 \le x < 1$,

$$F(x) = P\{\underline{x} \le x\} = P\{\underline{x} = 0\} = 0.8$$

Finally, for $x \ge 1$,

$$F(x) = P\{\underline{x} \le x\} = P\{\underline{x} = 0\} + P\{\underline{x} = 1\} = 1$$

The graphical depiction of the distribution function is shown in Figure 2.2 (left). The staircase-like shape reflects the fact that the stochastic variable is discrete.

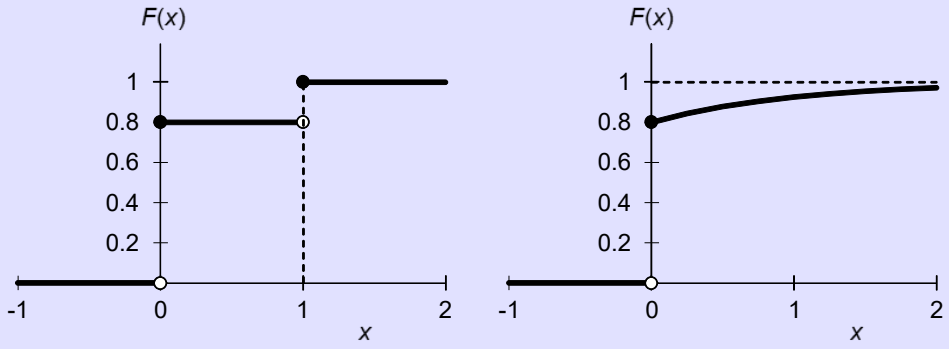

Figure 2.2 Distribution function of a stochastic variable representing events related to rainfall on a given day in a certain area and at a specific time of the year: **(left)** the dry or wet state; **(right)** the rainfall depth.

(b) The ground set is $\Omega = \mathbb{R}^+ \cup \{0\}$ and the stochastic variable $\underline{x}$ is given by the rule $x(\omega) = \omega$. Again, the distribution function of $\underline{x}$ will be $F(x) = P\{\underline{x} \le x\} = 0$ for $x < 0$ with a discontinuity at 0, so that $F(0^+) = P\{\underline{x} = 0\} = 0.8$. For $x \ge 0$ the distribution function will be continuous and increasing, approaching 1 as $x \to \infty$. To construct a plausible distribution function, without examining observations, we make an assumption that smaller values are more probable than higher and specifically that for two values x_1 and $x_2 > x_1$, the ratio of densities (expressing the ratio of probabilities according to l'Hôpital's rule) depends on the difference $x_2 - x_1$, i.e.,

$$\frac{f(x_1)}{f(x_2)} = g(x_2 - x_1)$$

where it is easy to see that the function $g(\,)$ should be given as $g(x) = f(0)/f(x)$. In turn, it can be shown (homework) that $f(x) = A\exp(-Bx)$ where A and B are constants. By integrating (according to equation (2.51)) we find:

$$F(x) = \frac{A}{B}(1 - \exp(-Bx)) + C$$

and, since $F(0^+) = 0.8$ and $F(\infty) = 1$, $C = 0.8$ and $A/B = 0.2$, thus:

$$F(x) = 0.2(1 - \exp(-Bx)) + 0.8$$

where B can be any positive number. An example is depicted in Figure 2.2 (right) for $B = 1$. The result is a modified exponential distribution (see Table 2.2), where the modification resulted from the fact that the distribution is not continuous everywhere but mixed. The same result could be derived by maximizing entropy (see Digression 2.L)

If this mathematical model is to represent a physical phenomenon, we must keep in mind that all probabilities depend on a specific location and a specific time of the year. So, the model cannot be a global representation of the wet and dry state of a day, nor of the rainfall depth. The model as formulated here is extremely simplified. It does not make any reference to the succession of dry or wet states on different days. This is not an error; it simply diminishes the predictive capacity of the model. A better model would describe separately the probability of a wet day following a wet day, a wet day following a dry day (we anticipate that the latter should be smaller than the former), etc. In addition, while the assumption made for the rainfall depth leading to a mixed exponential distribution seems plausible at a first glance, it does not fully correspond to the empirically observed behaviour. There are better models than the exponential.

2.7 Random number generation for stochastic simulation

One of the important scientific advances offered by stochastics in the last several decades is the Monte Carlo method, else known as stochastic simulation. It was originally developed for the numerical solution of integro-differential equations in Los Alamos in the framework of the Manhattan Project (Metropolis and Ulam, 1949). It can easily be shown (e.g. Niederreiter, 1992) that in high dimensional numerical integration (specifically for a number of dimensions $d > 4$), a stochastic (Monte Carlo) integration method (in which the function evaluation points are taken at random) is more accurate (for the same total number of evaluation points) than classical numerical integration (based on a grid representation of the integration space).

This gave importance to the much older concept of random numbers, whose first appearance in a scientific publication was Tippett's (1927) table, with 41 600 random digits taken from a 1925 census report. Before that (and even after; see Digression 3.G) random sampling was performed by means of dice and cards. Thus, Galton (1890) invented a set of three modified dice to generate samples from a normal distribution. "Student" (pseudonym of W.S. Gosset) in 1908 performed simulation experiments using 3000 cards (in 750 groups of size 4) to find the distribution of the t-statistic and of the correlation coefficient (see more information in Stigler, 2002).

A *sequence of random numbers* is a sequence of numbers x_i whose every statistical property is consistent with that of realizations from a sequence of independent identically distributed stochastic variables $\underline{x_i}$ with specified distribution function $F(x)$ (definition adapted from Papoulis, 1990). In turn, a *random number generator* is a device (typically computer algorithm) that generates a sequence of random numbers x_i with given distribution $F(x)$. Random number generation is also known as *Monte Carlo sampling*.

The basis of practically all random generators is the uniform distribution in [0,1] (see Table 2.2). A typical procedure for that distribution is the following:

- We generate a sequence of integers q_i from the recursive algorithm $q_i = (k\, q_{i-1} + c) \bmod m$ where k, c and m are appropriate integers (e.g. $k = 69\,069$, $c = 1$, $m = 2^{32} = 4\,294\,967\,296$ or alternatively $k = 7^5 = 16\,807$, $c = 0$, $m = 2^{31} - 1 = 2\,147\,483\,647$; Ripley, 1987, p. 39).
- We calculate the sequence of random numbers u_i with uniform distribution in [0,1] as $u_i = q_i/m$.

Obviously, this is a simple algorithm, purely deterministic. Why are the numbers it generates regarded as random? The answer is simple: Because if we do not know the algorithm and the initial condition (q_0 or q_{i-1}) we cannot predict these numbers. As most algorithms, like this one, are purely deterministic, sometimes the numbers are called pseudorandom. But this implies the idea that there exists another category of true or genuine random numbers. Even though in the literature references to true random numbers abound, this may reflect a misunderstanding of the notion of randomness and a dichotomic view of natural processes (cf. Koutsoyiannis, 2010; Dimitriadis et al., 2016). In any process of the macroscopic world, if we were able to know the "algorithm" (the system dynamics), and the initial conditions with full precision, the situation would be the

same as with the simple algorithm described. That we are unable to know precisely the algorithm of a physical process and the initial conditions does not make the numbers of different type.

A more recent algorithm for the generation of random numbers with uniform distribution is the so-called *Mersenne twister*, which is better due to its longer period (it is based on the Mersenne prime $2^{19\,937} - 1$) and has better equidistribution properties compared to linear congruential generators. It is available in most computer languages and software packages*.

Once we have a random generator for the uniform distribution, we can make one for any distribution $F(x)$. A direct (but sometimes time demanding) algorithm to produce random numbers x_i from *any* distribution $F(x)$ is given by:

$$x_i = F^{-1}(u_i) \tag{2.54}$$

where u_i is the sequence of random numbers with uniform distribution in [0,1]. This is very easy to apply in any computational environment.† However, for the most common distributions, there are much faster algorithms than this, which the interested reader can find in relevant probability books (e.g. Papoulis, 1990).

2.8 Expectation

Expectation is a key concept of stochastics, enabling a macroscopic view of a phenomenon or process in which the details are intentionally neglected. It converts a stochastic variable into a common one.

For a discrete stochastic variable $\underline{x}$, taking on the values $x_1, x_2, \ldots, x_J$ (where J could be ∞) with probability mass function $P_j \equiv P(x_j) = P\{x = x_j\}$, if $g(\underline{x})$ is an arbitrary function of $\underline{x}$ (so that $g(\underline{x})$ is a stochastic variable per se), we define the *expectation* or *expected value* or *mean* of $g(\underline{x})$ as:

$$\mathrm{E}[g(\underline{x})] \coloneqq \sum_{j=1}^{J} g(x_j)P(x_j) \tag{2.55}$$

Likewise, for a continuous stochastic variable $\underline{x}$ with density $f(x)$, the expectation is defined as:

$$\mathrm{E}[g(\underline{x})] \coloneqq \int_{-\infty}^{\infty} g(x)f(x)\mathrm{d}x \tag{2.56}$$

Expected values are common variables: for example, $\mathrm{E}[\underline{x}]$ and $\mathrm{E}[g(\underline{x})]$ are constants—neither functions of x nor of $\underline{x}$. That justifies the notation $\mathrm{E}[\underline{x}]$ instead of $\mathrm{E}(\underline{x})$ or $\mathrm{E}(x)$, which would imply functions of $\underline{x}$ or x.

---

* For example, for Excel (which by default includes the function rand) the Mersenne twister algorithm, called NtRand, can be found in www.ntrand.com/download/.

† For example, in Excel the function normsinv(rand()) generates random numbers from the normal distribution.

2.9 Classical moments and cumulants

For certain types of functions $g(\underline{x})$ we get very commonly used statistical parameters, as specified below:

- The *noncentral moment* of order q (or the q*th moment about the origin*):

$$g(\underline{x}) = \underline{x}^q, \qquad \mu'_q \coloneqq \mathrm{E}[\underline{x}^q] \tag{2.57}$$

- The *mean* (or the first moment):

$$g(\underline{x}) = \underline{x}, \qquad \mu \coloneqq \mu'_1 = \mathrm{E}[\underline{x}] \tag{2.58}$$

- The *central moment of order q*:

$$g(\underline{x}) = (\underline{x} - \mu)^q, \qquad \mu_q \coloneqq \mathrm{E}[(\underline{x} - \mu)^q] \tag{2.59}$$

  For $q = 0$ and 1 the central moments are respectively 1 and 0.

- The *variance*:

$$g(\underline{x}) = (\underline{x} - \mu)^2, \qquad \gamma \coloneqq \mathrm{E}[(\underline{x} - \mu)^2] \coloneqq \sigma^2 \tag{2.60}$$

  The variance is also denoted as $\mathrm{var}[\underline{x}]$; its square root σ (also denoted as $\mathrm{std}[\underline{x}]$) is called the standard deviation.

To distinguish the above quantities from other types of moments to be introduced below, we call them *classical moments*. Most used amongst the moments of order higher than two, are the third and fourth. If we standardize them by appropriate powers of σ to make them dimensionless, we get, respectively, the *coefficients of skewness* and *kurtosis*:

$$C_\mathrm{s} \coloneqq \frac{\mu_3}{\sigma^3}, \qquad C_\mathrm{k} \coloneqq \frac{\mu_4}{\sigma^4} \tag{2.61}$$

Other dimensionless indices are the ratios:

$$\frac{\mu}{\sigma}, \qquad \frac{\sigma}{\mu} =: C_\mathrm{v} \tag{2.62}$$

where the latter is called *coefficient of variation* (meaningful for nonnegative variables).

Central and noncentral moments are related to each other by:

$$\mu'_q = \sum_{i=0}^{q} \binom{q}{i} \mu^{q-i} \mu_i, \qquad \mu_q = \sum_{i=0}^{q} \binom{q}{i} (-\mu)^{q-i} \mu'_i \tag{2.63}$$

where $\mu_0 = \mu'_0 = 1, \mu_1 = 0, \mu'_1 = \mu$. (For proof of these relationships see Koutsoyiannis, 2025). For small q they take the following forms:

$$\mu'_2 = \sigma^2 + \mu^2, \qquad \mu'_3 = \mu_3 + 3\sigma^2\mu + \mu^3, \qquad \mu'_4 = \mu_4 + 4\mu_3\mu + 6\sigma^2\mu^2 + \mu^4 \tag{2.64}$$

and can be inverted as follows:

$$\sigma^2 = \mu'_2 - \mu^2, \qquad \mu_3 = \mu'_3 - 3\mu'_2\mu + 2\mu^3, \qquad \mu_4 = \mu'_4 - 4\mu'_3\mu + 6\mu'_2\mu^2 - 3\mu^4 \tag{2.65}$$

For ready reference, Table 2.3 provides the analytical expressions of the moments of some common distribution functions.

Table 2.3 Some common distributions of continuous variables and their moments (and cumulants when their expressions are simple).

Name, parameters, domain	Probability density or distribution function	Moments and cumulants
Uniform in $[a, b]$, $a \le x \le b$	$f(x) = \dfrac{1}{b-a}$	$\mu'_1 = \dfrac{a+b}{2}, \quad \mu_2 = \dfrac{(b-a)^2}{12}, \quad \mu'_q = \dfrac{b^{q+1}-a^{q+1}}{(q+1)(b-a)}$
Beta, $0 \le \underline{x} \le b$ $\zeta > 0, \varsigma > 0, \lambda > 0$	$f(x) = \dfrac{\left(\frac{x}{b}\right)^{\zeta-1}\left(1-\frac{x}{b}\right)^{\varsigma-1}}{\mathrm{B}(\zeta,\varsigma)}$	$\mu'_q = \dfrac{\Gamma(\zeta+\varsigma)\Gamma(q+\zeta)}{\Gamma(\zeta)\,\Gamma(q+\zeta+\varsigma)} b^q = \dfrac{\mathrm{B}(q+\zeta,\varsigma)}{\mathrm{B}(\zeta,\varsigma)}$
Exponential $\mu > 0, x \ge 0$	$f(x) = \dfrac{\mathrm{e}^{-\frac{x}{\mu}}}{\mu}$	$\mu'_1 = \mu, \quad \mu_2 = \mu^2, \quad \mu'_q = q!\,\mu^q,$ $\kappa_q = (q-1)!\,\mu^q$
Gamma $\zeta > 0, \lambda > 0, x \ge 0$	$f(x) = \dfrac{(x/\lambda)^{\zeta-1}\mathrm{e}^{-\frac{x}{\lambda}}}{\lambda\,\Gamma(\zeta)}$	$\mu'_1 = \zeta\lambda, \quad \mu_2 = \zeta\lambda^2, \quad \mu'_q = \dfrac{\Gamma(q+\zeta)}{\Gamma(\zeta)}\lambda^q,$ $\kappa_q = \zeta(q-1)!\,\lambda^q$
Weibull $\zeta > 0, \lambda > 0, x \ge 0$	$F(x) = 1 - \exp\left(-\left(\dfrac{x}{\lambda}\right)^{\zeta}\right)$	$\mu'_1 = \Gamma\left(1+\dfrac{1}{\zeta}\right)\lambda, \quad \mu_2 = \left(\Gamma\left(1+\dfrac{2}{\zeta}\right) - \Gamma\left(1+\dfrac{1}{\zeta}\right)^2\right)\lambda^2$ $\mu'_q = \Gamma\left(1+\dfrac{q}{\zeta}\right)\lambda^q$
Normal $\mu \in \mathbb{R}, \sigma > 0,$ $x \in \mathbb{R}$	$f(x) = \dfrac{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma}$	$\mu'_1 = \mu, \quad \mu_2 = \sigma^2, \quad \mu_q = \begin{cases} 0, & q \text{ odd} \\ \sigma^q(q-1)!!, & q \text{ even} \end{cases}$ $\kappa_q = \begin{cases} \mu'_1 = \mu, & q = 1 \\ \sigma^2 & q = 2 \\ 0 & q > 2 \end{cases}$
Lognormal ($\ln \underline{x}$ is $\mathrm{N}(\ln\lambda, \varsigma)$) $\varsigma > 0, \lambda > 0, x \ge 0$	$f(x) = \dfrac{\exp\left(-\frac{1}{2\varsigma^2}\left(\ln\left(\frac{x}{\lambda}\right)\right)^2\right)}{\sqrt{2\pi}\,\varsigma x}$	$\mu'_1 = \mathrm{e}^{\frac{\varsigma^2}{2}}\lambda, \quad \mu_2 = \mathrm{e}^{\varsigma^2}\left(\mathrm{e}^{\varsigma^2}-1\right)\lambda^2, \quad \mu'_q = \mathrm{e}^{\frac{q^2\varsigma^2}{2}}\lambda^q$
Pareto[1] $\xi > 0, \lambda > 0, x \ge 0$	$F(x) = 1 - \left(1+\xi\dfrac{x}{\lambda}\right)^{-\frac{1}{\xi}}$	$\mu'_1 = \dfrac{\lambda}{1-\xi}, \quad \mu_2 = \dfrac{\lambda^2}{(1-\xi)^2(1-2\xi)}$ $\mu'_q = \mathrm{B}\left(\dfrac{1}{\xi}-q, q+1\right)\dfrac{\lambda^q}{\xi^{q+1}}$
Pareto-Burr-Feller[1] (PBF)[2] $\zeta > 0, \xi > 0,$ $\lambda > 0, x \ge 0$	$F(x) = 1 - \left(1+\zeta\xi\left(\dfrac{x}{\lambda}\right)^{\zeta}\right)^{-\frac{1}{\xi\zeta}}$	$\mu'_q = \mathrm{B}\left(\dfrac{1}{\xi\zeta}-\dfrac{q}{\zeta}, \dfrac{q}{\zeta}+1\right)\dfrac{\lambda^q}{(\xi\zeta)^{\frac{q}{\zeta}+1}}$
Dagum[1] $\zeta > 0, \xi > 0,$ $\lambda > 0, x \ge 0$	$F(x) = \left(1+\dfrac{1}{\zeta\xi}\left(\dfrac{x}{\lambda}\right)^{-\frac{1}{\xi}}\right)^{-\zeta\xi}$	$\mu'_q = (\zeta\xi)^{1-\xi q}\,\mathrm{B}\big(1-\xi q, \xi(q+\zeta)\big)\lambda^q$
Extreme value type I (EV1) $\lambda > 0, x \in \mathbb{R}$	$F(x) = \exp\left(-\mathrm{e}^{-\frac{x}{\lambda}}\right)$	$\mu'_1 = \gamma\lambda, \quad \mu_2 = \dfrac{\pi^2\lambda^2}{6}, \quad \kappa_q = (-1)^q\psi^{(q-1)}(1)\lambda^q$
Extreme value type II (EV2)[1] $\xi > 0, \lambda > 0, x \ge 0$	$F(x) = \exp\left(-\left(\dfrac{x}{\lambda}\right)^{-\frac{1}{\xi}}\right)$	$\mu'_1 = \Gamma(1-\xi)\lambda, \mu_2 = (\Gamma(1-2\xi) - \Gamma(1-\xi)^2)\lambda^2$ $\mu'_q = \Gamma(1-q\xi)\lambda^q$

[1] The moments exist (have finite values) only for order $q < 1/\xi$; for larger q they are infinite.

[2] Also known as *Pareto III and IV*, *Burr XII* and *Feller*; for justification of the name PBF see Koutsoyiannis et al. (2018).

Another useful expectation is formed by choosing $g(\underline{x}) = \mathrm{e}^{t\underline{x}}$ for any t. The logarithm of the resulting expectation is called the *cumulant generating function*:

$$K(t) \coloneqq \ln \mathrm{E}[\mathrm{e}^{t\underline{x}}] \tag{2.66}$$

The power series expansion of the cumulant generating function i.e.:

$$K(t) = \sum_{q=1}^{\infty} \kappa_q \frac{t^q}{q!} \tag{2.67}$$

defines the *cumulants* κ_q. These are related to noncentral moments of similar order by (Smith, 1995):

$$\mu'_q = \sum_{i=0}^{q-1} \binom{q-1}{i} \kappa_{q-i} \mu'_i, \qquad \kappa_q = \mu'_q - \sum_{i=1}^{q-1} \binom{q-1}{i} \kappa_{q-i} \mu'_i \tag{2.68}$$

For small q they take the following forms:

$$\kappa_0 = \mu_1 = 0, \qquad \kappa_1 = \mu'_1 = \mu, \qquad \kappa_2 = \mu_2, \qquad \kappa_3 = \mu_3, \qquad \kappa_4 = \mu_4 - 3\mu_2^2 \tag{2.69}$$

The importance of cumulants lies in their homogeneity and additivity properties. Namely, for a stochastic variable that is the weighted sum of r independent variables $\underline{v}_i$, i.e., $\underline{x} = a_1\underline{v}_1 + \cdots + a_r\underline{v}_r$, the qth cumulant of $\underline{x}$ is given as

$$\kappa_q = a_1^q \, \kappa_q^{(v_1)} + \cdots + a_r^q \, \kappa_q^{(v_r)} \tag{2.70}$$

where $\kappa_q^{(v_l)}$ is qth cumulant of $\underline{v}_l$. This property is quite useful in stochastic simulation (see Koutsoyiannis and Dimitriadis, 2021; Koutsoyiannis, 2025).

Digression 2.K: Illustration of the first four classical moments and related statistical characteristics

The geometrical meaning of the first four classical moments is visualized in Figure 2.3. Essentially, the first moment, i.e. the mean, describes the abscissa of the centre of gravity of the shape defined by the probability density function and the horizontal axis (Figure 2.3a). It is also equivalent to the static moment of this shape about the vertical axis (given that the area of the shape equals 1). Often, the following quantities are alternatively used as location parameters:

- The *mode*, or most probable value, x_{m}, is the value of x for which the density $f(x)$ becomes maximum, if the stochastic variable is continuous, or, for discrete variables, the probability mass becomes maximum. If $f(x)$ has one, two or many local maxima, we say that the distribution is unimodal, bimodal or multi-modal, respectively.
- The *median*, $x_{0.5}$, is the value for which $P\{\underline{x} \le x_{0.5}\} \ge 1/2$ and $P\{\underline{x} \ge x_{0.5}\} \ge 1/2$. Thus, for a continuous stochastic variable, a vertical line at the median separates the graph of the density function into two equivalent parts each having an area of $1/2$.

Generally, the mean, the mode and the median are not identical unless the density has a symmetrical and unimodal shape.

The variance of a stochastic variable and its square root, the standard deviation, which has the same dimensions as the stochastic variable, quantify the scatter or dispersion of the probability density around the mean. Thus, a small variance shows a concentrated distribution (Figure 2.3b). The variance cannot be negative; its lowest possible value is zero. This corresponds to a variable that takes one value only (the mean) with absolute certainty. Geometrically, the

variance is equivalent to the moment of inertia about the vertical axis that passes from the centre of gravity of the shape defined by the probability density function and the horizontal axis.

Alternative measures of dispersion are provided by the so-called interquartile range, defined as the difference $x_{0.75} - x_{0.25}$, i.e., the difference between the 0.75 and 0.25 quantiles (or upper and lower quartiles) of the stochastic variable (which define an area in the density function equal to 0.5).

The third central moment is used as an indicator of skewness. A zero value indicates that the density is symmetric. This can be easily verified from the definition of the third central moment. If the third central moment is positive or negative, we say that the distribution is positively or negatively skewed respectively (Figure 2.3c). In a positively skewed unimodal distribution, $x_{\mathrm{m}} \leq x_{0.5} \leq \mu$; the reverse inequality holds for a negatively skewed distribution.

The fourth central moment is used as an indicator of kurtosis, a term which describes the "peakedness" of the probability density function around its mode. A reference value for kurtosis is provided by the normal distribution, which has $C_{\mathrm{k}} = 3$. Distributions with kurtosis greater than the reference value are called *leptokurtic* (acute, sharp) and typically have a heavy upper tail (see below), so that most of the variance is due to infrequent extreme deviations as opposed to frequent modestly-sized deviations. Distributions with kurtosis less than the reference value are called *platykurtic* (flat; Figure 2.3d).

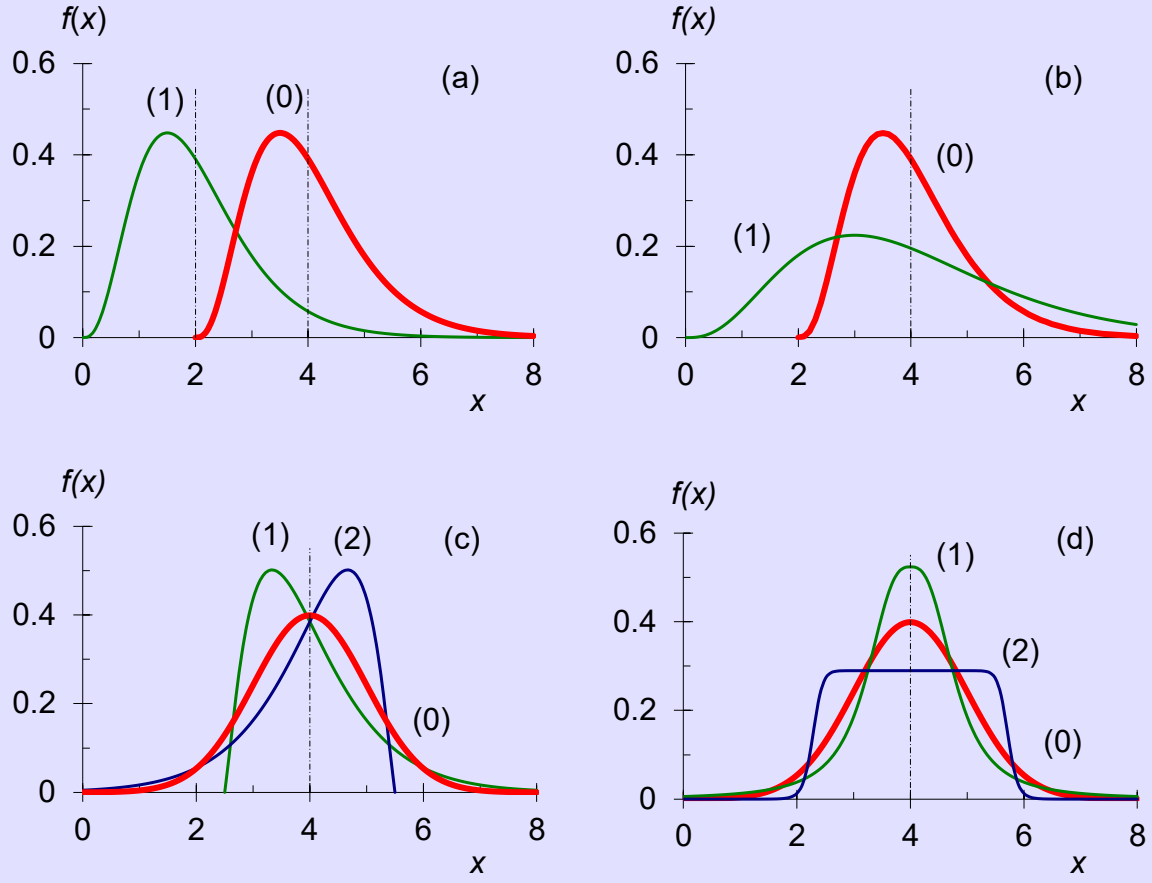

Figure 2.3 Graphical illustration of the geometrical interpretation of moments of a stochastic variable: **(a)** *Effect of the mean.* Curves (0) and (1) have means 4 and 2, respectively, whereas they both have standard deviation 1, coefficient of skewness 1 and coefficient of kurtosis 4.5. **(b)** *Effect of the standard deviation.* Curves (0) and (1) have standard deviation 1 and 2 respectively, whereas they both have mean 4, coefficient of skewness 1 and coefficient of kurtosis 4.5. **(c)** *Effect of the coefficient of skewness.* Curves (0), (1) and (2) have coefficients of skewness 0, +1.33 and −1.33, respectively, but they all have mean 4 and standard deviation 1; their coefficients of kurtosis are 3, 5.67 and 5.67, respectively. **(d)** *Effect of the coefficient of kurtosis.* Curves (0), (1) and (2) have coefficients of kurtosis 3, 5 and 2, respectively, whereas they all have mean 4, standard deviation 1 and coefficient of skewness 0.

2.10 Entropy as expectation

The enumeration of the ground set and hence the definition of a stochastic variable entails arbitrary choices and one could think of different options. In turn, expectations and moments depend on the option chosen. One may think of defining the function $g(\,)$ whose expectation is sought, in terms of the probability per se, i.e. $g(\underline{x}) = h(P(\underline{x}))$ for a discrete variable or $g(\underline{x}) = h\left(f(\underline{x})\right)$ for a continuous variable, where $h(\,)$ is any specified function. Among the several choices of $h(\,)$, most useful is the logarithmic function, which may render entropy as an expectation. Namely, the entropy of a stochastic variable is defined to be the entropy of the finest partition $\mathbb{V}$.

Specifically, for a discrete stochastic variable $\underline{x}$ taking integer values $x_j, = 1, \dots, J$ (where J could be infinite) with $P_j = P\{\underline{x} = x_j\}$ the entropy is:

$$\Phi[\underline{x}] \coloneqq \mathrm{E}[-\ln P(\underline{x})] = -\sum_{j=1}^{J} P_j \ln P_j \tag{2.71}$$

For a continuous stochastic variable $\underline{x}$ taking real values x in $(-\infty, \infty)$ the definition of the entropy of $\underline{x}$ is possible by means of the relative entropy. First, we take the finest partition for intervals of size δx, form the relative entropy for measure $B(x) = \int \beta(x)\,\mathrm{d}x$, and make the partition with elements $A_i = [(i-1)\delta x, i\delta x]$,

$$\begin{aligned} \Phi(\mathbb{A}||\mathrm{B})_{\delta x} &= -\sum_{i=-\infty}^{\infty} P(A_i) \ln\left(\frac{P(A_i)}{B(A_i)}\right) \\ &= -\sum_{i=-\infty}^{\infty} \Big(F(i\delta x) - F\big((i-1)\delta x\big)\Big) \ln\left(\frac{\Big(F(i\delta x) - F\big((i-1)\delta x\big)\Big)}{\Big(B(i\delta x) - B\big((i-1)\delta x\big)\Big)}\right) \end{aligned} \tag{2.72}$$

where $F(\,)$ is the distribution function. Then we take the limit of $\Phi(\mathbb{A}||\mathrm{B})_{\delta x}$ as $\delta x \to 0$ and set $F(i\delta x) - F\big((i-1)\delta x\big) = f(x)\delta x, B(i\delta x) - B\big((i-1)\delta x\big) = \beta(x)\delta x$:

$$\lim_{\delta x \to 0} \Phi(\mathbb{A}||\mathrm{B})_{\delta x} =: \Phi[\underline{x}] \coloneqq \mathrm{E}\left[-\ln\frac{f(\underline{x})}{\beta(\underline{x})}\right] = -\int_{-\infty}^{\infty} \ln\frac{f(x)}{\beta(x)} f(x)\mathrm{d}x \tag{2.73}$$

It is reminded that the background measure density $\beta(x)$ can be any density of a measure, such as a proper probability density (with integral equal to 1) or an improper one (meaning that its integral diverges). Most commonly, it is assumed to be an (improper) Lebesgue density, i.e. a constant with dimensions $[\beta(x)] = [f(x)] = [x^{-1}]$, so that the argument of the logarithm function be dimensionless. It can easily be shown that for $\beta(x) = \beta =$ constant, equation (2.73) can be expressed in a simpler manner in terms of the derivative of the quantile function $x(F)$ as:

$$\Phi[\underline{x}] \coloneqq \int_0^1 \ln\big(\beta\, x'(F)\big)\,\mathrm{d}F \tag{2.74}$$

This is useful for the numerical evaluation of $\Phi[\underline{x}]$, particularly when the quantile function is estimated empirically, provided that the estimated $x(F)$ is smooth, so that its derivative can be reliably estimated.

Entropy maximization is typically performed for the finest partition and for constraints suitable for each case. The standard constraints for all cases are:

- For discrete stochastic variables $\underline{x}$:

$$\sum_{j=1}^{J} P_j = 1 \tag{2.75}$$

- For continuous stochastic variables $\underline{x}$:

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1 \tag{2.76}$$

Digression 2.L: Illustration of the principle of maximum entropy

Here we illustrate the maximum entropy (ME) principle in a few simple cases. The examples may look trivial. However, we must have in mind that, as already mentioned in Digression 2.F, with the same reasoning we can infer more interesting things, such as the laws of thermodynamics (Chapter 6). The logic is the same: we maximize the uncertainty with respect to the state of a die or a water molecule.

(a) We thus start from the simple example of determining the probabilities of the outcomes of a die throw. These probabilities can readily be inferred by Theorem 4 in section 2.3, but here we will derive them from scratch for illustration. For the die the entropy is:

$$\Phi = \mathrm{E}\left[-\ln P\left(\underline{x}\right)\right] = -P_1 \ln P_1 - P_2 \ln P_2 - P_3 \ln P_3 - P_4 \ln P_4 - P_5 \ln P_5 - P_6 \ln P_6$$

Considering also the equality constraint:

$$P_1 + P_2 + P_3 + P_4 + P_5 + P_6 = 1$$

we form the objective function to maximize as:

$$\begin{aligned} L \coloneqq & -P_1 \ln P_1 - P_2 \ln P_2 - P_3 \ln P_3 - P_4 \ln P_4 - P_5 \ln P_5 - P_6 \ln P_6 \\ & + a(P_1 + P_2 + P_3 + P_4 + P_5 + P_6 - 1) \end{aligned}$$

where a is a Lagrange multiplier. We find the partial derivatives with respect to each of the variables and equate them to zero, obtaining:

$$\frac{\partial L}{\partial P_1} = -1 - \ln P_1 + a = 0 \ , \qquad \ldots \ , \qquad \frac{\partial L}{\partial P_6} = -1 - \ln P_6 + a = 0$$

Obviously, the solution of these equations yields the single maximum:

$$P_1 = P_2 = P_3 = P_4 = P_5 = P_6 = 1/6$$

The entropy is $\Phi = -6\ (1/6) \ln (1/6) = \ln 6$. In general, the entropy for J equiprobable outcomes is:

$$\Phi = \ln J \tag{2.77}$$

It is noted that entropy and information are complementary to each other. When we know (observe) that the outcome is i ($P_i = 1, P_j = 0$ for $j \neq i$), the entropy is zero.

In the above case of a fair die throw, the application of the ME principle is equivalent to the *principle of insufficient reason*. However, while the former is a variational law (equivalent to the solution of an optimization problem), the latter is formulated in terms of equations. As already stated in section 1.2, a single variational law is much more powerful than very many equations. Actually, from a variational law we derive as many equations as there are unknowns (even an infinite number of equations). And as we showed, in this case the variational ME principle entails the principle of insufficient reason, and thus there is no need at all to postulate the latter as an additional philosophical or scientific principle.

(b) To illustrate that the variational ME principle is more powerful than the principle of insufficient reason, we consider the following variant of the problem in which uniformity is a priori excluded. Specifically, we assume that the die is loaded and that we have prior information that $P_6 = 2P_1$. What is the probability that the outcome of a die throw will be i in this case?

For the ME optimization we only need to take into account the additional constraint, by adding to the objective function the term $b(P_6 - 2P_1)$ where b is an additional Lagrange multiplier. The solution of the optimization problem is a single maximum, $P_2 = P_3 = P_4 = P_5 = 0.1698$ (slightly >1/6), $P_1 = 0.1069, P_6 = 0.2139$. The entropy is $\Phi = 1.7732$, smaller than in the case of equiprobability, in which $\Phi = \ln 6 = 1.792$. The decrease of entropy in the loaded die derives from the additional information incorporated in the constraints.

(c) In another example we consider a roulette wheel which is not divided into pockets, but its outcome is a real number measured on a circular scale graded 0 to J. In this case our stochastic variable $\underline{x}$ is of continuous type. Assuming background density $\beta(\underline{x}) = 1$, the entropy is

$$\Phi[\underline{x}] = -\int_0^J \ln f(x) f(x) \mathrm{d}x$$

Considering also the constraint (2.76) with a Lagrange multiplier a, we should maximize:

$$L := -\int_0^J \ln f(x) f(x) \mathrm{d}x - a\left(\int_0^J f(x) \mathrm{d}x - 1\right)$$

Finding the partial derivative with respect to f and equating it to zero we obtain:

$$\frac{\partial L}{\partial f} = -1 - \ln f - a = 0$$

Hence $f = \exp(-1 - a) = \text{constant}$ and from the constraint we find that the entropy maximizing density is:

$$f(x) = \frac{1}{J} \tag{2.78}$$

and the entropy is:

$$\Phi = \ln J \tag{2.79}$$

This is the uniform distribution, given in Table 2.2. Notice that the expression of maximum entropy for a discrete stochastic variable (equation (2.77)) is identical to that of a continuous stochastic variable (equation (2.79)).

(d) If in the uniform distribution the upper bound J tends to ∞ (while the lower bound remains 0), it becomes improper ($f(x) = 0$). Therefore, in this case we need an additional constraint to find a proper distribution. The simplest one that we can think of is that the distribution has a specified mean μ, i.e.:

$$\int_0^\infty xf(x)\mathrm{d}x = \mu$$

The expression of the entropy is the same as in the example (c), but the objective function to maximize becomes:

$$A := -\int_0^\infty \ln f(x) f(x)\mathrm{d}x - a\left(\int_0^\infty f(x)\mathrm{d}x - 1\right) - b\left(\int_0^\infty xf(x)\mathrm{d}x - \mu\right)$$

Thus,

$$\frac{\partial A}{\partial f} = -1 - \ln f - a - bx = 0$$

and

$$f(x) = B\exp(-bx)$$

where from the two constraints we find, after the algebraic operations, that $B = b = 1/\mu$. This is the exponential distribution given in Table 2.2. It is very common in physics, as the mean constraint, from which it results, is omnipresent. For example, if $\underline{x}$ represents the kinetic energy of one of many particles moving in a box, we do not know the exact energy of each particle (which may change due to collisions, assumed to be elastic) but we may know the average μ, which is preserved according to the related physical principle (energy conservation). Consequently, the distribution of kinetic energy is exponential.

(e) If in the above example of moving particles we limit the motion to a straight line, and we choose not the kinetic energy but the velocity as our stochastic variable $\underline{x}$, which can be either positive or negative, the kinetic energy constraint is written as

$$\int_0^\infty x^2 f(x)\mathrm{d}x = \gamma$$

where γ is twice the average kinetic energy per unit mass. The objective function to maximize becomes:

$$A := -\int_0^\infty \ln f(x)\, f(x)\, \mathrm{d}x - a\left(\int_0^\infty f(x)\mathrm{d}x - 1\right) - b\left(\int_0^\infty x^2 f(x)\mathrm{d}x - \gamma\right)$$

Thus,

$$\frac{\partial A}{\partial f} = -1 - \ln f - a - bx^2 = 0$$

and

$$f(x) = B\exp(-bx^2)$$

where from the two constraints we find, after the algebraic operations, that $B = \sqrt{2\pi\gamma}$, $b = 1/2\gamma$. This is the normal distribution given in Table 2.2, with $\mu = 0$ and standard deviation $\sigma = \sqrt{\gamma}$.

In fact, all distributions of Table 2.2, except the Poisson distribution, turn out to be entropy maximizing distributions, either without a constraint or under a simple constraint in each case. The results are summarized in Table 2.4.

Table 2.4 Entropy of the most common distributions of Table 2.2, which turn out to be entropy maximizing distributions for Lebesgue background density ($\beta(x) = 1$) with simple constraints.

Name (and parameters)	Probability mass, density and distribution function	Corresponding entropy for unit background measure density
Discrete variable $\underline{x}$ with values x_j		
Discrete uniform, $x_j = 1, \dots, J$	$P(x_j) = 1/J$, $F(x) = \max(0, \min(\lfloor x \rfloor / J, 1))$	$\Phi[\underline{x}] = \ln J$ (the maximum among all distributions with $x_j = 1, \dots, J$)
Geometric $x_j = 0,1, \dots$ ($\mu > 0$)	$P(x_j) = \frac{1}{1+\mu}\left(\frac{\mu}{1+\mu}\right)^j$ $F(x) = \max\left(0, 1 - \frac{1}{1+\mu}\left(\frac{\mu}{1+\mu}\right)^{\lfloor x \rfloor}\right)$	$\Phi[\underline{x}] = \ln\left(\frac{(\mu+1)^{\mu+1}}{\mu^\mu}\right) \approx 1 + \ln(\mu + 1/\mathrm{e})$ (the maximum among all distributions with $x_j = 0,1, \dots$, and mean μ)
Continuous variable $\underline{x}$		
Uniform in $[0, J]$	$f(x) = \begin{cases} \frac{1}{J} & \text{for } 0 \le x \le J \\ 0 & \text{otherwise} \end{cases}$ $F(x) = \max(0, \min(x/J, 1))$	$\Phi[\underline{x}] = \ln J$ (the maximum among all distributions with domain $[0, \alpha]$)
Exponential ($\mu > 0$)	$f(x) = \begin{cases} \frac{\mathrm{e}^{-\frac{x}{\mu}}}{\mu} & x \ge 0 \\ 0 & x < 0 \end{cases}$ $F(x) = \begin{cases} 1 - \mathrm{e}^{-\frac{x}{\mu}} & \text{for } x \ge 0 \\ 0 & \text{for } x < 0 \end{cases}$	$\Phi[\underline{x}] = 1 + \ln \mu$ (the maximum among all distributions with domain $[0, \infty)$ and mean μ)
Normal ($\mu \in \mathbb{R}$, $\sigma > 0$)	$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ $F(x) = \frac{1}{2}\mathrm{erfc}\left(-\frac{x-\mu}{\sqrt{2}\sigma}\right)$	$\Phi[\underline{x}] = \frac{1}{2}(1 + \ln(2\pi)) + \ln \sigma = 1.419 + \ln \sigma$ (the maximum among all distributions with domain $(-\infty, \infty)$, mean μ and standard deviation σ)

Note: $\lfloor x \rfloor$ denotes the floor of the number x.

2.11 Maximum entropy distributions

In Digression 2.L we illustrated several simple cases of entropy maximization, in which we determined the entire probability mass or density function based on one or two constraints. We can generalize the result for a number of constraints of the form:

$$\mathrm{E}[g_i(\underline{x})] = \gamma_i \Leftrightarrow \int_{-\infty}^{\infty} g_i(x) f(x) \mathrm{d}x - \gamma_i = 0 \tag{2.80}$$

and for any background measure density $\beta(x)$. In this case, after incorporating the constraints to the entropy with Lagrange multipliers, the expression whose maximization is sought is:

$$L \coloneqq -\int_{-\infty}^{\infty} \ln\frac{f(x)}{\beta(x)} f(x)\mathrm{d}x - a\left(\int_{-\infty}^{\infty} f(x)\mathrm{d}x - 1\right) - \sum_i b_i \left(\int_{-\infty}^{\infty} g_i(x) f(x)\mathrm{d}x - \gamma_i\right) \quad (2.81)$$

Taking the partial derivative with respect to f and equating it to zero we find

$$-\ln\frac{f(x)}{\beta(x)} + 1 - a - \sum_i b_i g_i(x) = 0 \quad (2.82)$$

and, thus, the entropy maximizing density is:

$$f(x) = A\,\beta(x)\exp\left(-\sum_i b_i g_i(x)\right) \quad (2.83)$$

where $A \coloneqq \mathrm{e}^{1-a}$ is a constant.

As we have seen in Digression 2.L, some of the most typical distributions which are used in a variety of scientific fields can be derived from entropy maximization using a simple constraint. Here we will try to get a plethora of distributions again using a single constraint but both with a Lebesgue background measure and a generalized one.

The background measure density $\beta(x)$ shows how we measure the distances d between values of x. The Lebesgue measure corresponds to the Euclidean distance, $d(x, x') = |x - x'|$. However, most geophysical variables are non-negative physical quantities unbounded from above (e.g., precipitation, streamflow, temperature—absolute, expressed in kelvins). With positive physical quantities, often the Euclidean distance is not an ideal metric and so sometimes we use a logarithmic distance $d(x, x') = |\ln(x'/x)|$, as shown in the example below referring to precipitation depth:

	Euclidean distance	Logarithmic distance
$x = 0.1$ mm, $x' = 0.2$ mm	0.1 mm	$\ln 2 = 0.693$
$x = 100$ mm, $x' = 100.1$ mm	0.1 mm	$\ln 1.001 = 0.001$
$x = 100$ mm, $x' = 200$ mm	100 mm	$\ln 2 = 0.693$

Which of the second and third pairs of points is equidistant from the first one? In an attempt to merge (or unify) the Euclidean and logarithmic distance, we heuristically introduce (see Koutsoyiannis, 2014a) the following background measure density for nonnegative variables:

$$\beta(x) = \frac{1}{\lambda + x} \quad (2.84)$$

where λ is a characteristic scale parameter, which also serves as a physical unit for x. According to this, the distance of any point x from 0 is:

$$B_{\mathrm{LS}}(x) = \int_0^x \lambda\,\beta(s)\mathrm{d}s = \lambda \ln\left(1 + \frac{x}{\lambda}\right) \quad (2.85)$$

We will refer to $B_{\mathrm{LS}}(x)$ as the *log-shift measure*. An example plot for $B_{\mathrm{LS}}(x)$ is given in Figure 2.4. Its limiting properties are:

$$\lim_{x\to 0}\frac{B_{\mathrm{LS}}(x)}{x}=\lim_{\lambda\to\infty}\frac{B_{\mathrm{LS}}(x)}{x}=1,\qquad \lim_{x\to\infty}\left(\frac{B_{\mathrm{LS}}(x)}{\lambda\,\ln(x/\lambda)}\right)=\lim_{\lambda\to 0}\left(\frac{B_{\mathrm{LS}}(x)}{\lambda\,\ln(x/\lambda)}\right)=1 \qquad (2.86)$$

This means that for large λ or small x $B_{\mathrm{LS}}(x)$ behaves like the Lebesgue measure and for small λ or large x it behaves like a logarithmic transformation.

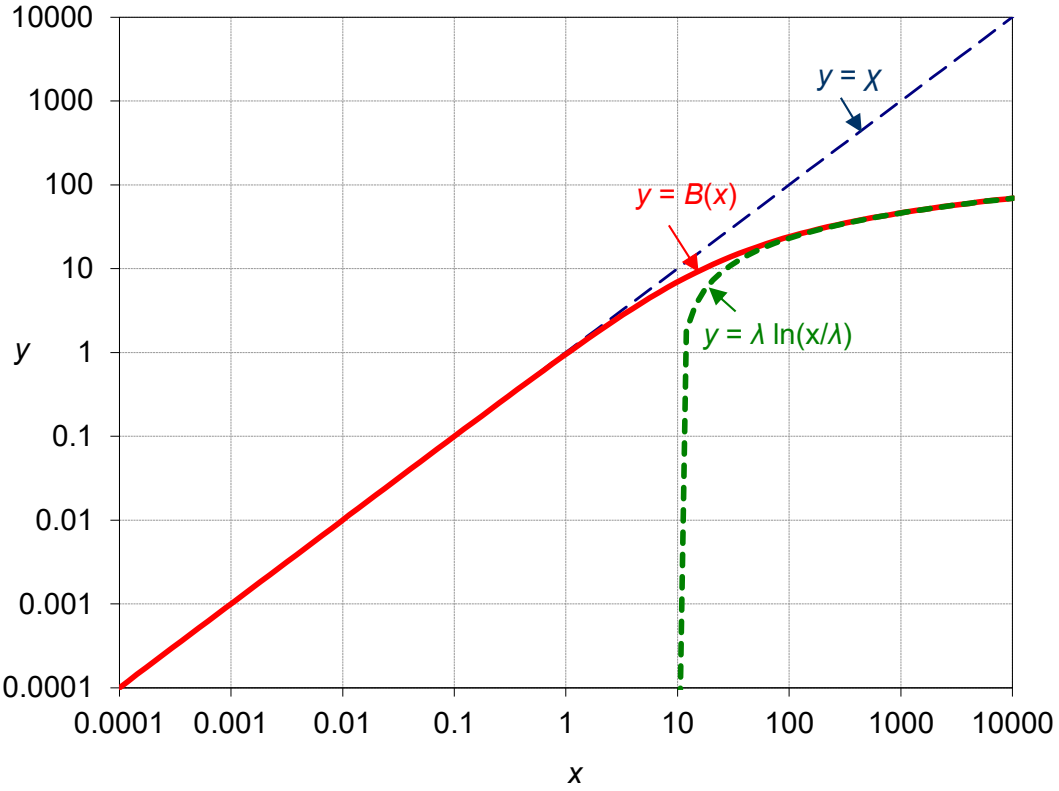

Figure 2.4 Illustration of the log-shift measure $y=B(x)=B_{\mathrm{LS}}(x)$; the example plot is for $\lambda=10$ and shows that for small x ($<\lambda/10$) $B(x)$is indistinguishable from x, while for large x ($>10\lambda$), $B(x)$ becomes a linear function of $\ln x$.

The distance between any two points x and x' is:

$$d(x,x')=|B(x')-B(x)|=\lambda\left|\ln\left(\frac{1+x'/\lambda}{1+x/\lambda}\right)\right| \qquad (2.87)$$

For small x values, i.e., $x<x'\ll\lambda$, the distance is $d(x,x')=\lambda\ln(1+(x'-x)/(\lambda+x))\approx x'-x$ (Euclidean distance). For large values, $\lambda\ll x<x', d(x,x')\approx\lambda\ln(x'/x)$ (logarithmic distance). We notice that both $B(x)$ and $d(x,x')$ have the same units as x (physical consistency).

We further generalize this measure to the following form:

$$B_{\mathrm{LPS}}(x)=\lambda\ln\left(1+\left(\frac{x}{\lambda}\right)^{c}\right),\qquad B'_{\mathrm{LPS}}(x)=\lambda\,\beta(x)=c\left(\frac{x}{\lambda}\right)^{c-1}/\left(1+\left(\frac{x}{\lambda}\right)^{c}\right) \qquad (2.88)$$

and we call this the *log-power-shift measure*. We note that the dimensions are $[B_{\mathrm{LPS}}(x)]=[x]=[\lambda]$ and so $\beta(x)=B'_{\mathrm{LPS}}(x)/\lambda$ with the derivative $B'_{\mathrm{LPS}}(x)$ being dimensionless.

Now, in the general solution (2.83) we use the background measure $B(x) = B_{\mathrm{LPS}}(x)$ with a single constraint, that is $\mathrm{E}[(B(x)/\lambda)^d] = \gamma$ and from (2.83) we get:

$$f(x) = A\frac{B'(x)}{\lambda}\exp\left(-b\left(\frac{B(x)}{\lambda}\right)^d\right) = \frac{A}{\lambda}\exp\left(-b\left(\frac{B(x)}{\lambda}\right)^d + \ln\big(B'(x)\big)\right) \tag{2.89}$$

After the algebraic operations we find the generalized maximum entropy distribution as:

$$f(x) = \frac{Ac}{\lambda}\left(\frac{x}{\lambda}\right)^{(c-1)}\left(1+\left(\frac{x}{\lambda}\right)^c\right)^{-1}\exp\left(-b\left(\ln\left(1+\left(\frac{x}{\lambda}\right)^c\right)\right)^d\right) \tag{2.90}$$

As a special case, as $\lambda \to \infty$, the log-power-shift background measure approaches the *power measure* $B_{\mathrm{P}}(x)$, and the quantities in (2.88) become:

$$B_{\mathrm{P}}(x) = \lambda\left(\frac{x}{\lambda}\right)^c, \qquad B'_{\mathrm{P}}(x) = c\left(\frac{x}{\lambda}\right)^{c-1} \tag{2.91}$$

This is justified by the fact that $\lim_{\lambda\to\infty} B(x)/B_{\mathrm{P}}(x) = 1$. Hence, the density of (2.89) becomes

$$f(x) = \frac{Ac}{\lambda}\left(\frac{x}{\lambda}\right)^{(c-1)}\exp\left(-b\left(\frac{x}{\lambda}\right)^{cd}\right) \tag{2.92}$$

Furthermore, when $c = 1$ we get the Lebesgue measure and obtain:

$$f(x) = \frac{A'}{\lambda}\exp\left(-b\left(\frac{x}{\lambda}\right)^d\right) \tag{2.93}$$

The densities (2.92) and (2.90) contain as special cases the most common distributions used in stochastics. These special cases are listed in Table 2.5 in terms of their densities $f(x)$ and distribution functions complements $\bar{F}(x) = 1 - F(x)$.

In particular, the density (2.92), which is derived from the power background measure, corresponds to a generalized gamma distribution, also listed in Table 2.5, after suitable transformation of its parameters. The density (2.90), which is derived from the log-shift background measure, does not yield a closed expression for $F(x)$ in its general case, and therefore is not listed in Table 2.5.

The distributions and the special cases resulting from equations (2.92) and (2.90) correspond to nonnegative stochastic variables, $\underline{x} \geq 0$. However, in some of the cases, in which the variable x appears in Table 2.5 raised to power 2, the extension to the whole real line is direct. The distributions of this type are earmarked as "half" in the table, and their "full" versions (valid for all real numbers) are derived by dividing the expressions given in the table by 2; these include the normal distribution.

Some additional distributions, resulting from entropy maximization with two constraints are given in Appendix 2-II.

One may wonder if it would be possible to derive all distributions listed in Table 2.5 using the Lebesgue background measure alone, along with simple constraints based on classical moments and some additional constraints on boundary conditions, such as specifying the tail indices $\xi > 0, \zeta > 1$. This question is examined in Appendix 2-III and the answer turns out to be negative: The density functions in question are not even local (let alone global) maximizers of entropy if that is based on Lebesgue background measure.

However, as seen above, with a different background measure, we can derive all distributions of Table 2.5.

Table 2.5 Maximum entropy distributions derived by equations (2.92), (2.93) and (2.90). (Expressions not to be used at face as different parameterizations are used throughout the book.)

Name	Parameters	$f(x)$	$\bar{F}(x) = 1 - F(x)$
Lebesgue background measure ($c = 1$)			
Exponential	$b = d = 1$	$\frac{1}{\lambda}\exp\left(-\frac{x}{\lambda}\right)$	$\exp\left(-\frac{x}{\lambda}\right)$
Half[1] normal	$d = 2, b = 1/2$	$\frac{2}{\lambda\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x}{\lambda}\right)^2\right)$	$\operatorname{erfc}\left(\frac{x}{\sqrt{2}\lambda}\right)$
Power background measure			
Gamma[2]	$d = 1/c, b = 1,$ $\zeta = c$	$\frac{1}{\lambda\,\Gamma(\zeta)}\left(\frac{x}{\lambda}\right)^{\zeta-1}\exp\left(-\frac{x}{\lambda}\right)$	$\frac{\Gamma_{x/\lambda}(\zeta)}{\Gamma(\zeta)}$
Weibull[3]	$b = d = 1, \zeta = c$	$\frac{\zeta}{\lambda}\left(\frac{x}{\lambda}\right)^{\zeta-1}\exp\left(-\left(\frac{x}{\lambda}\right)^{\zeta}\right)$	$\exp\left(-\left(\frac{x}{\lambda}\right)^{\zeta}\right)$
Extended half[1] normal[4]	$b = 1, d = 2/c,$ $\zeta = c$	$\frac{2}{\lambda\,\Gamma(\zeta/2)}\left(\left(\frac{x}{\lambda}\right)^2\right)^{\frac{\zeta}{2}-\frac{1}{2}}\exp\left(-\left(\frac{x}{\lambda}\right)^2\right)$	$\frac{\Gamma_y(\zeta)}{\Gamma(\zeta)},\ y = \left(\frac{x}{\lambda}\right)^2$
Generalized gamma[5]	$b = 1, \zeta = c,$ $\varsigma = \zeta/cd$	$\frac{\zeta}{\lambda\,\varsigma\,\Gamma(\varsigma)}\left(\frac{x}{\lambda}\right)^{\zeta-1}\exp\left(-\left(\frac{x}{\lambda}\right)^{\zeta/\varsigma}\right)$	$\frac{\Gamma_y(\varsigma)}{\Gamma(\varsigma)},\ y = \left(\frac{x}{\lambda}\right)^{\zeta/\varsigma}$
Log-shift background measure ($c = 1$)			
Pareto[6]	$d = 1, \xi = 1/b$	$\frac{1}{\lambda\xi}\left(1+\frac{x}{\lambda}\right)^{-\frac{1}{\xi}-1}$	$\left(1+\frac{x}{\lambda}\right)^{-\frac{1}{\xi}}$
Lognormal	$d = 2, b = 1/2$	$\frac{2}{\sqrt{2\pi}\,\lambda}\frac{\exp\left(-\frac{1}{2}\left(\ln\left(1+\frac{x}{\lambda}\right)\right)^2\right)}{1+x/\lambda}$	$\operatorname{erfc}\left(\frac{1}{\sqrt{2}}\ln\left(1+\frac{x}{\lambda}\right)\right)$
Log-power-shift background measure			
Pareto-Burr-Feller (PBF)[7]	$d = 1, \zeta = c,$ $\xi = 1/bc$	$\frac{1}{\lambda\xi}\left(\frac{x}{\lambda}\right)^{\zeta-1}\left(1+\left(\frac{x}{\lambda}\right)^{\zeta}\right)^{-\frac{1}{\zeta\xi}-1}$	$\left(1+\left(\frac{x}{\lambda}\right)^{\zeta}\right)^{-\frac{1}{\zeta\xi}}$
Generalized log-gamma[8]	$b = 1, \zeta = c, \varsigma = d$	$\frac{\zeta\varsigma}{\Gamma(1/\varsigma)\,\lambda}\frac{\exp\left(-\left(\ln\left(1+\left(\frac{x}{\lambda}\right)^{\zeta}\right)\right)^{\varsigma}\right)}{\left(\frac{x}{\lambda}\right)^{1-\zeta}\left(1+\left(\frac{x}{\lambda}\right)^{\zeta}\right)}$	$\frac{\Gamma_z(1/\varsigma)}{\Gamma(1/\varsigma)},$ $z = \left(\ln\left(1+\left(\frac{x}{\lambda}\right)^{\zeta}\right)\right)^{\varsigma}$

[1] Distributions named "half" have their "full" version whose density $f(x)$ and exceedance $\bar{F}(x)$ are obtained by dividing those given in the table by 2. The "half" version corresponds to $x \geq 0$, while the "full" version is supported on the whole real line. The lognormal distribution has also a "full" version for $x \geq -\lambda$.

[2] Special cases: *Chi-squared* and *Erlang*.

[3] Special case: *Rayleigh*. Antisymmetric case (in which $F(x) \leftarrow \bar{F}(x)$): *Fréchet*.

[4] Also known as *Chi*.

[5] Special cases: *Maxwell–Boltzmann*, *Maxwell–Jüttner*, *Nakagami*. Antisymmetric: *Inverse-chi-squared*, *Inverse-gamma*, *Lévy*.

[6] More precisely, *Pareto II* or *Lomax*.

[7] Also known as *Pareto III and IV*, *Burr XII* and *Feller*. Antisymmetric: *Dagum* (often referred to in hydrology as kappa (Mielke, 1973; Mielke and Johnson, 1973; Hosking, 1994) but it is totally different from the kappa distribution used in other fields—see Appendix 2-II.).

[8] For $d = 1$ it becomes PBF with upper-tail index $\xi = 1/c$. For $d > 1$, $\xi = 0$ (all moments exist). For $d < 1$, $\xi = \infty$ (no moment exists).

2.12 Tails: heavy-tailed and light-tailed distributions

There is a substantial difference between the distributions corresponding to equation (2.92) and (2.93) on the one hand and (2.90) on the other. Specifically, the former are light-tailed and the latter heavy-tailed. In heavy-tailed distributions for any $t > 0$ (however small) the following limit diverges to infinity:

$$\lim_{x\to\infty} \mathrm{e}^{tx}\bar{F}(x) = \infty \tag{2.94}$$

In turn, a heavy-tailed distribution is characterized by the so-called *upper-tail index* (or, if there is no risk of ambiguity, simply *tail index*), defined as that number ξ for which the following asymptotic equation holds true:

$$\lim_{x\to\infty} x^{1/\xi}\bar{F}(x) = l_2 \tag{2.95}$$

where l_2 is a nonzero and finite constant. The distributions listed in Table 2.5 under the titles *Log-shift background measure* and *Log-power-shift background measure* are heavy tailed. Those distributions in which a parameter ξ appears have upper-tail index ξ (e.g., Pareto, Pareto-Burr-Feller). The remaining (e.g., lognormal) have upper-tail index zero (except a specific case of the generalized log-gamma, shown in the table footnotes, whose upper-tail index is infinite). At the same time, the moments of heavy-tailed distributions also diverge beyond a certain order, i.e., $\mathrm{E}\left[\underline{x}^q\right] = \infty$ for all $q > 1/\xi$. The distributions with zero upper-tail index, such as the lognormal distribution, have all their moments finite. For that reason, they are often regarded as light-tailed. However, the lognormal distribution clearly satisfies (2.94) and therefore according to this definition is heavy tailed.

In a similar manner, we can define a *lower-tail index*. Whenever the domain of the distribution is the entire line of real numbers, we must replace ∞ with $-\infty$ and $x^{1/\xi}$ with $(-x)^{1/\xi}$. However, usually we deal with nonnegative quantities (lower bounded by 0) and, in this case, we need a different way to define the lower-tail index. Specifically, the lower-tail index is that number ζ for which the following asymptotic equation holds true:

$$\lim_{x\to 0} x^{-\zeta}F(x) = l_3 \tag{2.96}$$

where l_3 is again a nonzero and finite constant*. Those distributions listed in Table 2.5, in which a parameter ζ appears, have lower-tail index ζ (e.g., Gamma, Weibull, Pareto-Burr-Feller). Using l'Hôpital's rule, we see that $\lim_{x\to 0} x^{-\zeta}F(x) = \lim_{x\to 0} x^{1-\zeta}f(x)/\zeta$. Thus, if $\zeta < 1$, the density $f(x)$ should necessarily be a decreasing function, at least close to the origin, with $\lim_{x\to 0} f(x) = \infty$. In contrast, when $\zeta > 1$, the density $f(x)$ is an increasing function close to the origin, with $f(0) = 0$, and is usually bell-shaped. The particular case $\zeta = 1$ is characteristic of the exponential and Pareto distributions, where $f(0)$ is finite and the density $f(x)$ is a decreasing function.

* It would be more natural to use $1/\zeta$ instead of ζ in (2.96) so that it be more consistent with (2.95). However, we have used that convention so that the parameterization of common distributions, such as Gamma and Weibull, be similar to the one dominating in the statistical literature.

Table 2.6 summarizes the above cases and extends them to all possible upper and lower tails and their indices. Notice that a single function, namely the odds function, $\Psi(x)$, can be used to calculate and visualize (on a log-log plot) both tails.

Table 2.6 Definitions of tail indices for the different cases of tail behaviour.

Characteristic	Definition of tail index[1]	Determination of tail index[2]
Upper bounded by c_{U}, tail index ζ'	$\lim_{x\to c_{\mathrm{U}}} (c_{\mathrm{U}}-x)^{-\zeta'}\overline{F}(x) = l_1$	$\zeta' = \overline{F}^{\#}_{c_{\mathrm{U}}}(0) = -\Psi^{\#}_{c_{\mathrm{U}}}(0)$
Upper unbounded, tail index ξ	$\lim_{x\to\infty} x^{1/\xi}\,\overline{F}(x) = l_2$	$\xi = -1/\overline{F}^{\#}(\infty) = 1/\Psi^{\#}(\infty)$
Lower bounded by c_{L}, tail index ζ	$\lim_{x\to c_{\mathrm{L}}} (x-c_{\mathrm{L}})^{-\zeta}F(x) = l_3$	$\zeta = F^{\#}_{c_{\mathrm{L}}}(0) = \Psi^{\#}_{c_{\mathrm{L}}}(0)$
Lower unbounded, tail index ξ'	$\lim_{x\to-\infty} (-x)^{1/\xi'}F(x) = l_4$	$\xi' = -1/F^{\#}(-\infty) = -1/\Psi^{\#}_{c_{\mathrm{U}}}(0)$

[1] $l_i, i = 1, \ldots, 4$ are nonzero and finite constants.

[2] $F^{\#}(c)$ is the log-log derivative (LLD; see section 3.6) of the function $F(x)$ at the point $x = c$ and $F_c(x) \coloneqq F(x + c)$. Likewise for $\Psi^{\#}(c)$.

Digression 2.M: The geophysical importance of heavy-tailed distributions

In classical statistical mechanics the Lebesgue measure is used as background measure. As a consequence, a constrained mean results in exponential distribution which notably has coefficient of variation $\sigma/\mu = 1$. However, in several geophysical processes, most notably rainfall, when the time scale is small (e.g., daily or hourly), the empirical σ/μ is greater than 1, which means that the exponential distribution is not suitable. One may think that adding one more constraint would fix the problem. The natural choice seems to be to constrain entropy maximization by both the mean μ and the variance σ^2. However, this does not work as, for nonnegative stochastic variables, entropy maximization with Lebesgue background measure cannot yield $\sigma/\mu > 1$. In other words, the exponential distribution is the upper limit.

The next solution to try is either to use a trickier (less natural) constraint, to change the definition of entropy (using a generalized definition; see Digression 2.E), or to change the background measure. The first two cases have been studied in Koutsoyiannis (2005a) and Papalexiou and Koutsoyiannis (2012) and the last one in Koutsoyiannis (2017). Whatever the choice may be, the result is practically the same: a heavy-tailed distribution. The easiest way to derive that distribution is by the framework described above, using the log-shift background measure and a single constraint, the mean of the distance function. The resulting Pareto distribution has $\sigma/\mu = 1/\sqrt{1-2\xi} > 1$.

In other words, by changing the background measure from Lebesgue to log-shift, the light-tailed exponential distribution switches to the heavy-tailed Pareto one. The theoretical framework otherwise remains unaffected—the same probabilistic definition of entropy is used in both cases. But the change in the derived distribution may have important consequences in the design and management related to extreme events.

We illustrate this based on real world data. Specifically, we use the daily rainfall data of Bologna, Italy (44.50°N, 11.35°E, 53.0 m). The available dataset is one of the longest daily rainfall records worldwide. The time series of observations is available online in the frame of the Global Historical Climatology Network – Daily (GHCN-Daily; Menne et al., 2012)[1]. It is uninterrupted for the period 1813-2007, 195 years in total. More recent daily data for 2008-2018 are provided by the online data repository Dext3r of ARPA Emilia Romagna, Rete di monitoraggio RIRER.[2] With these additional data, the record length becomes 206 years.

Figure 2.5 depicts the daily time series as well the (right-aligned) moving averages and moving maxima for a time window of 10 years, representing the 10-year climatic (time-averaged) values. The most spectacular behaviour shown in the figure is the changing climate (see definition of climate in Chapter 8): The 10-year climatic average daily rainfall has been changing between a

minimum of 1.2 mm (having occurred in the 1820s) and a maximum of 2.5 mm (having occurred at the decade ending in 1902)—more than twice the minimum. At the same time the 10-year climatic value of the maximum daily rainfall has varied between a minimum of 48.5 mm (having occurred in the 1820s) and a maximum of 155.7 mm (having occurred in the 1930s)—more than three times the minimum. These changes do not follow a linear pattern but have the form of long-term non-periodic fluctuations, up and down. In the most recent years, after 1950, there is a roughly increasing trend in both climatic indices, but such increasing trends were also observed before 1900, followed by drops thereafter. These trendy-looking behaviour reflects the Hurst-Kolmogorov dynamics (see section 3.12), omnipresent in geophysics.

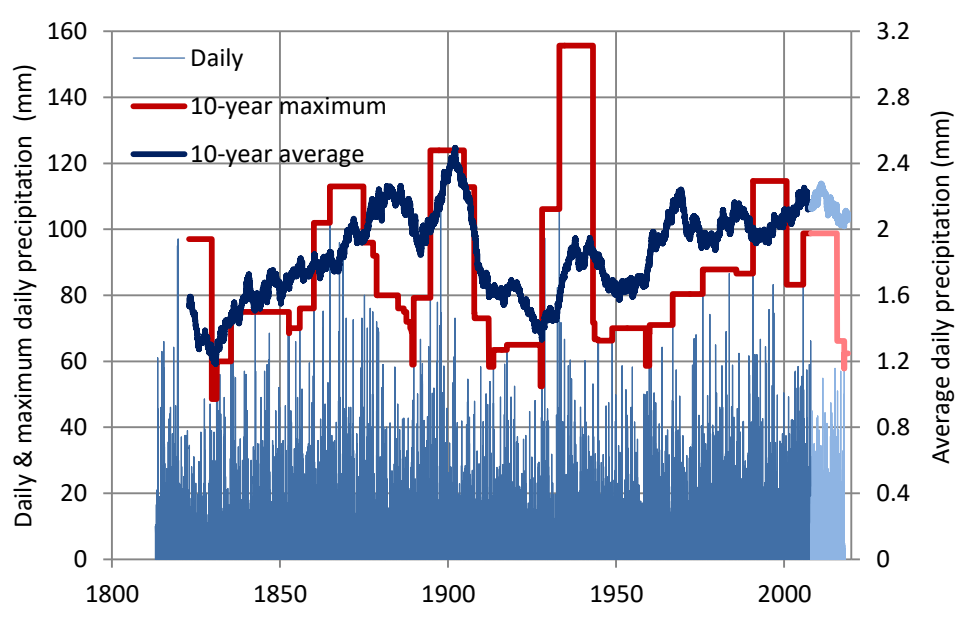

Figure 2.5 Plot of the time series of daily rainfall in Bologna, along with moving averages and moving maxima for a time window of 10 years (right-aligned, i.e., the value plotted at a specific year is the average or the maximum of the previous 10 years). The lines in darker colour represent the GHCN time series while those in lighter colour represent the newer data which are not included in the GHCN time series.

During the 206 years of observations there were 19 426 rain days, all of which are used in the modelling. The nonzero rainfall depths of all 19 426 days are plotted against their empirical return periods in Figure 2.6. The return period of an observed value x is related to the probability of exceedance by $T(x) = D/\bar{F}(x)$ (equation (2.42)) where D would be 1 d if all days were considered, while, after neglecting the zero rain days, D should be adapted to $D = 365.25 \times 206 / 19\,426 = 3.87$ d. The empirical return period for the sample of 19 426 daily rainfall values is estimated from equation (2.45).

The 19 426 values range between 0.1 and 155.7 mm, with a mean of 7.2 mm. In the exponential distribution the single parameter λ equals the mean, which allows plotting of the theoretical curve corresponding to it in Figure 2.6. Clearly, the comparison with the empirical points of the figure indicates a bad performance of the exponential model. In contrast, the Pareto model, also plotted in Figure 2.6 looks suitable. It is admirable that a model with only two parameters (the upper-tail index ξ and the scale parameter λ) can make such a good fitting on so many observations made over 206 years. The parameter values, $\xi = 0.11$ and $\lambda = 7.78$ mm (with parameterization as in Table 2.3), have been estimated by a least squares method to minimize the error between the empirical and theoretical return period. The empirical return period has been assigned by the method described in Digression 2.I. The good performance of the Pareto distribution suggests that the hypothesis of a log-shift background measure, along with the principle of maximum entropy, leads to a good predictive capacity.

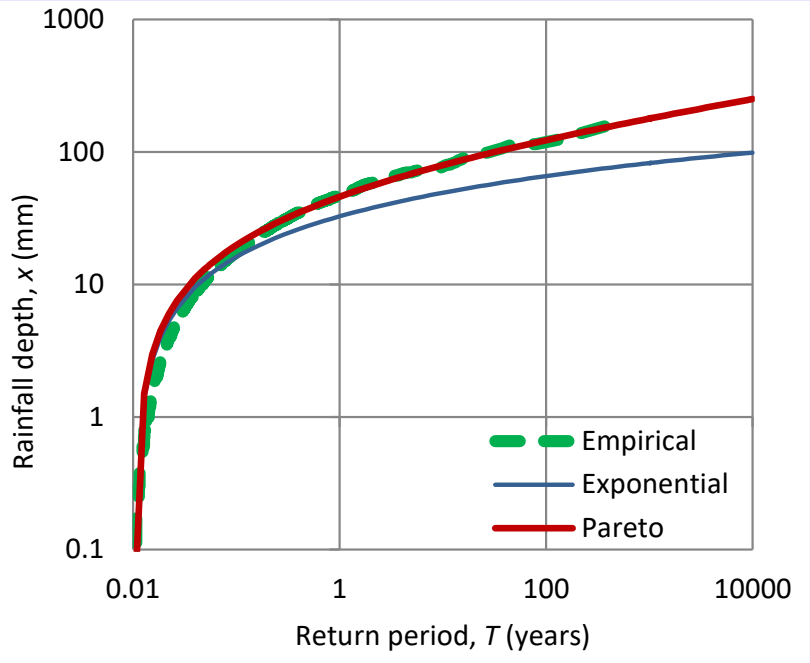

Figure 2.6 Rainfall depth vs. return period for Bologna based on 19 426 daily rainfall depths observed throughout the observation period of 206 years.

Now, comparing the behaviour of the light-tailed exponential distribution with the heavy-tailed Pareto distribution, and both with the empirical distribution, we clearly see that the former severely underestimates the magnitude of the extremes. Notably, for a return period of 10 000 years, which is typically used in the engineering design of major projects such as dams, Figure 2.6 shows that the exponential distribution predicts a rainfall depth of ~100 mm, a value that was in fact exceeded seven times in the 206-year available record. On the other hand, the Pareto distribution predicts a value of ~250 mm, 2.5 times higher (and it becomes even higher if we also take into account the dependence structure of rainfall; Koutsoyiannis, 2025). Thus, inappropriate model selection based on inappropriate theoretical considerations can have substantial consequences when it comes to practical application. Sooner or later, Nature will expose how the inappropriate they are (e.g. by frequent exceedances of design values). In such cases, one could re-examine the theory (even though an alternative and more popular practice is to blame external agents and find good scapegoats).

Indeed, in the 20th century, the light-tailed distributions constituted the dominant theoretical model in research and engineering practice. And given the substantial underestimation of extremes by this model, its failure (and its severe consequences) should not come as a surprise. By now, both theoretical advances and accumulated empirical evidence have shaken this model and pointed to heavy-tailed distributions.

In addition, Koutsoyiannis (2004a, 2005a, 2007) discussed several theoretical reasons that favour the heavy tailed distributions over the exponential case, which are consistent with the above considerations related to the log-shift background measure. Furthermore, as already discussed (Chapter 1, the omnipresence of change and the non-static climate are consistent with heavy-tailed distributions, as will be further illustrated in Digression 3.I.

---

[1] GHCN Version 3; data retrieved on 2019-02-17 from https://climexp.knmi.nl/gdcnprcp.cgi?WMO=ITE00100550.

[2] Data retrieved on 2019-02-17 from http://www.smr.arpa.emr.it/dext3r/. In particular, the data from the station Bologna Idrografico (coordinates 44.499883°N, 11.346156°E, 84.0 m, practically the same as those given for the GHCN station (except a 31 m difference in the elevation, perhaps indicating that this particular station is located at the roof of a building), were used except for year 2008 for which no data are provided for this station. For this year, as well as for very few days with missing values in other years, the daily precipitation values of the station Bologna Urbana (44.500754°N, 11.328789°E, 78.0 m) were used instead.

2.13 Two variables: joint distribution and joint moments

The above sections have been devoted to concepts of probability pertaining to the analysis of a single variable $\underline{x}$. Often, however, the simultaneous modelling of two (or more) variables is necessary. Let the pair of stochastic variables $\left(\underline{x}, \underline{y}\right)$ defined on two ground sets (Ω_x, Ω_y), respectively. The intersection (simultaneous occurrence) of the two events $\{\underline{x} \leq x\}$ and $\left\{\underline{y} \leq y\right\}$, denoted as $\left\{\underline{x} \leq x, \underline{y} \leq y\right\}$ is an event of the sample space $\Omega_{xy} = \Omega_x \times \Omega_y$. Based on the latter event, we can define the *joint probability distribution function* of $\left(\underline{x}, \underline{y}\right)$ as a function of the real variables (x, y):

$$F_{\underline{x}\underline{y}}(x, y) \coloneqq P\left\{\underline{x} \leq x, \underline{y} \leq y\right\} \tag{2.97}$$

The subscripts $\underline{x}\underline{y}$ can be omitted if there is no risk of ambiguity.

If $F_{\underline{x}\underline{y}}$ is differentiable, then the function:

$$f_{\underline{x}\underline{y}}(x, y) \coloneqq \frac{\partial^2 F_{\underline{x}\underline{y}}(x, y)}{\partial x \, \partial y} \tag{2.98}$$

is the *joint probability density function* of the two variables. Obviously, the following equation holds:

$$F_{\underline{x}\underline{y}}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{\underline{x}\underline{y}}(u, v) \mathrm{d}u \mathrm{d}v \tag{2.99}$$

The functions:

$$F_{\underline{x}}(x) \coloneqq P\{\underline{x} \leq x\} = \lim_{y \to \infty} F_{\underline{x}\underline{y}}(x, y), \qquad F_{\underline{y}}(y) \coloneqq P\left\{\underline{y} \leq y\right\} = \lim_{x \to \infty} F_{\underline{x}\underline{y}}(x, y) \tag{2.100}$$

are called the *marginal probability distribution functions* of $\underline{x}$ and $\underline{y}$, respectively. Also, the *marginal probability density functions* can be determined as

$$f_{\underline{x}}(x) = \int_{-\infty}^{\infty} f_{\underline{x}\underline{y}}(x, y) \mathrm{d}y, \qquad f_{\underline{y}}(y) = \int_{-\infty}^{\infty} f_{\underline{x}\underline{y}}(x, y) \mathrm{d}x \tag{2.101}$$

Similar to the univariate case, we can define the expected value of any given function $g\left(\underline{x}, \underline{y}\right)$ of the stochastic variables $\underline{x}$ and $\underline{y}$ by

$$\mathrm{E}\left[g\left(\underline{x}, \underline{y}\right)\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) \; f_{\underline{x}\underline{y}}(x, y) \mathrm{d}x \mathrm{d}y \tag{2.102}$$

In this manner, we define the (*noncentral* or *about the origin*) *joint moment* of orders p, q as:

$$\mu'_{pq} \coloneqq \mathrm{E}\left[\underline{x}^p \underline{y}^q\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q \, f_{\underline{x}\underline{y}}(x,y)\mathrm{d}x\mathrm{d}y \tag{2.103}$$

as well as the *joint central moment* of orders p, q:

$$\mu_{pq} \coloneqq \mathrm{E}\left[\left(\underline{x}-\mu_x\right)^p \left(\underline{y}-\mu_y\right)^q\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(\underline{x}-\mu_x\right)^p \left(\underline{y}-\mu_y\right)^q f_{\underline{x}\underline{y}}(x,y)\mathrm{d}x\mathrm{d}y \tag{2.104}$$

If $p = 0$ or $q = 0$, we get the *marginal moments* (e.g., means, $\mu_x \coloneqq \mu'_{10}, \mu_y \coloneqq \mu'_{01}$; variances, $\mathrm{var}[\underline{x}] \coloneqq \mathrm{E}\left[\left(\underline{x}-\mu_x\right)^2\right] \equiv \mu_{20} \equiv \gamma_x \equiv \sigma_x^2$, $\mathrm{var}\left[\underline{y}\right] = \mu_{02} \equiv \gamma_y \equiv \sigma_y^2$, etc.). The lowest-order joint central moment is the *covariance*:

$$\mathrm{cov}\left[\underline{x},\underline{y}\right] \coloneqq \mathrm{E}\left[\left(\underline{x}-\mu_x\right)\left(\underline{y}-\mu_y\right)\right] \equiv \mu_{11} \equiv \sigma_{xy} = \mathrm{E}\left[\underline{x}\,\underline{y}\right] - \mathrm{E}[\underline{x}]\mathrm{E}\left[\underline{y}\right] \tag{2.105}$$

A dimensionless index derived from covariance is the *correlation coefficient*:

$$r_{xy} \coloneqq \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{2.106}$$

which obeys the inequality:

$$-1 \le r_{xy} \le 1 \tag{2.107}$$

where the values –1 and 1 indicate *fully anti-correlated* (*fully negatively correlated*) and *fully (positively) correlated* variables. Here is the mathematical proof about why this happens: We start from the obvious relationship $\mathrm{E}\left[\left(\underline{x}+\underline{y}\right)^2\right] = \mathrm{E}[\underline{x}^2] + \mathrm{E}\left[\underline{y}^2\right] + 2\mathrm{E}\left[\underline{x}\,\underline{y}\right]$, observing that terms involving squares are nonnegative quantities. We assume, without loss of generality, $\mathrm{E}[\underline{x}] = \mathrm{E}\left[\underline{y}\right] = 0$, so that $\mathrm{E}[\underline{x}^2] = \sigma_x^2, \mathrm{E}\left[\underline{y}^2\right] = \sigma_y^2, \mathrm{E}\left[\underline{x}\,\underline{y}\right] = \sigma_{xy}$. Thus, we get $\sigma_x^2 + \sigma_y^2 + 2\sigma_{xy} \ge 0$ or $\sigma_{xy}/\sigma_x\sigma_y \ge -(1/2)\left(\sigma_x/\sigma_y + \sigma_y/\sigma_x\right) = -(1/2)\,(a + 1/a)$, where $a \coloneqq \sigma_x/\sigma_y \ge 0$. It is easy to see that $(a + 1/a)$ has minimum value 2, so that $\sigma_{xy}/\sigma_x\sigma_y \ge -1$. Furthermore, $\mathrm{E}\left[\left(\underline{x}-\underline{y}\right)^2\right] = \mathrm{E}[\underline{x}^2] + \mathrm{E}\left[\underline{y}^2\right] - 2\mathrm{E}\left[\underline{x}\,\underline{y}\right]$ and, likewise, $\sigma_{xy}/\sigma_x\sigma_y \le (1/2)\left(\sigma_x/\sigma_y + \sigma_y/\sigma_x\right) = (1/2)(a + 1/a) \le 1$.

The particular case where:

$$\sigma_{xy} = r_{xy} = 0 \Leftrightarrow \mathrm{E}\left[\underline{x}\,\underline{y}\right] = \mathrm{E}[\underline{x}]\mathrm{E}\left[\underline{y}\right] \tag{2.108}$$

defines *uncorrelated variables*. *Independent variables* are necessarily uncorrelated, but independence is a stricter concept whose definition is:

$$F_{\underline{x}\underline{y}}(x,y) = F_{\underline{x}}(x)F_{\underline{y}}(y), \qquad f_{\underline{x}\underline{y}}(x,y) = f_{\underline{x}}(x)f_{\underline{y}}(y) \tag{2.109}$$

The joint entropy is defined in an analogous manner with that in the univariate case (section 2.10) and corresponds to the product partition (section 2.3). For discrete stochastic variables the entropy is:

$$\Phi\left[\underline{x},\underline{y}\right] \coloneqq \mathrm{E}\left[-\ln P\left(\underline{x},\underline{y}\right)\right] = -\sum_{i,j} P_{ij} \ln P_{ij} \tag{2.110}$$

where $P_{ij} \coloneqq P\left\{\underline{x} = x_i, \underline{y} = y_j\right\}$. For continuous stochastic variables it is:

$$\Phi\left[\underline{x},\underline{y}\right] \coloneqq \mathrm{E}\left[-\ln\frac{f\left(\underline{x},\underline{y}\right)}{\beta\left(\underline{x},\underline{y}\right)}\right] = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \ln\frac{f(x,y)}{\beta(x,y)} f(x,y)\mathrm{d}x\,\mathrm{d}y \tag{2.111}$$

2.14 Conditional densities and expectations

Of particular interest are the so-called *conditional probability distribution function* and *conditional probability density function* of $\underline{x}$ for a specified value of $\underline{y} = y$; these are given by:

$$F_{\underline{x}|\underline{y}}(x|y) \coloneqq \frac{\int_{-\infty}^{x} f_{\underline{x}\underline{y}}(\xi,y)\mathrm{d}\xi}{f_{\underline{y}}(y)}, \qquad f_{\underline{x}|\underline{y}}(x|y) \coloneqq \frac{f_{\underline{x}\underline{y}}(x,y)}{f_{\underline{y}}(y)} \tag{2.112}$$

respectively. Switching $\underline{x}$ and $\underline{y}$ we obtain the conditional functions of $\underline{y}$.

The *conditional expected value* of a function $g(\underline{x})$ for a specified value of $\underline{y} = y$ is defined by

$$\mathrm{E}\left[g(\underline{x})|y\right] \coloneqq \mathrm{E}\left[g(\underline{x})\middle|\underline{y} = y\right] = \int_{-\infty}^{\infty} g(x) f_{\underline{x}|\underline{y}}(x|y)\mathrm{d}x \tag{2.113}$$

An important quantity of this type is the conditional expected value of $\underline{x}$:

$$\mathrm{E}\left[\underline{x}|y\right] \coloneqq \mathrm{E}\left[\underline{x}\middle|\underline{y} = y\right] = \int_{-\infty}^{\infty} x f_{\underline{x}|\underline{y}}(x|y)\mathrm{d}x \tag{2.114}$$

Likewise, the conditional variance is

$$\mathrm{var}\left[\underline{x}|y\right] \coloneqq \mathrm{E}\left[\left(\underline{x} - \mathrm{E}\left[\underline{x}|y\right]\right)^2\middle|\underline{y} = y\right] = \int_{-\infty}^{\infty} \left(x - \mathrm{E}\left[\underline{x}|y\right]\right)^2 f_{\underline{x}|\underline{y}}(x|y)\mathrm{d}x \tag{2.115}$$

and can also be written as

$$\mathrm{var}\left[\underline{x}|y\right] \coloneqq \mathrm{E}\left[\underline{x}^2|y\right] - \left(\mathrm{E}\left[\underline{x}|y\right]\right)^2 \tag{2.116}$$

It is obvious from the definition (2.113) that the conditional expectation $\mathrm{E}\left[g(\underline{x})|y\right]$ is a function of the real variable y, call it $h(y)$, rather than a constant. If we do not specify the value y of the stochastic variable $\underline{y}$ in the condition, then the quantity $\mathrm{E}\left[g(\underline{x})\middle|\underline{y}\right] = h\left(\underline{y}\right)$ becomes a function of the stochastic variable $\underline{y}$. Hence, it is a stochastic variable itself. Its own expected value is:

$$\mathrm{E}\left[\mathrm{E}\left[g(\underline{x})\middle|\underline{y}\right]\right] = \int_{-\infty}^{\infty} \mathrm{E}[g(\underline{x})|y] f_{\underline{y}}(y)\mathrm{d}y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x) f_{\underline{x}\underline{y}}(x,y)\,\mathrm{d}x\mathrm{d}y \tag{2.117}$$

where we have utilized (2.113) and (2.112). As a result,

$$\mathrm{E}\left[\mathrm{E}\left[g(\underline{x})\middle|\underline{y}\right]\right] = \mathrm{E}[g(\underline{x})] \tag{2.118}$$

This can be readily generalized for a function of two stochastic variables, i.e.,

$$\mathrm{E}\left[\mathrm{E}\left[g\left(\underline{x},\underline{y}\right)\middle|\underline{y}\right]\right] = \mathrm{E}\left[g\left(\underline{x},\underline{y}\right)\right] \tag{2.119}$$

The entropy of the finest partition, as derived in section 2.10 and for the bivariate case in equations (2.110) and (2.111), is an expectation and thus we can also define conditional forms of entropy which are quite useful. Thus, for a specified value of $\underline{y} = y$ and for a discrete stochastic variable the entropy is:

$$\Phi[\underline{x}|y] \coloneqq \mathrm{E}[-\ln P(\underline{x}|y)] = -\sum_{i,j} P_{i|j} \ln P_{i|j} \tag{2.120}$$

where $P_{i|j} \coloneqq P\left\{\underline{x} = x_i | \underline{y} = y_j\right\}$. This corresponds to the event-conditional entropy (section 2.3). For a continuous stochastic variable it is:

$$\Phi[\underline{x}|y] \coloneqq \mathrm{E}\left[-\ln\frac{f(\underline{x}|y)}{\beta(\underline{x})}\right] = -\int_{-\infty}^{\infty} \ln\frac{f(x|y)}{\beta(x)} f(x|y)\mathrm{d}x \tag{2.121}$$

These quantities depend on the specified conditioning value y of $\underline{y}$. However, we can define a global conditional entropy for an unspecified value of $\underline{y}$:

$$\Phi\left[\underline{x}|\underline{y}\right] \coloneqq \mathrm{E}\left[\mathrm{E}\left[-\ln\ P\left(\underline{x}|\underline{y}\right)|\underline{y}\right]\right] = \mathrm{E}\left[-\ln P\left(\underline{x}|\underline{y}\right)\right] = -\sum_{j}\sum_{i} P_{i|j} \ln P_{i|j}\, P_j \tag{2.122}$$

Here we have simplified the notation to $\Phi\left[\underline{x}|\underline{y}\right]$ instead of the more accurate $\mathrm{E}\left[\Phi\left[\underline{x}|\underline{y}\right]\right]$. This corresponds to partition-conditional entropy (section 2.3). For continuous stochastic variables it is:

$$\begin{aligned}\Phi\left[\underline{x}|\underline{y}\right] &\coloneqq \mathrm{E}\left[\mathrm{E}\left[-\ln\frac{f\left(\underline{x}|\underline{y}\right)}{\beta(\underline{x})}|\underline{y}\right]\right] = \mathrm{E}\left[-\ln\frac{f\left(\underline{x}|\underline{y}\right)}{\beta(\underline{x})}\right] \\ &= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \ln\frac{f(x|y)}{\beta(x)} f(x|y) f(y)\mathrm{d}x\,\mathrm{d}y = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \ln\frac{f(x|y)}{\beta(x)} f(x,y)\mathrm{d}x\,\mathrm{d}y\end{aligned} \tag{2.123}$$

From inequality (2.24) we have

$$\Phi\left[\underline{x}|\underline{y}\right] \leq \Phi[\underline{x}] \tag{2.124}$$

Another distinction we have to stress is that:

$$\Phi\left[\underline{x}|\underline{y}\right] \neq \Phi[\underline{x}|y] \tag{2.125}$$

because the quantity on the right-hand side is generally a function of y while that on the left is not. An interesting exception is the case of a bivariate normal distribution in which $\Phi[\underline{x}|y]$ turns out to be a constant rather than a function of y ($\Phi[\underline{x}|y] =: \Phi_c[\underline{x}] \leq \Phi[\underline{x}]$). Generally, we should stress that:

- conditional expectations like $\mathrm{E}[\underline{x}|y]$ are deterministic functions of the conditioning value y;
- conditional expectations like $\mathrm{E}[\underline{x}|\underline{y}]$ are stochastic variables, depending on $\underline{y}$;
- expectations of conditional expectations, as in $\mathrm{E}\left[\mathrm{E}\left[g\left(\underline{x},\underline{y}\right)\left|\underline{y}\right.\right]\right]$ and $\Phi\left[\underline{x}|\underline{y}\right]$, are constants.

These remarks have to be added to the notes of Digression 2.H about the importance of notation.

Digression 2.N: Does information decrease entropy?

It is intuitive to think that, if a stochastic variable $\underline{x}$ has some relationship with another stochastic variable $\underline{y}$ then observation of the value of $\underline{y}$ would decrease our uncertainty on $\underline{x}$. (The contrary of this, referred to independent variables, has been used to define entropy). As entropy is a formal quantification of uncertainty, this can be formally stated as follows: the conditional entropy of $\underline{x}$ given information on $\underline{y}$ cannot be greater than the unconditional entropy of $\underline{x}$. However, this simple truth is sometimes contradicted in scientific texts, the reason being the inattentive use of concepts and notation. We will illustrate them with the following example.

In Digression 2.C and Digression 2.D we studied the probabilities of the dry and wet (rain) state in an area. Continuing this study, we now introduce the stochastic variables $\underline{x}$ and $\underline{y}$ for today's and yesterday's state, respectively, with $\{\underline{x} = 0\}$ and $\{\underline{x} = 1\}$ representing a dry and wet state of today, respectively, and likewise for yesterday. We assume for illustration the conditional probabilities:

$$P\left\{\underline{x} = 1|\underline{y} = 1\right\} = 0.4, \qquad P\left\{\underline{x} = 1|\underline{y} = 0\right\} = 0.15$$

from which it directly follows that

$$P\left\{\underline{x} = 0|\underline{y} = 1\right\} = 0.6, \qquad P\left\{\underline{x} = 0|\underline{y} = 0\right\} = 0.85$$

and after some simple calculations (see Digression 2.D) we also find the marginal probabilities to be

$$P\{\underline{x} = 0\} = 0.8, \qquad P\{\underline{x} = 1\} = 0.2$$

Hence the unconditional entropy is:

$$\Phi[\underline{x}] = \mathrm{E}[-\ln P(\underline{x})] = -0.8\ln 0.8 - 0.2\ln 0.2 = 0.500$$

while the entropy conditional on yesterday being dry is:

$$\Phi\left[\underline{x}|\underline{y} = 0\right] = \mathrm{E}\left[-\ln P\left(\underline{x}\middle|\underline{y} = 0\right)\right] = -0.85\ln 0.85 - 0.15\ln 0.15 = 0.423$$

and that conditional on yesterday being wet is:

$$\Phi\left[\underline{x}|\underline{y} = 1\right] = \mathrm{E}\left[-\ln P\left(\underline{x}\middle|\underline{y} = 1\right)\right] = -0.6\ln 0.6 - 0.4\ln 0.4 = 0.673$$

These correspond to the event-conditional entropies (section 2.3). It is true that the added information that yesterday was a wet day increased the entropy from 0.5 (when there was no information) to 0.673. This happened because the probabilities of the two states, which initially were 0.8 vs. 0.2, far from the equiprobability (0.5) in which the entropy is maximized, have now approached each other (0.6 vs. 0.4). That is why the event-conditional entropy has increased.

However, this happens for that particular value, $\underline{y} = 1$. When we consider all values (we have two here), on the average the conditional entropy (partition-conditional entropy) is

$$\Phi\left[\underline{x}|\underline{y}\right] = 0.423 \times 0.8 + 0.673 \times 0.2 = 0.473 < 0.500$$

In other words, the reply to the question in the above title is: Yes, information decreases entropy, but we must be attentive about the correct use of the probabilistic concepts.

2.15 Many variables

All the above theoretical analyses can easily be extended to more than two stochastic variables. For instance, the distribution function of the n stochastic variables $\underline{x}_1, \dots, \underline{x}_n$ is

$$F_{\underline{x}_1,\dots,\underline{x}_n}(x_1, \dots, x_n) := P\{\underline{x}_1 \le x_1, \dots, \underline{x}_n \le x_n\} \tag{2.126}$$

and is related to the n-dimensional probability density function by

$$F_{\underline{x}_1,\dots,\underline{x}_n}(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} f_{\underline{x}_1,\dots,\underline{x}_n}(\xi_1, \dots \xi_n)\mathrm{d}\xi_n \cdots \mathrm{d}\xi_1 \tag{2.127}$$

The variables $\underline{x}_1, \dots, \underline{x}_n$ are independent if:

$$F_{\underline{x}_1,\dots,\underline{x}_n}(x_1, \dots, x_n) = F_{\underline{x}_1}(x_1) \cdots F_{\underline{x}_n}(x_n) \tag{2.128}$$

A useful rule to mention is the so-called chain rule, which allows for the expression of joint densities as products of conditional densities:

$$f(x_1, \dots, x_n) = f(x_n|x_{n-1}, \dots, x_1) \cdots f(x_2|x_1)f(x_1) \tag{2.129}$$

where for notational brevity we have omitted the subscripts of functions $f(\,)$, as these resemble the arguments of the functions. A direct consequence that allows the evaluation of entropy is

$$\Phi\left[\underline{x}_1, \dots, \underline{x}_n\right] = \Phi\left[\underline{x}_n|\underline{x}_{n-1}, \dots, \underline{x}_1\right] + \cdots + \Phi\left[\underline{x}_2|\underline{x}_1\right] + \Phi\left[\underline{x}_1\right] \tag{2.130}$$

The expected values and moments are defined in a similar manner to the case with two variables, and all properties discussed in section 2.13 are likewise generalized for functions of many variables.

If we integrate $f(x_1, \dots, x_n)$ with respect to some of the variables, we obtain the joint density of the remaining variables. For example

$$f(x_1, x_3) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4)\,\mathrm{d}x_2\mathrm{d}x_4 \tag{2.131}$$

and since

$$f(x_1, x_2, x_3, x_4) = f(x_1, x_3|x_2, x_4)f(x_2, x_4) \tag{2.132}$$

we obtain

$$f(x_1, x_3) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_3 | x_2, x_4)\, f(x_2, x_4)\, \mathrm{d}x_2 \mathrm{d}x_4 \tag{2.133}$$

which can also be written as an expected value, i.e.,

$$f(x_1, x_3) = \mathrm{E}\big[f\big(x_1, x_3 \big| \underline{x}_2, \underline{x}_4\big)\big] \tag{2.134}$$

where the conditioning variables $\underline{x}_2, \underline{x}_4$ are taken as stochastic variables and the conditioned ones are taken as values.

2.16 Linear combinations of stochastic variables

A consequence of the definition of the expected value is the relationship

$$\mathrm{E}\big[c_1 g_1(\underline{x}_1, \underline{x}_2) + c_2 g_2(\underline{x}_1, \underline{x}_2)\big] = c_1 \mathrm{E}\big[g_1(\underline{x}_1, \underline{x}_2)\big] + c_2 \mathrm{E}\big[g_2(\underline{x}_1, \underline{x}_2)\big] \tag{2.135}$$

where c_1 and c_2 are any constants, and g_1 and g_2 any functions. This property can be extended to any number of functions g_i. Applying it to the weighted sum of two variables we obtain

$$\mathrm{E}\big[a_1 \underline{x}_1 + a_2 \underline{x}_2\big] = a_1 \mathrm{E}\big[\underline{x}_1\big] + a_2 \mathrm{E}\big[\underline{x}_2\big] \tag{2.136}$$

Likewise, we can calculate the variance of the weighted sum. After some algebraic operations we get

$$\mathrm{var}\big[a_1 \underline{x}_1 + a_2 \underline{x}_2\big] = a_1^2 \mathrm{var}\big[\underline{x}_1\big] + a_2^2 \mathrm{var}\big[\underline{x}_2\big] + 2 a_1 a_2 \mathrm{cov}\big[\underline{x}_1, \underline{x}_2\big] \tag{2.137}$$

It is more difficult to calculate the probability distribution function of such combinations. As an example, for the simplest case, the sum $\underline{z} = \underline{x}_1 + \underline{x}_2$ of two independent variables $\underline{x}_1$ and $\underline{x}_2$ has density:

$$f_{\underline{z}}(z) = \int_{-\infty}^{\infty} f_{\underline{x}_1}(z - x_2)\, f_{\underline{x}_2}(x_2) \mathrm{d}x_2 \tag{2.138}$$

The right-hand side is known as the *convolution integral* of $f_{\underline{x}_1}(x)$ and $f_{\underline{x}_2}(x)$. For nonnegative variables it takes the form:

$$f_{\underline{z}}(z) = \int_{0}^{z} f_{\underline{x}_1}(z - x_2)\, f_{\underline{x}_2}(x_2) \mathrm{d}x_2, \qquad z > 0 \tag{2.139}$$

2.17 Variance-based correlation and the climacogram

While covariance and its equivalent standardized form, i.e., correlation, have been the most customary tools to characterize dependence, they are neither the only nor the most effective ones. Assuming two stochastic variables $\underline{x}_1$ and $\underline{x}_2$ (possibly representing different physical quantities) with means μ_i ($i = 1,2$), standard deviations σ_i, covariance σ_{12} and correlation coefficient r_{12}, we may form a different type of a correlation coefficient

and covariance by examining a weighted sum of the two variables. Namely, we examine the average of the variables $\underline{x}_i$ after standardizing them with their standard deviations σ_i, which is necessary if they represent different physical quantities, in order to make them compatible for addition. From (2.137) we obtain the variance of this average which is:

$$\mathrm{var}\left[\frac{1}{2}\left(\frac{\underline{x}_1}{\sigma_1}+\frac{\underline{x}_2}{\sigma_2}\right)\right]=\frac{1}{2}+\frac{1}{2}\mathrm{cov}\left[\frac{\underline{x}_1}{\sigma_1},\frac{\underline{x}_2}{\sigma_2}\right]=\frac{1}{4}\mathrm{E}\left[\left(\frac{\underline{x}_1-\mu_1}{\sigma_1}+\frac{\underline{x}_2-\mu_2}{\sigma_2}\right)^2\right] \tag{2.140}$$

where we can recognize in the middle term the correlation coefficient r_{12}. Defining

$$\rho_{12}\coloneqq\mathrm{var}\left[\frac{1}{2}\left(\frac{\underline{x}_1}{\sigma_1}+\frac{\underline{x}_2}{\sigma_2}\right)\right],\qquad \gamma_{12}\coloneqq\sigma_1\sigma_2\rho_{12}=\mathrm{var}\left[\frac{1}{2}\left(\sqrt{\frac{\sigma_2}{\sigma_1}}\underline{x}_1+\sqrt{\frac{\sigma_1}{\sigma_2}}\underline{x}_2\right)\right] \tag{2.141}$$

we find from (2.140) that

$$\rho_{12}\coloneqq\frac{1+r_{12}}{2},\qquad \gamma_{12}\coloneqq\frac{\sigma_1\sigma_2+\sigma_{12}}{2} \tag{2.142}$$

Obviously, the same information as in r_{12} is contained in ρ_{12}, which lies in the interval [0, 1] with the values 0, 1/2, 1 representing fully anti-correlated, uncorrelated and fully correlated variables, respectively. Consequently, γ_{12} lies in the interval $[0, \sigma_1\sigma_2]$ with the values $0, \sigma_1\sigma_2/2, \sigma_1\sigma_2$ representing fully anti-correlated, uncorrelated and fully correlated variables, respectively.

The power of the notion of ρ_{12} and γ_{12} is in the fact that, unlike r_{12}, they can be readily expanded to many variables to provide a macroscopic (or bulk) indicator of correlation among all of them. Considering a number $\kappa > 0$ of stochastic variables, in the customary case where all have identical variances $\gamma_1 = \sigma^2$, we write:

$$\rho_\kappa\coloneqq\mathrm{var}\left[\frac{\underline{X}_\kappa}{\kappa\sigma}\right]=\frac{\gamma_\kappa}{\gamma_1},\qquad \gamma_\kappa\coloneqq\mathrm{var}\left[\frac{\underline{X}_\kappa}{\kappa}\right]=\gamma_1\rho_\kappa,\qquad \underline{X}_\kappa\coloneqq\underline{x}_1+\cdots+\underline{x}_\kappa \tag{2.143}$$

The quantity γ_κ as a function of the time scale κ is termed the *climacogram*, and ρ_κ is a dimensionless (standardized) climacogram. They range in the intervals $(0, \gamma_1)$ and (0, 1), respectively, with the highest value representing full correlation ($\underline{x}_1+\cdots+\underline{x}_\kappa=\kappa\underline{x}_1+c$, where c is a constant) and the lowest representing deterministic linear dependence, i.e. the condition that $\underline{x}_1+\cdots+\underline{x}_\kappa=c$). In the case of independence, $\gamma_\kappa\coloneqq\gamma_1/\kappa$ and $\rho_\kappa\coloneqq 1/\kappa$.

2.18 Limiting distributions and the central limit theorem

As we have seen in section 2.16, it is rather difficult to calculate the distribution function of the sum of two stochastic variables from the distributions of the constituents. This difficulty increases as the number of constituents increases. However, if this number becomes quite large, paradoxically the problem becomes easier—this is the ease of macroscopization. A central role in resolving this paradox is played by the *central limit*

theorem[*], one of the most important in probability theory. It concerns the limiting distribution function of a sum of stochastic variables–constituents, which, under some conditions (irrespective of the distribution functions of the constituents) is always the same: the celebrated *normal distribution*. This is the most commonly used distribution in probability theory as well as in all scientific disciplines and, as we have seen in section 2.11, it is also derived from the principle of maximum entropy.

Let $\underline{x}_i$ $(i = 1, \dots, n)$ be stochastic variables whose sum $\underline{z}_n \coloneqq \underline{x}_1 + \cdots + \underline{x}_n$ has mean μ_z and variance σ_z^2. The central limit theorem states that, under some general conditions (see below), as n tends to infinity, the distribution of $\underline{z}$ will tend to the normal distribution (also known as Gauss or Gaussian distribution and denoted as $\mathrm{N}(\mu_z, \sigma_z)$), i.e.,

$$\lim_{n\to\infty} F_{\underline{z}_n}(z) = \int_{-\infty}^{z} \frac{1}{\sigma_z\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{z-\mu_z}{\sigma_z}\right)^2} \tag{2.144}$$

and in addition, if $\underline{x}_i$ are continuous variables, the density function of $\underline{z}_n$ also has a limit:

$$\lim_{n\to\infty} f_{\underline{z}_n}(z) = \frac{1}{\sigma_z\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{z-\mu_z}{\sigma_z}\right)^2} \tag{2.145}$$

We observe in (2.144) and (2.145) that the limits of the functions $F_{\underline{z}_n}(z)$ and $f_{\underline{z}_n}(z)$ do not depend on the distribution functions of $\underline{x}_i$, that is, the result is always the same. Thus, provided that the conditions for the applicability of the theorem hold, (a) we can know the macroscopic behaviour (the distribution function of the sum) without knowing details of the constituents, and (b) precisely the same distribution will describe any variable that is a sum of a large number of components. Here lies the great importance of the normal distribution in all sciences (mathematical, physical, social, economic, etc., as well as stochastics per se). Additionally, we recall (Digression 2.L and Table 2.4) that the normal distribution also emerges from the principle of maximum entropy: for constant (Lebesgue) background density and for domain $(-\infty, \infty)$, it yields the maximum entropy among all distributions with specified (constrained) mean and standard deviation.

Therefore, we can consider the central limit theorem to be another manifestation of the principle of maximum entropy. As a system grows to involve many degrees of freedom, uncertainty increases towards its maximum possible value, resulting in a normal distribution.

In practice, the convergence for $n \to \infty$ can be regarded as an approximation if n is sufficiently large. But how large does n need to be to ensure that the approximation is satisfactory? Generally, the literature suggests that a value $n = 30$ is satisfactory. However, this varies depending on the (joint) distribution of the constituents $\underline{x}_i$. Figure 2.7 gives a graphical illustration of the convergence to the normal distribution of the sum of n independent variables. Clearly, if the distribution of $\underline{x}_i$ is symmetric (left panel, with

[*] The term was most likely introduced by Pólya in 1920. A first version of the theorem was formulated and proved by Laplace in 1810 while at about the same time Gauss studied the normal distribution in characterizing measurement or model errors. Earlier, in 1733, de Moivre had introduced the normal distribution as an approximation of the binomial distribution (Fischer, 2010).

uniform distribution of $\underline{x}_i$), the convergence is rapid (even for $n = 3$) but if it is asymmetric (right panel, with exponential distribution of $\underline{x}_i$) a value higher than 32 (the highest n shown in the plot) is needed for a satisfactory convergence. In case of dependent $\underline{x}_i$ with positive correlation, the convergence is slower and a much larger n is needed.

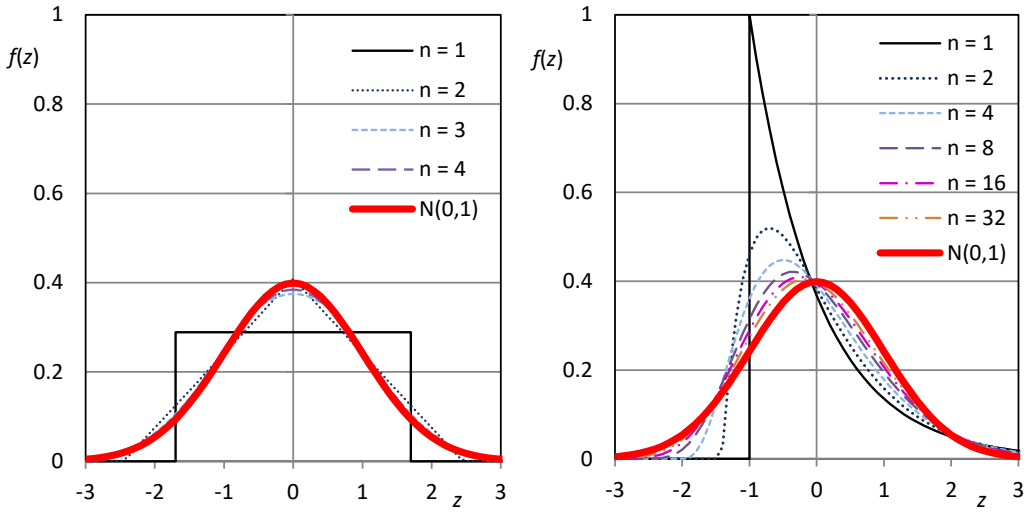

Figure 2.7 Convergence of the sum of independent identically distributed stochastic variables to the normal distribution (thick line). The thin continuous lines represent the probability density of the constituent variables $\underline{x}_i$, which have mean 0 and standard deviation 1. In the **left** panel the density is uniform on the interval $\left(-\sqrt{3}, \sqrt{3}\right)$ with $f(x) = \frac{1}{2\sqrt{3}}$ and in the **right** panel exponential, $f(x) = \mathrm{e}^{-x-1}, x \geq -1$ (the parameters are chosen so as to have mean 0 and standard deviation 1). The dotted lines represent the densities of the sums $\underline{z}_n := \left(\underline{x}_1 + \cdots + \underline{x}_n\right)/\sqrt{n}$ for the values of n indicated in the legend. (The division of the sum by $\sqrt{n}$ improves the presentation of the curves, as all $\underline{z}_n$ have the same mean and variance, 0 and 1, respectively, and does not affect the essentials of the central limit theorem.) Equations (2.53) and (2.138) were used to produce the graph.

The conditions for the validity of the central limit theorem are general enough and are met in many practical situations. Conditions with particular interest are the following (e.g. Papoulis, 1990, p. 215):

a. the variables $\underline{x}_i$ are independent identically distributed with finite third moment;
b. the variables $\underline{x}_i$ are bounded from above and below with nonzero variance;
c. the variables $\underline{x}_i$ are independent with finite third moment and the variance of $\underline{z}_n$ tends to infinity as n tends to infinity.

The theorem has been extended for variables $\underline{x}_i$ that are interdependent, but each one effectively dependent on a finite number of other variables. Gnedenko and Kolmogorov (1949) proved an extended version of the theorem, according to which the sum of n stochastic variables with heavy tailed distributions with upper-tail index $\xi > 1/2$, therefore having infinite variance, will tend to the so-called Lévy alpha-stable distribution, as $n \to \infty$. If $\xi < 1/2$, the standard central limit theorem holds, i.e., the sum converges to the Gaussian distribution, which is a special case of the Lévy alpha-stable distribution. In geophysics, the standard theorem suffices because we can justifiably

assume that those processes have finite variance: an infinite variance would presuppose an infinite quantity of energy to materialize, which is absurd.

Most geophysical processes (e.g., rainfall and streamflow) have skewed distributions at fine time scales, and therefore the normal distribution is not a suitable model at these scales. However, the normal distribution describes with satisfactory accuracy those variables which refer to longer time scales such as the annual. Thus, the annual rainfall depth in an area with wet climate is the sum of many (e.g., 50-100) rainfall events during the year. However, this is not valid for rainfall in dry areas. Likewise, the annual runoff volume passing through a river section can be regarded as the sum of 365 daily volumes. These are not independent, but the central limit theorem can be applicable again.

Nonetheless, it should be stressed that the convergence to the normal distribution concerns the body of the distribution. For example, what is depicted in Figure 2.7 is about the body of the distribution. What happens to the upper tail behaviour, i.e., to the extremes? Apparently, once the theoretical conditions of validity are satisfied, the theoretical result should hold true. However, this may not be of any help in practice as the convergence in the tail is much slower. Figure 2.8 (left) shows that the convergence in the tail is indeed slow for the exponential distribution, much slower than that of the body of the same distribution shown in Figure 2.7 (right). The coefficient of skewness for the sum of 32 $\underline{x}_i$ is rather small (0.35) indicating a rather satisfactory approximation by the normal distribution. However, Figure 2.8 (left) shows that for $\bar{F} = 0.001$ the distribution of the sum of 32 $\underline{x}_i$ is by an order of magnitude larger than that of the normal distribution.

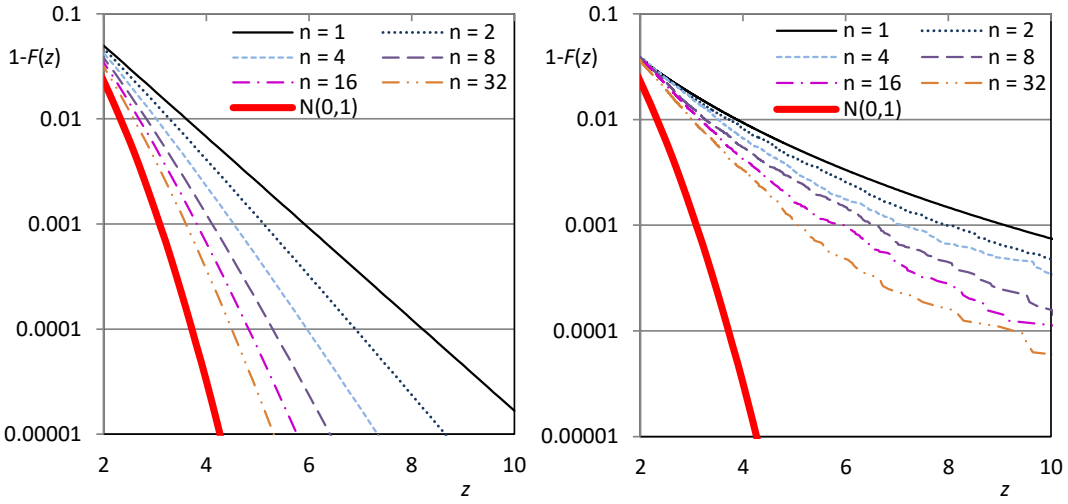

Figure 2.8 Convergence of the sum of independent identically distributed stochastic variables to the normal distribution (thick line) with focus on the upper tail. The thin continuous lines represent the exceedance probability of the constituent variables $\underline{x}_i$, which have mean 0 and standard deviation 1. In the **left** panel the distribution is exponential with density $f(x) = \mathrm{e}^{-x-1}, x > -1$ as in the right panel of Figure 2.7) and in the **right** panel Pareto with upper-tail index 1/4 and exceedance probability $\bar{F}(x) = \left(4/3 + \frac{x\sqrt{2}}{3}\right)^{-4}, x \geq -1/\sqrt{2}$ (the parameters are chosen so that the mean is 0 and standard deviation 1). The dotted lines represent the exceedance probability of the sums $\underline{z}_n := (\underline{x}_1 + \cdots + \underline{x}_n)/\sqrt{n}$ for the values of n indicated in the legend. The left graph was produced from equations (2.53) and (2.138), and the right graph by a numerical (Monte Carlo) method (as analytical integration is impossible beyond $n = 2$), and this is the reason for the curly shape of the curves.

For heavy tailed distributions, there are differences of several orders of magnitude as shown for the Pareto distribution in Figure 2.8 (right). The upper-tail index of this Pareto distribution is 1/4, which means that the moments below the fourth order exist and therefore the necessary conditions for the central limit theorem are satisfied. Despite that, the approximation of the distribution upper tail is unsatisfactory. Actually, it can be easily understood that, as the moments for order ≥ 4 of $\underline{x}_i$ are infinite, the same will hold for the sum of any finite number of $\underline{x}_i$, while all the moments of the limiting normal distribution are finite. This conflict, along with the fact that the behaviour of extremes is closely connected to the high order moments of a distribution (see Koutsoyiannis, 2025) suggests that we must be very attentive in the application of the theorem to geophysical processes, particularly because these processes seem to exhibit heavy tails and long-range dependence.

2.19 Limiting extreme value distributions

By analogy with the central limit theorem referring to the sum or the average of many variables, limiting distributions may also arise, as $n \to \infty$, for the maximum of these variables, $\underline{y}_n \coloneqq \max(\underline{x}_1, \dots, \underline{x}_n)$, whose exact distribution function for independent and identically distributed $\underline{x}_n$ is:

$$F_{\underline{y}_n}(y) = \left(F_{\underline{x}}(y)\right)^n \tag{2.146}$$

The relevant theory was developed in the 20th century. Fréchet (1927) was the first to identify one of the asymptotic distributions of maxima, which bears his name. Fisher and Tippett (1928) showed that there are only three possible limiting distributions for extremes, while von Mises (1936) identified sufficient conditions for convergence to the three limiting laws. Gnedenko (1943) established solid foundations for the asymptotic theory of extremes providing the precise conditions for the weak convergence to the limiting laws. In this respect, it is worth noting the celebrated book by Gumbel (1958), who was one of the pioneers promoting and applying the formal theory to geophysical and engineering practice. The theory is concisely presented in a review paper by Davison and Huser (2015). Assuming that $\underline{x}_i$ are independent and identically distributed, there exist a real number ξ and sequences of real numbers $\lambda_n > 0$ and ε_n such that the rescaled maximum $\underline{y}_n^* \coloneqq \max(\underline{x}_1, \dots, \underline{x}_n)/\lambda_n - \varepsilon_n$ has limiting distribution, as $n \to \infty$:

$$H(y) \coloneqq F_{\underline{y}_\infty^*}(y) = \exp\left(-\left(1 + \xi\left(\frac{y}{\lambda} - \varepsilon\right)\right)^{-1/\xi}\right), \qquad \xi y \geq \xi\lambda\left(\varepsilon - \frac{1}{\xi}\right) \tag{2.147}$$

Here $\lambda > 0$ is a scale parameter, ε is a dimensionless location parameter and ξ is a shape parameter, identical to the upper-tail index.

The parameter ξ has a unique value, which is precisely the same with the upper-tail index of the parent distribution, but the parameters λ and ε are not unique. They can be chosen as per convenience (different choices will lead to appropriate modification of the sequences λ_n and ε_n). A natural choice is $\varepsilon = 0, \lambda = 1$. A more customary option is to choose

a large n for which convergence has been achieved to a satisfactory degree, for that n set $\lambda_n = 1$ and $\varepsilon_n = 1$ (so that $\underline{y}_n^* \equiv \underline{y}_n = \max(\underline{x}_1, \dots, \underline{x}_n)$ without any rescaling) and calculate λ and ε from equation (2.147). With this aim (and given that, for finite n, (2.147) is an approximation and not an exact relationship) we choose two points x_1 and x_2 and equate $F(x)^n$ with $H(x)$ at these points. For mathematical convenience we can choose the two points so that $-x_1/\lambda + \varepsilon = 0, -x_2/\lambda + \varepsilon = -1$, or $x_1 = \lambda\varepsilon, x_2 = \lambda\varepsilon + \lambda$. Hence, $F(\lambda\varepsilon)^n = \mathrm{e}^{-1}, F(\lambda\varepsilon + \lambda)^n = \mathrm{e}^{-(1+\xi)^{-\frac{1}{\xi}}}$. Solving for λ and ε we find:

$$\lambda = F^{-1}\left(\mathrm{e}^{-\frac{1}{(1+\xi)^{\frac{1}{\xi}}n}}\right) - F^{-1}\left(\mathrm{e}^{-\frac{1}{n}}\right), \qquad \varepsilon = \frac{F^{-1}\left(\mathrm{e}^{-\frac{1}{n}}\right)}{\lambda} \tag{2.148}$$

where for $\xi \to 0$, $(1+\xi)^{-1/\xi} \to 1/\mathrm{e}$. This is usually done unconsciously, for example when we study annual maxima of daily values and fit $H(y)$ of equation (2.147) on the annual maxima directly, without even deriving it from $F_{\underline{x}}(x)$.

Depending on the value of ξ, the limiting distribution in equation (2.147), known as the generalized extreme value (GEV) distribution, is a compact expression including three cases with different behaviours:

- For $\xi = 0$, GEV takes the following form, known as the Gumbel distribution or extreme value type I (EV1) distribution:

$$H(y) = \exp(-\exp(-y/\lambda + \varepsilon)) \tag{2.149}$$

  This is a light-tailed distribution without an upper or lower bound.
- For $\xi > 0$, the distribution is known as Fréchet or extreme value type II (EV2) and has a lower bound at $\lambda\psi - \lambda/\xi$. This is a heavy-tailed distribution with upper-tail index ξ.
- In case that $\xi < 0$ the distribution is known as the reverse Weibull or the extreme value type III (EV3) distribution. This has an upper bound for y at $\lambda\psi - \lambda/\xi$.

The GEV has the property to be max-stable, meaning that maxima from this distribution, after linear transformation, have the same distribution. More formally, Fréchet's necessary condition for max-stability is this: For any $n \in \mathbb{N}$ and $y \in \mathbb{R}$, there exist real numbers $a_n > 0$ and b_n such that:

$$\big(H(a_n y + b_n)\big)^n = H(y) \tag{2.150}$$

In fact, GEV is the only distribution satisfying this condition.

A specific parent distribution $F(x)$ belongs to the so-called (*max*-)*domain of attraction* of one of the three limiting laws, in the sense that the distribution of rescaled maxima from this parent distribution is this particular limiting law. Formal mathematical conditions determining a parent distribution's domain of attraction were formulated by von Mises (1936) and Gnedenko (1943). The practical result is that heavy-tailed distributions with upper-tail index $\xi > 0$ (e.g., Pareto, Pareto-Burr-Feller, Student and its extensions, generalized log-gamma and generalized beta prime) belong to the domain of

attraction of EV2. Light-tailed distributions (e.g. exponential, gamma, Weibull, normal and their generalizations) as well as heavy-tailed distributions with upper-tail index $\xi = 0$ (e.g. lognormal) belong to the domain of attraction of EV1. In the domain of attraction of EV3 belong distributions bounded from above (e.g. uniform).

Because of its upper bound, EV3 is not an appropriate model for geophysical extremes, for Nature has no upper limits (unless dictated by a conservation law). The values of ξ which we expect to see in geophysical processes are in the range $(0, 1/2)$ so that the variance be finite, as already discussed in section 2.18. The exact value of the upper-tail index is important to specify in engineering design. The major question in this regard is how the value of an extreme quantity y grows as the probability of exceedance $\overline{H}(y)$ decreases tending to zero. To put the question the other way, at which rate does y tend to infinity as the probability of exceedance tends to zero? The Gumbel distribution represents the mathematically proven lower limit to the rate of this growth. The alternative is the Fréchet law which represents a higher rate of growth. The two options may lead to substantial differences in design quantities for high return periods. As already discussed, the Fréchet law which has a positive upper-tail index is the more realistic option.

When we are interested in minima, we can follow the same procedure observing that $\underline{z}_n \coloneqq \min(\underline{x}_1, \dots, \underline{x}_n) = -\max(-\underline{x}_1, \dots, -\underline{x}_n)$. Consequently, we have $P\{\underline{z}_n \le z\} = 1 - P\{\max(-\underline{x}_1, \dots, -\underline{x}_n) \le -z\}$ and hence the limiting distribution is

$$G(z) \coloneqq F_{\underline{z}^*_\infty}(y) = 1 - \exp\left(-\left(1 + \xi\left(-\frac{z}{\lambda} - \varepsilon\right)\right)^{-1/\xi}\right), \qquad \xi z \le \xi\lambda\left(\frac{1}{\xi} - \varepsilon\right) \tag{2.151}$$

Again, we have three cases: (a) $\xi = 0$, corresponding to the Gumbel (EV1) distribution of minima, i.e.,

$$G(z) = 1 - \exp(-\exp(z/\lambda + \varepsilon)) \tag{2.152}$$

(b) $\xi > 0$, corresponding to the reverse Fréchet distribution, which has upper bound $\lambda(1/\xi - \varepsilon)$ and a heavy lower tail, and (c), $\xi < 0$, corresponding to the Weibull distribution, which has lower bound $\lambda(1/\xi - \varepsilon)$ and a light upper tail.

While most of the above mathematical developments have assumed independent stochastic variables, the results can be approximately valid even in the case of variables dependent in time. Specifically, Leadbetter (1983) demonstrated that, under mild conditions, maxima of dependent series follow the same form of distributional limit laws as those of independent series. However, the dependence changes the location and scale parameters (Davison and Huser, 2015) in such a manner as if $H(y)$ was replaced by $\big(H(y)\big)^\theta$, where $\theta \in (0,1]$ is the so-called *extremal index*. It can be seen that this replacement is equivalent to a change of the parameters λ and ε, while ξ remains the same. Also, the rate of convergence to the limit becomes slower in the case of dependence. Phenomenologically, dependence in time of a process causes clustering or grouping of extreme events. Unfortunately dependence in time is quite often misinterpreted as

nonstationarity. This may explain the recent increase in the number of publications detecting nonstationarity in extremes (cf. Koutsoyiannis and Montanari, 2015).

Here it should be stressed that, if compared to the central limit theorem, which is characterized by a fast convergence to the limit (except in the extremes, as seen in Figure 2.8), the convergence to the max-stable distribution may be much slower in certain cases. The rate of convergence to the limit of distributions belonging to the domain of attraction of EV2 is generally satisfactory. However, for those belonging to the domain of attraction of EV1, such as the normal and lognormal distributions, the rate is desperately slow. The meaning of a slow convergence in real-world applications, where n is finite and often small, is that the approximation of EV1 to the actual distribution of maxima is not satisfactory. Thus, it may be preferable to approximate the actual distribution of maxima of variables with distributions belonging to the domain of attraction of EV1 by the EV2 distribution, as illustrated in Digression 2.O.

Digression 2.O: How well do limiting distributions approximate exact distributions?

For independent identically distributed variables, the exact distribution of maxima is $F(x)^n$ (equation (2.146)). To approximate the exact distribution by the GEV we use equation (2.148). As an example, for the maxima from the standard normal distribution, F_{N}, approximated by the EV1 we get:

$$\lambda = F_{\mathrm{N}}^{-1}\left(\mathrm{e}^{-\frac{1}{\mathrm{e}n}}\right) - F_{\mathrm{N}}^{-1}\left(\mathrm{e}^{-\frac{1}{n}}\right), \qquad \varepsilon = \frac{1}{\lambda} F_{\mathrm{N}}^{-1}\left(\mathrm{e}^{-\frac{1}{n}}\right)$$

As a second example, for the Pareto distribution, $F(x) = 1 - (1+x)^{-1/\xi}$, approximated by the EV2 we get:

$$\lambda = \left(1 - \mathrm{e}^{-\frac{1}{n(1+\xi)^{1/\xi}}}\right)^{-\xi} - \left(1 - \mathrm{e}^{-\frac{1}{n}}\right)^{-\xi}, \qquad \varepsilon = \frac{1}{\lambda}\left(\left(1 - \mathrm{e}^{-\frac{1}{n}}\right)^{-\xi} - 1\right)$$

We have applied this approximation for $n = 10$, 100 and 1000 for the normal and the Pareto distributions which belong to the domain of attraction of EV1 and EV2, respectively. The results are shown graphically in Figure 2.9, along with the case $n = 1$, i.e., the parent distribution per se for the sake of comparison.

The results are very good for the Pareto distribution and very bad for the normal distribution. Even for $n = 1000$, the EV1 severely overestimates the actual probability of exceedance. One may think of using the EV3 instead of EV1 for the approximation of the normal distribution. However, this is not advisable because the EV3, despite giving a better approximation, entails an upper bound to extremes which distorts a fundamental characteristic of the modelled phenomenon.

Likewise, Figure 2.10 provides similar information for the lognormal distribution with mean and standard deviation of $\ln x$ equal to 0 and 1, respectively (denoted LN(0,1)). Like the normal, it belongs to the domain of attraction of EV1. Here the approximation is even worse than in the normal case but now the EV1 underestimates the exact probability of exceedance. For that reason, we could use EV2 as a better approximation (without having the problem of artificially inducing an upper bound). As seen in the right panel of Figure 2.10 this latter approximation is quite satisfactory.

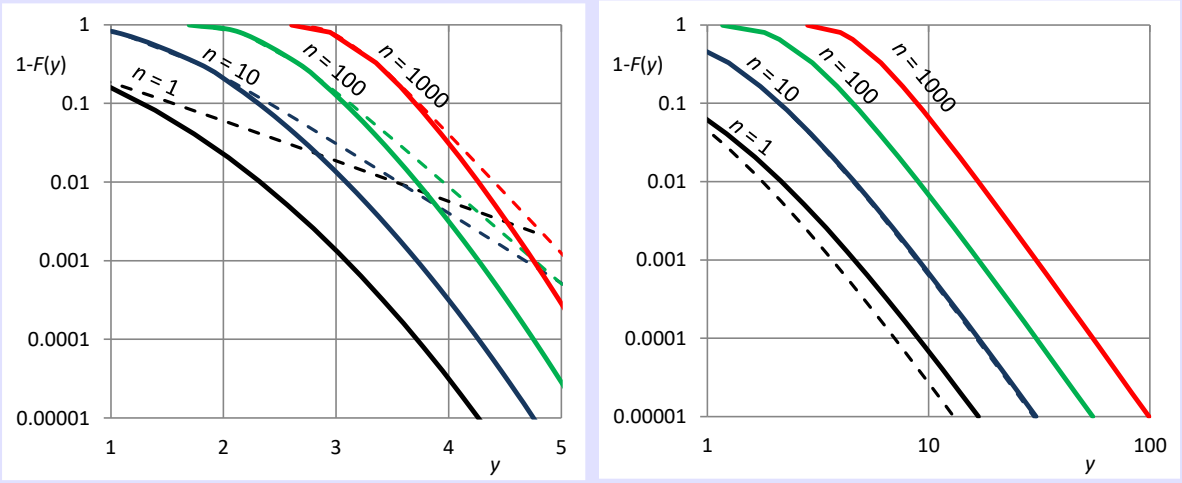

Figure 2.9 Approximation of the true distribution of the maximum of n independent identically distributed variables (continuous lines) by the limiting extreme value distribution (dashed lines). (**left)** The parent distribution is the standard normal, N(0,1), and the approximating distribution is the EV1. **(right)** The parent distribution is the Pareto, $F(x) = 1 - (1 + x)^{-1/\xi}$, with $\xi = 0.25$ and the approximating distribution is the EV2 with the same ξ.

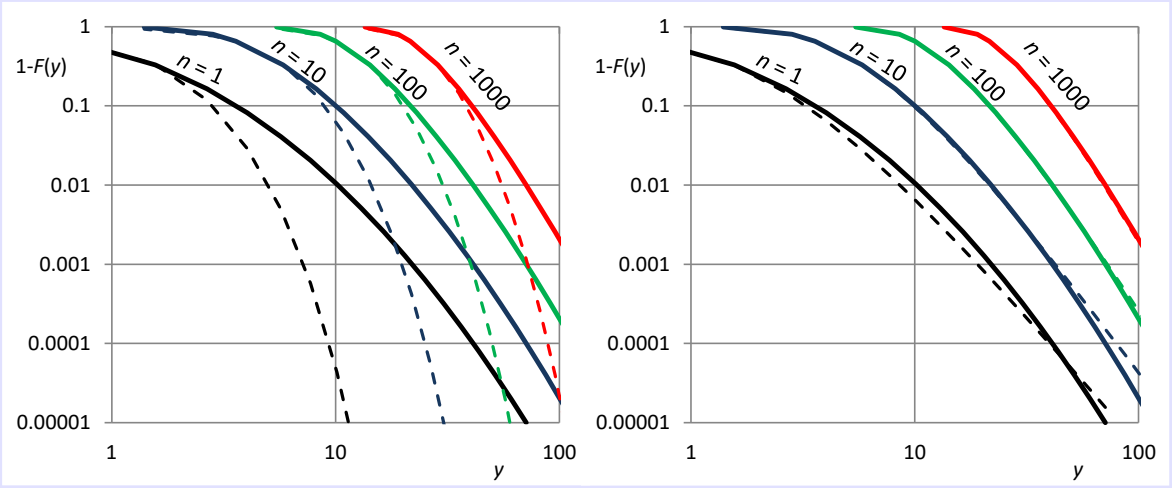

Figure 2.10 Approximation of the true distribution of the maximum of n independent variables with lognormal distribution LN(0,1) (continuous lines) by the limiting extreme value distribution (dashed lines), which is **(left)** EV1 and **(right)** EV2 with $\xi = 0.3/n^{0.07}$, which was found after a numerical investigation and fitting a power function of n by minimizing the fitting error.

2.20 Relationship of parent and extreme value distribution

Because of problems originating from the slow rate of convergence of the actual distribution to GEV (particularly EV1), it may be a good idea not to use the limiting distributions in practical applications but, instead, to model the tails of the parent distribution or even the entire parent distribution. Yet the theory of max-stable distributions retains its usefulness for inferring the tail behaviour of the parent distribution.

Specifically, the upper tail behaviour of the parent distribution is described by the conditional distribution function:

$$F\left(x\middle|\underline{x}>u\right)=\ P\{\underline{x}\le x|\underline{x}>u\}=\frac{F(x)-F(u)}{1-F(u)} \tag{2.153}$$

for a value of the threshold u that is sufficiently large. Now, assuming that the parameterization of $H(x)$ with regard to λ and ε has been made with reference to a specific large n, as described for the derivation of (2.148), we choose u so that the exceedance probability $1 - F(u)$ equals $1/n$. (This is a very common choice as will be discussed in Digression 2.P). Thus, $F\left(x\middle|\underline{x}>u\right)=n(F(x)-1)+1$ or:

$$1-F\left(x\middle|\underline{x}>u\right)=n\left(1-F(x)\right) \tag{2.154}$$

On the other hand, we can write for $F(x)$ approaching 1,

$$-\ln H(x)\approx-\ln\left(F(x)\right)^n=-n\ln F(x)\approx n\left(1-F(x)\right) \tag{2.155}$$

because $\ln F(x)=\ln(1-\left(1-F(x)\right)=-\left(1-F(x)\right)-\left(1-F(x)\right)^2-\cdots$ and for $F(x)$ approaching 1 we can keep the first term only. Hence, combining (2.154) and (2.155) we find:

$$F\left(x\middle|\underline{x}>\lambda\varepsilon\right)=1+\ln H(x)=1-\left(1+\xi\left(\frac{x}{\lambda}-\varepsilon\right)\right)^{-1/\xi},\qquad x\ge\lambda\varepsilon \tag{2.156}$$

where we equated u with $\lambda\varepsilon$ for consistency (i.e. to make $F\left(u\middle|\underline{x}>u\right)=0$). This is the Pareto distribution for $\xi > 0$ while for $\xi = 0$ we get the exponential form:

$$F\left(x\middle|\underline{x}>\lambda\varepsilon\right)=1+\ln H(x)=1-\exp\left(\frac{x}{\lambda}-\varepsilon\right),\qquad x\ge\lambda\varepsilon \tag{2.157}$$

Furthermore, for values of x large enough to make $H(x)$ approach 1, we can use the same logic to get $\ln H(x)\approx H(x)-1$ and hence

$$F\left(x\middle|\underline{x}>\lambda\varepsilon\right)\approx H(x) \tag{2.158}$$

This approximation error does not exceed ~1% for $H(x) > 0.99$ and ~5% for $H(x) > 0.9$.

We can generalize the above analysis for different values of the threshold u. In this case the resulting functional form remains the same, with the same upper-tail index, but the location and scale parameters differ, i.e. (Davison and Huser, 2015):

$$F\left(x|\underline{x}>u\right)=1-\left(1+\xi\left(\frac{x}{\lambda_u}-\varepsilon_u\right)\right)^{-\frac{1}{\xi}} \tag{2.159}$$

where

$$\lambda_u\coloneqq\lambda\left(1+\xi(u/\lambda-\varepsilon)\right),\qquad\varepsilon_u\coloneqq\frac{u}{\lambda_u} \tag{2.160}$$

It is readily confirmed that if we set $u=\lambda\varepsilon$ in (2.159) and (2.160) we recover (2.156). However, this equation is valid only for large values of u (unless the unconditional $F(x)$ is Pareto, in which case it is valid for any u).

A final note that may be relevant in some analyses is this. If the value of n in $\underline{y}_n \coloneqq \max(\underline{x}_1, \dots, \underline{x}_n)$, is not constant but a stochastic variable with Poisson distribution with mean ν, while $\underline{x}_i$ are independent, then the conditional distribution of $\underline{y}_n$ on specified n remains $F_{\underline{y}_n}(y|n) = \left(F_{\underline{x}}(y)\right)^n$ but for unspecified n the unconditional distribution becomes (Todorovic and Zelenhasic, 1970):

$$F_{\underline{y}}(y) = \exp\left(-\nu\left(1 - F(y)\right)\right) \tag{2.161}$$

This resembles (2.155) with the difference that it is now exact rather than approximate.

As already discussed above and in section 2.19, due to the problems of the limiting extreme value distributions, it is preferable to focus the studies of extremes on the parent distributions and primarily their upper tails. From the above theoretical discussion, we have reasons to expect a parent distribution upper tail of Pareto type, or at least exponential, but each time this should be verified by observations. Nowadays, there is an abundance of geophysical data on daily and subdaily scales and there is no need to extract annual or seasonal maxima. Instead, we should use the entire observational record or at least the values over some threshold. If the available observations are originally given in terms of time-block (e.g., annual) maxima, it may seem pertinent to refer to extreme value distribution. However, again it is possible to model the parent process, estimating its parameter from time-block maxima (see Koutsoyiannis, 2025; section 6.22). Advantages of studying the distribution of the parent variable rather than the distribution of maxima are discussed in Digression 2.P

Digression 2.P: Block maxima vs. values over threshold vs. complete record

Traditionally, geophysical records have been analysed in either of two ways. The most frequent is to choose the highest of all recorded values for a given time period or "block" (typically one year) and form a statistical sample (commonly referred to as "block maxima") with size equal to the number of blocks (typically years) of the record. The other is to form a sample of values over a threshold (here abbreviated as VOT but sometimes referred to as "peaks-over-threshold"—POT) with all recorded values over a certain threshold irrespective of the time they occurred. Usually, the threshold is chosen high enough, so that the sample size is again equal to the number of years of the record. This however is not necessary: it can well be set equal to zero, so that all recorded values be included in the sample (the complete sample). However, a high threshold simplifies the study and helps focus attention on the distribution upper tail. In addition, this choice simplifies the mathematical expression (compare equations (2.156) and (2.159)), leading to the identity of the parameters of the distributions of block-maxima and values over threshold.

Furthermore, studying the complete series of observations has the advantage of utilizing the entire information that is available (Volpi et al., 2019). Indeed, extracting maxima over some period results in waste of information because other extreme observations should also be informative about extremes. Such information (e.g., the second-largest value of a period, which can be higher than another period's largest value) is retained even if we use the values over threshold instead of the entire series of observations.

Furthermore, design quantities should naturally correspond to the parent distribution, rather than the artificially induced maxima over an arbitrarily defined time period. This favours the use of the parent distribution. As we have seen (equation (2.158)) the two are almost equal for large design values, but for lower ones there are differences. Thus, even if our analyses are based on time-block extremes ($H(y)$), the results should eventually be converted to the parent

distribution ($F(x)$) before they are used for design. The above considerations provide the necessary mathematical support for such conversion.

The most important reason favouring the study of the complete record over that of block maxima and VOT is that only the former provides faithful information about time dependence of the underlying process. Such dependence may be marked and possibly of long-range type. Neglecting dependence results in underestimation of extremes. On the other hand, the procedure of extracting block maxima leads to severe distortion of the dependence structure (Iliopoulou and Koutsoyiannis, 2019; Koutsoyiannis, 2025), whereas the concept of taking values over threshold relies on a tacit assumption of time independence, which may be inappropriate, particularly for the streamflow process (Lombardo et al., 2019).

Appendix 2-I: Proof of theorems for the entropy definition

Proof of Theorem 1

Step 1.1 For didactic purposes, in this step we examine the simplest case of a bipartition. Let $U \subset \mathbb{R}^2$ containing the line $x + y = 1$, and a function $\Phi \in C^2(U)$ with properties of symmetry and non-interaction:

$$\Phi(x, y) = \Phi(y, x), \qquad \frac{\partial^2 \Phi}{\partial x\, \partial y} \equiv 0 \tag{2.162}$$

Non-interaction implies that $\partial_x \Phi(x, y)$ depends only on x: $\partial_x \Phi(x, y) = g'(x)$ for some function $g(x)$. We integrate in x, with an arbitrary function of y, $H(y)$, as constant of integration, to get $\Phi(x, y) = g(x) + H(y)$. By virtue of symmetry, setting $x = x_0$ we may write $g(x_0) + H(y) = g(y) + H(x_0)$ which yields $H(y) = g(y) + c$, where $c \coloneqq H(x_0) - g(x_0)$. Hence

$$\Phi(x, y) = \varphi(x) + \varphi(y) \tag{2.163}$$

where $\varphi(x) = g(x) + c/2$.

Step 1.2 We will prove the general case, $\Phi(x_1, \dots, x_n) = \sum_{i=1}^{n} \varphi(x_i)$, where $\Phi \in C^2(U)$, $U \subset \mathbb{R}^n$ containing the hyperplane $x_1 + \cdots + x_n = 1$. For each $j \neq n$, $\partial^2 F / \partial x_j\, \partial x_n = 0$, and so $\partial_{x_n} F$ does not depend on any x_j. Therefore, we can write $\partial_{x_n} \Phi(x_1, \dots, x_n) = g_n'(x_n)$. Integrating in x_n, with an arbitrary function of the remaining variables, $H_{n-1}(x_1, \dots, x_{n-1})$, as constant of integration, we get

$$\Phi(x_1, \dots, x_n) = g_n(x_n) + H_{n-1}(x_1, \dots, x_{n-1}) \tag{2.164}$$

Taking joint partial derivatives in the above for any $1 \leq i < j \leq n - 1$, we find

$$\frac{\partial^2 H_{n-1}}{\partial x_i\, \partial x_j} = \frac{\partial^2 \Phi}{\partial x_i\, \partial x_j} = 0 \tag{2.165}$$

Thus H_{n-1} satisfies the same vanishing mixed-partial postulate in dimension $n - 1$ and, in a similar manner, we can write

$$H_{n-1}(x_1, \dots, x_{n-1}) = g_{n-1}(x_{n-1}) + H_{n-2}(x_1, \dots, x_{n-2}) \tag{2.166}$$

Proceeding inductively, we find

$$\Phi(x_1, \dots, x_n) = g_n(x_n) + g_{n-1}(x_{n-1}) + \cdots + g_1(x_1) \tag{2.167}$$

where $g_1(x_1) = H_1(x_1)$. Enrolling symmetry, we conclude that $g_n(x) = g_{n-1}(x) = \cdots = g_1(x) = \varphi(x)$. Setting $x_i = P(A_i)$ we conclude in Theorem 1.

Proof of Theorem 2

Step 2.1 We start with the observation that, since Postulate 4 holds for any pair of partitions $\mathbb{A}, \mathbb{A}'$, it also holds in the simplest nontrivial case of bipartitions. Let $\mathbb{A} = \{A, \overline{A}\}$ be a bipartition of Ω, with probabilities $P(A) =: x$ and $P(\overline{A}) = 1 - x$. The entropy is

$$\Phi[\mathbb{A}] = \varphi(x) + \varphi(1 - x) \tag{2.168}$$

We consider two identical ground sets Ω, Ω', and two identical partitions thereof, $\mathbb{A} = \{A, \overline{A}\}, \mathbb{A}' = \{A', \overline{A}'\}$ with $A' = A$. The product partition is $\mathbb{A} \otimes \mathbb{A}' = \{AA', A\overline{A}', \overline{A}A', \overline{AA}'\}$ and constitutes a partition of the set $\Omega \times \Omega'$. Let $y = P(AA')$. Since $P(AA') + P(A\overline{A}') = P(A) = x, P(AA') + P(\overline{A}A') = P(A') = x$, we have $P(A\overline{A}') = P(\overline{A}A') = x - y$. Furthermore, since $P(\overline{A}A') + P(\overline{AA}') = P(\overline{A}) = 1 - x$, we have $P(\overline{AA}') = 1 - x - (x - y) = 1 - 2x + y$. Hence the entropy of the product partition is

$$\Phi(\mathbb{A} \otimes \mathbb{A}') = \varphi(y) + 2\varphi(x - y) + \varphi(1 - 2x + y) \tag{2.169}$$

Step 2.2 The probability y which extremizes $\Phi(\mathbb{A} \otimes \mathbb{A}')$ has derivative zero, i.e.

$$\frac{\partial \Phi(\mathbb{A} \otimes \mathbb{A}')}{\partial y} = \varphi'(y) - 2\varphi'(x - y) + \varphi'(1 - 2x + y) = 0 \tag{2.170}$$

Step 2.3 By Postulate 4, for independent A, A', $y = P(AA') = P(A)P(A') = x^2$. Hence:

$$\varphi'(x^2) - 2\varphi'(x - x^2) + \varphi'((1 - x)^2) = 0 \tag{2.171}$$

We note here that independence of A, A' implies independence of all other pairs. Specifically, we have

$$\begin{aligned} &P(A\overline{A}') = P(\overline{A}A') = x - x^2 = x(1 - x) = P(A)P(\overline{A}') = P(\overline{A})P(A'), \\ &P(\overline{AA}') = 1 - 2x + x^2 = (1 - x)^2 = P(\overline{A})P(\overline{A}') \end{aligned} \tag{2.172}$$

Step 2.4 The functional equation (2.171) should be valid for any x in [0,1]. Consequently, a solution to this equation will also be a solution to equation

$$\varphi'(x^2) - 2\varphi'(x) + \varphi'(1) = 0 \tag{2.173}$$

The latter is obtained by taking a very small x in (2.171), so that we can omit all but the least order polynomial terms, and is easier to handle (for an analogy, cf. Apostol, 1967, chapter 6). By setting $x = \mathrm{e}^{-s}$, $g(s) = f'(e^{-s}) - \varphi'(1)$, we get the linearly homogeneous functional equation

$$g(2s) = 2g(s) \tag{2.174}$$

whose solution is

$$g(s) = k\, s \tag{2.175}$$

where k is any real. Pathological discontinuous solutions of the equation are excluded by virtue of Postulate 1. By inverting the transformation, we get

$$\varphi'(x) = \varphi'(1) - k \ln x\,, \qquad \varphi(x) = \int_0^x \varphi'(q)\mathrm{d}q = x\,(\varphi'(1) + k - k \ln x) \tag{2.176}$$

From Postulate 2, we have $\Phi(\Omega) = \varphi(1) = 1(\varphi'(1) + k) = 0$, so that $\varphi'(1) = -k$ and finally:

$$\varphi'(x) = -k - k\ln x\,, \qquad \varphi(x) = -k\,x\ln x \tag{2.177}$$

Step 2.5 We verify that the thus determined $\varphi(x)$ satisfies the original functional equation in step 2.3. We have

$$\begin{aligned}\varphi'(x^2) - 2\varphi'(x - x^2) + \varphi'((1-x)^2) \\ = (-k - k\ln x^2) + (2k + 2k\ln(x - x^2)) + (-k - k\ln((1-x)^2)) \\ = k\ln\frac{(x-x^2)^2}{x^2(1-x)^2} = k\ln 1 = 0\end{aligned} \tag{2.178}$$

which validates the solution.

Step 2.6 By virtue of Theorem 1, the entropy of $\mathbb{A}$ is

$$\Phi(\mathbb{A}) = \varphi(x) + \varphi(1-x) = -k\,x\ln x - k\,(1-x)\ln(1-x) \tag{2.179}$$

Since (by Postulate 1) $\Phi(\mathbb{A}) \geq 0$, we must set $k > 0$. We check whether the extremum $\Phi(\mathbb{A})$ is maximum or minimum. The second derivative is $\varphi''(x) = -k/\big(x(1-x)\big) < 0$, which means that we have a maximum.

Step 2.7 We consider two arbitrary partitions $\mathbb{A}, \mathbb{A}'$ of Ω, Ω' with n, n' elements, respectively, and we will show that that the uniquely defined function $\varphi(x) = -k\,x\ln x$ for bipartitions satisfies Postulate 4 for any $\mathbb{A}, \mathbb{A}'$. Specifically, we will show that entropy maximization with this $\varphi(x)$ results in independent partitions. If $x_i = P(A_i), x'_j = P\big(A'_j\big), y_{ij} = P(A_i A'_j)$ then

$$x_i = \sum_{j=1}^{n'} y_{ij}\,, \qquad x'_j = \sum_{i=1}^{n} y_{ij} \tag{2.180}$$

The total entropy is

$$\Phi(\mathbb{A}\otimes\mathbb{A}') = -k\sum_{i=1}^{n}\sum_{j=1}^{n'} x_{ij}\ln x_{ij} \tag{2.181}$$

The functional that maximizes $\Phi(\mathbb{A}\otimes\mathbb{A}')$ with constrained probabilities as above and Lagrange multipliers λ_i, λ'_j is

$$L = -k\sum_{i=1}^{n}\sum_{j=1}^{n'} y_{ij}\ln y_{ij} - \sum_{i=1}^{n}\lambda_i\left(x_i - \sum_{j=1}^{n'} y_{ij}\right) - \sum_{j=1}^{n'}\lambda'_j\left(x_j - \sum_{i=1}^{n} y_{ij}\right) \tag{2.182}$$

and has derivatives

$$\frac{\partial L}{\partial y_{ij}} = k\big(-1 - \ln y_{ij}\big) - \lambda_i - \lambda'_j = 0, \quad \frac{\partial L}{\partial \lambda_i} = -x_i + \sum_{j=1}^{n'} y_{ij} = 0, \quad \frac{\partial L}{\partial \lambda_i} = -x_j + \sum_{i=1}^{n} y_{ij} = 0 \tag{2.183}$$

Hence

$$y_{ij} = \exp\left(-\frac{\lambda_i}{k}\right)\exp\left(-\frac{\lambda'_j}{k}\right)\exp(-1) = \exp\left(-\frac{\lambda_i}{k} - \frac{1}{2}\right)\exp\left(-\frac{\lambda'_j}{k} - \frac{1}{2}\right) \tag{2.184}$$

Setting

$$\lambda_i = -k\left(\ln x_i + \frac{1}{2}\right), \qquad \lambda'_j = -k\left(\ln x'_j + \frac{1}{2}\right) \tag{2.185}$$

we find that

$$y_{ij} = x_i x'_j \tag{2.186}$$

which means that the partitions $\mathbb{A}, \mathbb{A}'$ are independent. It is easy to see that the constraints are satisfied with these values of λ_i, λ'_j and hence the values set for λ_i, λ'_j are the correct ones.

Step 2.8 The constant k can be any positive real, but by convention (and for convenience) we set $k = 1$. This completes the proof of Theorem 2.

Proof of Theorem 3

We take the first element A_1 of $\mathbb{A} = \{A_1, A_2, \dots, A_n\}$ and partition it into two, $B, A_1 - B$, so that $\mathbb{B} = \{B, A_1 - B, A_2, \dots, A_n\}$. The two entropies are

$$\begin{aligned} \Phi(\mathbb{A}) &\coloneqq \varphi\big(P(A_1)\big) + \sum_{i=2}^{n} \varphi\big(P(A_i)\big), \\ \Phi(\mathbb{B}) &\coloneqq \varphi\big(P(B)\big) + \varphi\big(P(A_1 - B)\big) + \sum_{i=2}^{n} \varphi\big(P(A_i)\big) \end{aligned} \tag{2.187}$$

and hence

$$\Phi(\mathbb{B}) - \Phi(\mathbb{A}) = \varphi\big(P(B)\big) + \varphi\big(P(A_1 - B)\big) - \varphi\big(P(A_1)\big) \tag{2.188}$$

The right-hand side is non-negative, because φ is a concave function (corollary 2.3).

Proof of Theorem 4

If all P_i are unknown and no constraint is imposed on them, except that their sum is 1, the Lagrangian to be extremized is

$$L = -\sum_{i=1}^{n} P(A_i) \ln P(A_i) - \lambda\left(1 - \sum_{i=1}^{n} P(A_i)\right) \tag{2.189}$$

Then $\partial L / \partial P(A_i) = 0$ results in $-1 - \ln P(A_i) + \lambda = 0$ or $P(A_i) = \exp(\lambda - 1)$. Hence all $P(A_i)$ are equal to each other, $P(A_i) = 1/n$.

Proof of Theorem 5

For any positive numbers a_i, b_i, such that $a_1 + \cdots + a_n = 1$, $b_1 + \cdots + b_n \leq 1$, the following inequality holds true (Papoulis, 1991, p. 548):

$$-\sum_{i=1}^{n} a_i \ln a_i \leq -\sum_{i=1}^{n} a_i \ln b_i \tag{2.190}$$

with equality corresponding to the identity $a_i = b_i$. Hence, if $\sum_{i=1}^{n} B(A_i) \leq 1$, then

$$\Phi(\mathbb{A}||\mathrm{B}) \coloneqq -\sum_{i=1}^{n} P(A_i) \ln\left(\frac{P(A_i)}{B(A_i)}\right) = -\sum_{i=1}^{n} P(A_i) \ln P(A_i) + \sum_{i=1}^{n} P(A_i) \ln B(A_i) \leq 0 \tag{2.191}$$

with equality if $B(A_i) = P(A_i)$. In other words, if $\sum_{i=1}^{n} B(A_i) = 1$ the maximum possible relative entropy $\Phi(\mathbb{A}||\mathrm{B}) = 0$ and if for some i, $B(A_i) \neq P(A_i)$, $\Phi(\mathbb{A}||\mathrm{B}) < 0$.

However, if $\sum_{i=1}^{n} B(A_i) > 1$ (so $\ln \sum_{i=1}^{n} B(A_i) > 0$), then

$$\begin{aligned}\Phi(\mathbb{A}||\mathrm{B}) &\coloneqq -\sum_{i=1}^{n} P(A_i)\left(\ln\left(\frac{P(A_i)}{B(A_i)/\sum_{i=1}^{n} B(A_i)}\right) - \ln\sum_{i=1}^{n} B(A_i)\right) \\ &\leq 0 - \sum_{i=1}^{n} P(A_i)\left(-\ln\sum_{i=1}^{n} B(A_i)\right) = \ln\sum_{i=1}^{n} B(A_i)\end{aligned} \tag{2.192}$$

Proof of Theorem 6
See proof of Theorem 2 in Papoulis (1991, p. 550).

Proof of Theorem 7
See proof of Theorem 4 in Papoulis (1991, p. 551).

Proof of Theorem 8
See proof of Theorem 1 in Papoulis (1991, p. 550).

Appendix 2-II: Entropy maximizing distributions with two constraints

To get a different family of distributions from those in Table 2.5 we add a second constraint, namely $\mathrm{E}\left[\ln\left(B'(\underline{x})\right)\right] = \gamma_2$, while in order not to increase the number of parameters we set $d = 1$. Then equation (2.89) switches to

$$f(x) = \frac{A}{\lambda}\exp\left(-b\frac{B(x)}{\lambda} + \ln\big(B'(x)\big) + b_2\ln\big(B'(x)\big)\right) \tag{2.193}$$

which can be written as

$$f(x) = \frac{A}{\lambda}\exp\left(-b\,\frac{B(x)}{\lambda} + \ln(B'(x)^e)\right) \tag{2.194}$$

where $e \coloneqq 1 + b_2$. After the algebraic operations we find the maximum entropy distribution as:

$$f(x) = \frac{Ac^e}{\lambda}\left(\frac{x}{\lambda}\right)^{(c-1)e}\left(1 + \left(\frac{x}{\lambda}\right)^c\right)^{-b-e} \tag{2.195}$$

This is listed in Table 2.7, after reparameterization, as the generalized (power transformed) beta prime distribution (where the standard beta prime corresponds to $c = 1$). Two special cases of it are also listed there. The generalized gamma and the generalized beta prime distributions were also studied in Koutsoyiannis (2005a,c, where additional information for some of their characteristics are provided) and Papalexiou and Koutsoyiannis (2012).

Appendix 2-III: Irreconcilability of heavy tails and Lebesgue measure

We will prove that maximizing entropy with the Lebesgue measure and constrained mean cannot result in heavy tails. First we develop a method to check if a specified density is a local maximum of entropy under constrained mean. Then we check it with the exponential distribution (light tailed) to see that it works. Subsequently, we apply it to the Pareto distribution to prove that this is not a local entropy maximizer for a Lebesgue background measure.

Step 1: We consider a probability density $f(x)$ and a perturbation $g(x)$ of it close to its mean, namely in the interval $(\mu - \delta, \mu + \delta)$ with a small δ so that it does not affect the asymptotic behaviour at the tails (because it is local at the neighbourhood of the mean). Furthermore, we choose $g(x)$ so as not to affect the integral constraints, i.e.:

Table 2.7 Special cases of maximum entropy distributions given by equation (2.195). (Expressions not to be used at face as different parameterizations may be used throughout the book.)

Name	Parameters	$f(x)$	$\bar{F}(x) = 1 - F(x)$
Half Student[1]	$c = 2, e = 0$ $\xi = 1/(b + e - 1)$	$\frac{2}{\lambda\,\mathrm{B}\left(\frac{1}{2}, \frac{1}{2\xi}\right)} \left(1 + \left(\frac{x}{\lambda}\right)^2\right)^{-\frac{1}{2}-\frac{1}{2\xi}}$	$\frac{\mathrm{B}_{\frac{y}{1+y}}\left(\frac{1}{2}, \frac{1}{2\xi}\right)}{\mathrm{B}\left(\frac{1}{2}, \frac{1}{2\xi}\right)}, \quad y = \left(\frac{x}{\lambda}\right)^2$
Half extended Student	$c = 2, \zeta = e + 1,$ $\xi = 1/(b + e - 1)$	$\frac{2\left(\left(\frac{x}{\lambda}\right)^2\right)^{\frac{\zeta}{2}-\frac{1}{2}} \left(1 + \left(\frac{x}{\lambda}\right)^2\right)^{-\frac{\zeta}{2}-\frac{1}{2\xi}}}{\lambda\,\mathrm{B}\left(\frac{\zeta}{2}, \frac{1}{2\xi}\right)}$	$\frac{\mathrm{B}_{\frac{y}{1+y}}\left(\frac{\zeta}{2}, \frac{1}{2\xi}\right)}{\mathrm{B}\left(\frac{\zeta}{2}, \frac{1}{2\xi}\right)}, \quad y = \left(\frac{x}{\lambda}\right)^2$
Generalized beta prime (GBP)[2]	$\varsigma = c,$ $\zeta = (c - 1)e + 1),$ $\xi = 1/(b + e - 1)$	$\frac{\varsigma \left(\frac{x}{\lambda}\right)^{\zeta-1} \left(1 + \left(\frac{x}{\lambda}\right)^{\varsigma}\right)^{-\frac{\zeta}{\varsigma}-\frac{1}{\varsigma\xi}}}{\lambda\,\mathrm{B}\left(\frac{\zeta}{\varsigma}, \frac{1}{\varsigma\xi}\right)}$	$\frac{\mathrm{B}_{\frac{y}{1+y}}\left(\frac{\zeta}{\varsigma}, \frac{1}{\varsigma\xi}\right)}{\mathrm{B}\left(\frac{\zeta}{\varsigma}, \frac{1}{\varsigma\xi}\right)}, \quad y = \left(\frac{x}{\lambda}\right)^{\varsigma}$

[1] Also known as *Tsallis* or *1-particle kappa distribution* (Olbert, 1968; Livadiotis and McComas 2013). Special case: Cauchy.
[2] Special case: Beta prime (for $\varsigma = 1$).

$$\int_{\mu-\delta}^{\mu+\delta} g(x)\,\mathrm{d}x = 0, \qquad \int_{\mu-\delta}^{\mu+\delta} x\,g(x)\,\mathrm{d}x = 0 \tag{2.196}$$

Note that these also imply that for any constant c

$$\int_{\mu-\delta}^{\mu+\delta} (x + c) g(x)\,\mathrm{d}x = 0 \tag{2.197}$$

The entropy of the perturbed distribution will be

$$\begin{aligned} \Phi_g[\underline{x}] &= -\int_{\mu-\delta}^{\mu+\delta} \ln\big(f(x) + g(x)\big)\big(f(x) + g(x)\big)\mathrm{d}x \\ &= -\int_{\mu-\delta}^{\mu+\delta} \left(\ln f(x) + \ln\left(1 + \frac{g(x)}{f(x)}\right)\right)\big(f(x) + g(x)\big)\mathrm{d}x \end{aligned} \tag{2.198}$$

Since $g(x)/f(x) \ll 1$ (more precisely, for $g(x)/f(x) \to 0$), we have $\ln(1 + g(x)/f(x)) = g(x)/f(x)$. Hence

$$\begin{aligned} \Phi_g[\underline{x}] &= -\int_{\mu-\delta}^{\mu+\delta} \left(\ln f(x) + \frac{g(x)}{f(x)}\right)\big(f(x) + g(x)\big)\mathrm{d}x \\ &= -\int_{\mu-\delta}^{\mu+\delta} \ln f(x)\,\big(f(x) + g(x)\big)\mathrm{d}x - \int_{\mu-\delta}^{\mu+\delta} \frac{g(x)}{f(x)}\big(f(x) + g(x)\big)\mathrm{d}x \end{aligned} \tag{2.199}$$

The first of the latter two terms equals

$$\Phi[\underline{x}] - \int_{\mu-\delta}^{\mu+\delta} \ln f(x)\, g(x)\, \mathrm{d}x \tag{2.200}$$

and the second term equals

$$-\int_{\mu-\delta}^{\mu+\delta} g(x)\mathrm{d}x - \int_{\mu-\delta}^{\mu+\delta} \frac{g(x)^2}{f(x)}\mathrm{d}x = -\int_{\mu-\delta}^{\mu+\delta} \frac{g(x)^2}{f(x)}\mathrm{d}x \tag{2.201}$$

Finally, the perturbation in entropy is

$$\Delta\Phi = \Phi_g[\underline{x}] - \Phi[\underline{x}] = -\int_{\mu-\delta}^{\mu+\delta} \ln f(x)\, g(x)\, \mathrm{d}x - \int_{\mu-\delta}^{\mu+\delta} \frac{g(x)^2}{f(x)}\mathrm{d}x \tag{2.202}$$

Step 2: As we have seen, the exponential distribution maximizes entropy for constrained mean and therefore no perturbation $g(x)$ exists that could lead to $\Delta\Phi > 0$. We check if the above method confirms this. The probability density and its logarithm are

$$f(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}}, \qquad -\ln f(x) = \ln\lambda + \frac{x}{\lambda} \tag{2.203}$$

The first integral in the rightmost expression in (2.202) is

$$\int_{\mu-\delta}^{\mu+\delta} \left(\ln\lambda + \frac{x}{\lambda}\right) g(x)\, \mathrm{d}x \tag{2.204}$$

and because of (2.196) it evaluates to zero. Hence:

$$\Delta\Phi = -\int_{\mu-\delta}^{\mu+\delta} \frac{g(x)^2}{f(x)}\mathrm{d}x \le 0 \tag{2.205}$$

because $g(x)^2 \ge 0$ and $f(x) \ge 0$.

Step 3: A single counterexample suffices to show that distributions other than exponential are not local extremes. To this aim we consider a Pareto distribution specified as follows:

$$f(x) = 2(1+x)^{-3}, \qquad -\ln f(x) = 3\ln(1+x) - \ln 2\,, \qquad \mu = 1 \tag{2.206}$$

Substituting these in (2.202) we get

$$\begin{aligned} \Delta\Phi &= \int_{1-\delta}^{1+\delta} 3\ln(1+x)\, g(x)\, \mathrm{d}x - \int_{1-\delta}^{1+\delta} \frac{g(x)^2}{2(1+x)^{-3}}\mathrm{d}x \\ &= 3\int_{1-\delta}^{1+\delta} \ln(1+x)\, g(x)\, \mathrm{d}x - \frac{1}{2}\int_{1-\delta}^{1+\delta} (1+x)^3 g(x)^2 \mathrm{d}x \end{aligned} \tag{2.207}$$

We consider the power expansion of the function $g(x)$:

$$g(x) = a_0 + a_1 x + a_2 x^2 + O(x^3) \tag{2.208}$$

We neglect the high order terms, calculate the integrals in (2.196), equate them to zero and obtain

$$g(x) = a_2\left((x-1)^2 - \frac{\delta^2}{3}\right) \tag{2.209}$$

where both a_2 and δ are chosen to be small.

The first integral in the rightmost term of (2.207), is

$$3\int_{1-\delta}^{1+\delta} \ln(1+x)\, g(x)\, \mathrm{d}x = \frac{4}{3}a_2\left(\delta(\delta^2-6) - 3(\delta^2-4)\tanh^{-1}\left(\frac{\delta}{2}\right)\right) \tag{2.210}$$

and the second integral is

$$-\frac{1}{2}\int_{1-\delta}^{1+\delta} (1+x)^3 g(x)^2 \mathrm{d}x = -\frac{8}{315}a_2^2\delta^5(28+11\delta^2) \tag{2.211}$$

The latter is of higher order than the former for both a_2 and δ, which are small, and thus it can be neglected. Hence the sign of $\Delta\Phi$ will depend on the signs of a_2 and of the quantity

$$A \coloneqq \delta(\delta^2-6) - 3(\delta^2-4)\tanh^{-1}\left(\frac{\delta}{2}\right) \tag{2.212}$$

The Taylor series expansion of A is

$$A = -\frac{\delta^5}{20} - \frac{3\delta^7}{560} - \frac{\delta^9}{1344} + O(\delta)^{11} \tag{2.213}$$

It is then evident that the sign of A, and hence of $\Delta\Phi$ can be either positive or negative, depending on δ, and hence the inequality $\Delta\Phi \leq 0$ does not hold true for any δ.

Chapter 3. Stochastic processes and quantification of change

3.1 Definitions

A deterministic worldview is founded on the idea of sharp exactness. A deterministic mathematical description of a system uses common variables (e.g. x) which are represented as numbers. The change of the system state is represented as a *trajectory* $x(t)$, which is the sequence of a system's states x as time t changes. Changes in time are studied using the concept of a dynamical system with certain dynamics. The term *dynamics* denotes a transformation S_t which maps its initial state $x(0)$ in the trajectory of a dynamical system (at time 0) to its current state $x(t)$ (at time t), that is, $x(t) = S_t\big(x(0)\big)$ (Lasota and Mackey 1994).

In an indeterministic worldview there is uncertainty or randomness, where the latter term simply means unpredictability or intrinsic uncertainty. Thus, to study change according to this approach we use the notion of a *stochastic process*. This is defined as an arbitrarily (usually infinitely) large family of stochastic variables $\underline{x}(t)$ (Papoulis, 1991). To each one of them there corresponds an index t, which takes values from an *index set* T, most often referring to time. The time t can be either *discrete* (when T is the set of integers, $\mathbb{Z}$) or continuous (when T is the set of real numbers, $\mathbb{R}$). Thus, we have respectively a *discrete-time* or a *continuous-time* stochastic process. As natural time runs continuously, the faithful representation of a natural process needs a model formulated for continuous time in order to avoid the risk of making artificial constructs. Even so, the discrete-time representation is necessary in simulation. Typically, the discrete time representation $\underline{x}_\tau$ is derived from the continuous time representation $\underline{x}(t)$ as the temporal average:

$$\underline{x}_\tau \coloneqq \frac{1}{D} \int_{(\tau-1)D}^{\tau D} \underline{x}(t)\mathrm{d}t \tag{3.1}$$

where $\tau \in \mathbb{Z}$ represents the continuous-time interval $[(\tau-1)D, \tau D]$ and D is the time step. Notice that we use different notation in the continuous and discrete time representation, in the latter case denoting time as a subscript. Each of the stochastic variables $\underline{x}(t)$ or $\underline{x}_\tau$ can be either discrete (e.g. the wet or dry state of a day) or continuous (e.g. the rainfall depth). Thus, we have respectively a *discrete-state* or a *continuous-state* stochastic process.

The index set can also be a vector space rather than the real line or the set of integers. This is the case for instance when we assign a stochastic variable (e.g. rainfall depth) to each geographical location (a two dimensional vector space) or to each location and time instance (a three-dimensional vector space). Stochastic processes with multidimensional index sets are also known as stochastic (or random) fields.

A realization $x(t)$ of a stochastic process $\underline{x}(t)$, which is a common (numerical) function of time t, is known as a *sample function*. Typically, a realization can be known (simulated) at countable time instances, i.e. in discrete time (not in continuous time, even in a continuous-time process). Likewise, observation of a natural process is also made in

discrete time. A sequence of simulated or observed values is called a *time series*. Clearly then, a time series is a *finite* sequence of *numbers*, whereas a stochastic process is a family of *stochastic variables*, infinitely many for discrete time processes and uncountably infinitely many for continuous time processes. A large body of literature does not make this distinction and confuses stochastic processes with time series (see Digression 3.F).

3.2 Distribution function and moments

The distribution function of the stochastic variable $\underline{x}(t)$ i.e.,

$$F(x;t) \coloneqq P\{\underline{x}(t) \leq x\} \tag{3.2}$$

is called *first-order distribution function* of the process. Likewise, the *second-order distribution function* is:

$$F(x_1, x_2; t_1, t_2) \coloneqq P\{\underline{x}(t_1) \leq x_1, \underline{x}(t_2) \leq x_2\} \tag{3.3}$$

and the (multivariate) n*th order distribution function* is:

$$F(x_1, x_2, \dots, x_n; t_1, t_2, \dots, t_n) \coloneqq P\{\underline{x}(t_1) \leq x_1, \underline{x}(t_2) \leq x_2, \dots, \underline{x}(t_n) \leq x_n\} \tag{3.4}$$

A stochastic process is completely determined if we know the nth order distribution function for any n. The nth order probability density function of the process is derived by taking the derivatives of the distribution function with respect to all x_i.

The moments (and quantities related to them) are defined in the same manner as in sections 2.13 and 2.15. Of particular interest are the following:

1. The process *mean*, i.e. the expected value of the variable $\underline{x}(t)$:

$$\mu(t) \coloneqq \mathrm{E}[\underline{x}(t)] = \int_{-\infty}^{\infty} x f(x;t)\mathrm{d}x \tag{3.5}$$

2. The process *variance*, i.e. the variance of the variable $\underline{x}(t)$:

$$\gamma_0(t) \coloneqq \mathrm{var}[\underline{x}(t)] = \int_{-\infty}^{\infty} \big(x - \mu(t)\big)^2 f(x;t)\mathrm{d}x \tag{3.6}$$

3. The process *autocovariance*, i.e. the covariance of the stochastic variables $\underline{x}(t)$ and $\underline{x}(t+h)$:

$$c(t;h) \coloneqq \mathrm{cov}[\underline{x}(t), \underline{x}(t+h)] = \mathrm{E}\left[\Big(\underline{x}(t) - \mu(t)\Big)\Big(\underline{x}(t+h) - \mu(t+h)\Big)\right] \tag{3.7}$$

   where $c(t;0) \equiv \gamma_0(t)$.

4. The process *autocorrelation*, i.e., the correlation coefficient of the stochastic variables $\underline{x}(t)$ and $\underline{x}(t+h)$):

$$r(t;h) \coloneqq \mathrm{corr}[\underline{x}(t), \underline{x}(t+h)] = \frac{c(t;h)}{\sqrt{\gamma_0(t)\gamma_0(t+h)}} \tag{3.8}$$

Additional characteristics will be given in section 3.5.

3.3 Stationarity

The term *process* has been introduced into the scientific vocabulary as synonymous with change, as evident in Kolmogorov's (1931) pioneering paper, in which he introduced the term *stochastic process*. This paper starts by stating "*A physical process* [is] *a change of a certain physical system*".

It is common in science to try to identify invariant properties within change (Koutsoyiannis 2011a). For example, in the absence of an external force, the position of a body in motion changes in time but the velocity is unchanged (Newton's first law). If a constant force is present, then the velocity changes but the acceleration is constant (Newton's second law). If the force changes, e.g. the gravitational force changing with the changing distance of the orbiting planets, then the acceleration is no longer constant, but other invariant properties emerge such as the angular momentum (Newton's law of gravitation; see also Koutsoyiannis 2011a).

Evidently, the notion of a stochastic process was invented to describe the irregular changes in natural systems which are more complex than the above, and which are impossible to model deterministically or predict their future evolution in full detail and with precision. Here, the great scientific achievement has been the invention of macroscopic descriptions, which replace the modelling of the details. This is essentially done using stochastics. Herein lies the essence and usefulness of the stationarity concept, which seeks invariant properties in complex systems (Koutsoyiannis, 2011a, 2014a; Koutsoyiannis and Montanari, 2015). Following Kolmogorov (1931, 1938) and Khintchine (1934), a process is stationary if its statistical properties are invariant to a shift of time origin, i.e. $\underline{x}(t)$ and $\underline{x}(t+h)$ have the same (multivariate) distribution for any t and a specified h. Furthermore, following Kolmogorov (1947), a process is called *wide-sense stationary* if its mean is constant and its autocovariance depends only on time differences, i.e.:

$$\mathrm{E}\big[\underline{x}(t)\big] = \mu \ (= \text{constant}), \qquad \mathrm{cov}\big[\underline{x}(t), \underline{x}(t+h)\big] = c(h) \tag{3.9}$$

A strict-sense stationary process is also wide-sense stationary, but the inverse is not true.

A process that is not stationary is called nonstationary. In a nonstationary process one or more statistical properties are *deterministic functions of time*. A typical case of a nonstationary process is a cumulative process whose mean is proportional to time. For instance, let us assume that the rainfall intensity at a geographical location and time of year is modelled as a stationary process $\underline{x}(t)$, with mean μ. Let us further denote $\underline{X}(t)$ the rainfall depth collected in a large container (a cumulative raingauge) at time t and assume that at the time origin, $t = 0$, the container is empty, so that $\underline{X}(t) = \int_0^t \underline{x}(s)\mathrm{d}s$. It is easy then to understand that $\mathrm{E}\big[\underline{X}(t)\big] = \mu t$. This is a deterministic (linear) function of time t and thus $\underline{X}(t)$ is a nonstationary process.

We should stress that stationarity and nonstationarity are properties of a process, not of a sample function or time series. There is some confusion in the literature about this, as a lot of studies assume that a time series is either stationary or not, or can reveal whether the process is stationary or not. As a general rule, to characterize a process as

nonstationary, it suffices to show that a specific statistical property is a *deterministic* function of time (as in the above example of the raingauge), but this cannot be inferred from a time series alone. A time series formed from observations of a natural process cannot be stationary or nonstationary.

Stochastic processes describing periodic phenomena, such as those affected by the annual cycle of Earth, are nonstationary. For instance, the daily temperature at a mid-latitude location could not be regarded as a stationary process. It could be modelled as a special kind of a nonstationary process with characteristics depending on time in a periodical manner (as periodic functions of time). Such processes are called *cyclostationary* processes.

3.4 Ergodicity

Stationarity is also related to another important stochastic concept, *ergodicity*.* Its importance derives from the fact that ergodicity is a prerequisite to making inference from data, that is, *induction*—the Aristotelian *ἐπαγωγή* (*epagoge*). This is a type of inference weaker than *deduction*—the Aristotelian *ἀπόδειξις* (*apodeixis*) —albeit very useful when deduction is not possible (see Digression 4.A).

In dynamical systems, by definition (e.g. Mackey, 2003, p. 48), ergodicity is the property of a system whose invariant sets under the dynamic transformation are all trivial (have zero probability). In other words, in an ergodic transformation starting from any point, the trajectory of the system state will visit all other points, without being trapped to a certain subset. The ergodic theorem (Birkhoff, 1931; Khintchine, 1933; see also Mackey, 2003, p. 54), allows for the redefining of ergodicity within the stochastics domain (Papoulis, 1991, p. 427; Koutsoyiannis 2010) in the following manner: A stochastic process $\underline{x}(t)$ is ergodic if the time average of any (integrable) function $g\left(\underline{x}(t)\right)$, as time tends to infinity, equals the true (ensemble) expectation, i.e.:

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T g\left(\underline{x}(t)\right)\mathrm{d}t = \mathrm{E}\left[g\left(\underline{x}(t)\right)\right], \qquad \lim_{T\to\infty}\frac{1}{T}\sum_{\tau=0}^{T} g\left(\underline{x}_\tau\right) = \mathrm{E}\left[g\left(\underline{x}(t)\right)\right] \tag{3.10}$$

for a process in continuous or in discrete time, respectively.

The right-hand side in the above equations represents the true average, also known as the *ensemble average*, whereas the left-hand side represents the time average, for the limiting case of infinite time. The left-hand side in each of the two equations (3.10) is a stochastic variable (as a sum or integral of stochastic variables) and is not a function of the time *t*. Hence, the right-hand side should not be a function of the time *t*, i.e. the process should be stationary. Furthermore, the right-hand side is a number, not a stochastic variable. Equating a number with a stochastic variable implies that the stochastic variable

* The concept of ergodicity was first conceived by Boltzmann (1884/85), who coined the terms *ergode* and *isodic*, both of which are etymologized from Greek words but which ones exactly is uncertain. Most probably, *ergodic* comes from the Greek *ἔργον* (ergon = work) and *ὁδός* (hodos = pathway). According to another interpretation, the second noun is *εἶδος* (eidos = form, kind, nature), or the whole word is a transliteration of the Greek adjective *ἐργώδης* (ergodes = laborious, troublesome; see Mathieu 1988).

has zero variance. This is precisely the condition that makes a process ergodic. And this allows the estimation (i.e. approximate calculation) of the true but unknown property $\mathrm{E}\left[g\left(\underline{x}(t)\right)\right]$ from the time average of $g\left(\underline{x}(t)\right)$, that is, from the available data. Without ergodicity inference from data would not be possible.

A stochastic process for which it can be shown that the property (3.10) holds true for the particular case that $g\left(\underline{x}(t)\right) = \underline{x}(t)$, whose expectation is the mean ($\mathrm{E}\left[\underline{x}(t)\right] = \mu$), is called *mean-ergodic*. The property could be extended to multivariate functions, e.g. $g\left(\underline{x}(t), \underline{y}(t)\right)$, and thus we can speak about *covariance-ergodic* processes. Further information, including conditions that should hold for ergodicity can be found in Papoulis (1991).

Now, if the system that is modelled in a stochastic framework has deterministic dynamics (meaning that a system input will give a single system response), then a theorem applies (Mackey 2003, theorem 4.5 p. 52), according to which a dynamical system with dynamics $S_t(x)$ has a stationary probability density *if and only if* it is ergodic. Therefore, a stationary system is also ergodic and vice versa, and a nonstationary system is also non-ergodic and vice versa. Here we note that even if a system has deterministic dynamics, again it is legitimate to use a stochastic description, replacing the study of the evolution of system states $S_t(x)$ with the evolution of probability densities of states $f(x;t)$. One reason to prefer the stochastic description over the purely deterministic description is that the former includes quantification of uncertainty, whereas deterministic dynamics does not eliminate uncertainty (Koutsoyiannis 2010). Furthermore, we clarify that the deterministic description through the transformation $S_t(x)$ is fully compatible with a stochastic description that is stationary and ergodic, according to the theorem stated above: While the system state is changing in time t according to the transformation $S_t(x)$, its statistical properties (and the probability density $f(x;t)$) can be constant in time (i.e. $f(x)$).

If the system dynamics is stochastic (a single input could result in multiple outputs), then ergodicity and stationarity do not necessarily coincide. However, recalling that a stochastic process is a model and not part of the real world, we can always conveniently devise a stochastic process that is ergodic, provided we exclude nonstationarity. In conclusion, from a practical point of view ergodicity can generally be assumed when there is stationarity and the variance of the time averaged process tends to zero as the time of averaging tends to infinity, while this assumption is fully justified by the theory if the system dynamics is deterministic. Conversely, if nonstationarity is assumed, then ergodicity cannot hold, thus disallowing inference from data. This contradicts the basic premise in geosciences, where data are the only reliable information for building models and making inference and prediction.

Digression 3.A: Misuses of stationarity and ergodicity

Despite having a central role in stochastics, the concepts of stationarity and ergodicity have been widely misunderstood and broadly misused (Montanari and Koutsoyiannis, 2014; Koutsoyiannis and Montanari, 2015). In an attempt to find trends everywhere, according to the popular motto

"stationarity is dead" (Milly et al., 2008), trend analysis of hydroclimatic processes is more fashionable today than ever before (Iliopoulou and Koutsoyiannis, 2020). The notion of a trend, as a fundamental constituent of time series, is very old, but it is fundamentally problematic (Koutsoyiannis, 2020), despite its popularity.

Ironically, most of these studies use time series data to estimate statistical properties, as if the process were ergodic, while at the same time their cursory estimates falsify the ergodicity hypothesis. The correct tactic, even when dealing with provably nonstationary and nonergodic processes and our study is based on data, is to convert the process to a stationary and ergodic one before trying to make any inference from the data.

As an example, assuming that we deal with the cumulative rainfall process $\underline{X}(t)$, used as an example of a nonstationary process in section 3.3, we convert the process into a stationary one in discrete time by $\underline{x}_\tau \coloneqq \underline{X}(\tau D) - \underline{X}\big((\tau-1)D\big)$, where D is a time step, and perform the same transformation to the time series data. Then we can use the x_τ data to make inferences.

As a second example related to trends, let us examine a statement such as: "By analysing the time series x_τ (where τ denotes discrete time), we concluded that it is nonstationary, and we identified an increasing trend with slope b." This is an incorrect statement and can be corrected in the following manner: "We analysed the time series x_τ based on the modelling assumption that the stochastic process $\underline{x}_\tau - b\tau$ is stationary and ergodic, which enabled the estimation of the slope b." The latter statement respects the fact that we always need stationarity and ergodicity to make inference from data. It also avoids using the vague term "trend", which, despite being trendy, has no scientific definition. Finally, it reveals the fact that the entire setting is just a modelling assumption—not anything objective, related to physical reality.

3.5 Second-order characteristics of stochastic processes

Along with the definition of a stochastic process (section 3.1), we have also provided that of the autocovariance function (section 3.2), an important characteristic of the second-order distribution function of a stochastic process. However, there are other second-order characteristics that are useful to study, as they have certain properties that help the understanding and the simulation of stochastic processes.

Before defining them, starting with the process of interest $\underline{x}(t)$, we will better explain the concepts of the cumulative process $\underline{X}(t)$ and the discrete-time process $\underline{x}_\tau$, which have already been introduced. As graphically shown in Figure 3.1, the cumulative process is defined as:

$$\underline{X}(t) \coloneqq \int_0^t \underline{x}(u)\mathrm{d}u \tag{3.11}$$

where obviously $\underline{X}(0) \equiv 0$. If $\underline{x}(t)$ aims to represent a natural process, then $\underline{X}(t)$ should necessarily be nonstationary. However, by time averaging (dividing the cumulative process by time) and differencing, we may construct a stationary process over any time scale D, provided that $\underline{x}(t)$ is stationary. With the help of the cumulative process, the discrete-time representation of the process (equation (3.1)) can be written as:

$$\underline{x}_\tau(D) \coloneqq \frac{1}{D}\int_{(\tau-1)D}^{\tau D} \underline{x}(u)\mathrm{d}u = \frac{\underline{X}(\tau D) - \underline{X}\big((\tau-1)D\big)}{D} \tag{3.12}$$

The argument (D) in $\underline{x}_\tau(D)$ denotes the time step of discretization. In cases where we use a single discretization step and there is no ambiguity we will omit it, writing $\underline{x}_\tau$. We can also define discrete-time processes in multiples of D, say κD, where κ is an integer:

$$\underline{x}_\tau^{(\kappa)} := \underline{x}_\tau(\kappa D) := \frac{1}{\kappa D} \int_{(\tau-1)\kappa D}^{\tau\kappa D} \underline{x}(u)\mathrm{d}u = \frac{\underline{X}(\tau\kappa D) - \underline{X}\big((\tau-1)\kappa D\big)}{\kappa D} \tag{3.13}$$

Obviously, the discrete-time process $\underline{x}_\tau^{(\kappa)}$ is the time average (at scale $k = \kappa D$), of the discrete-time process $\underline{x}_\tau$ (at scale equal to the time step D):

$$\underline{x}_\tau^{(\kappa)} = \frac{1}{\kappa} \sum_{j=(\tau-1)\kappa+1}^{\tau\kappa} \underline{x}_j \tag{3.14}$$

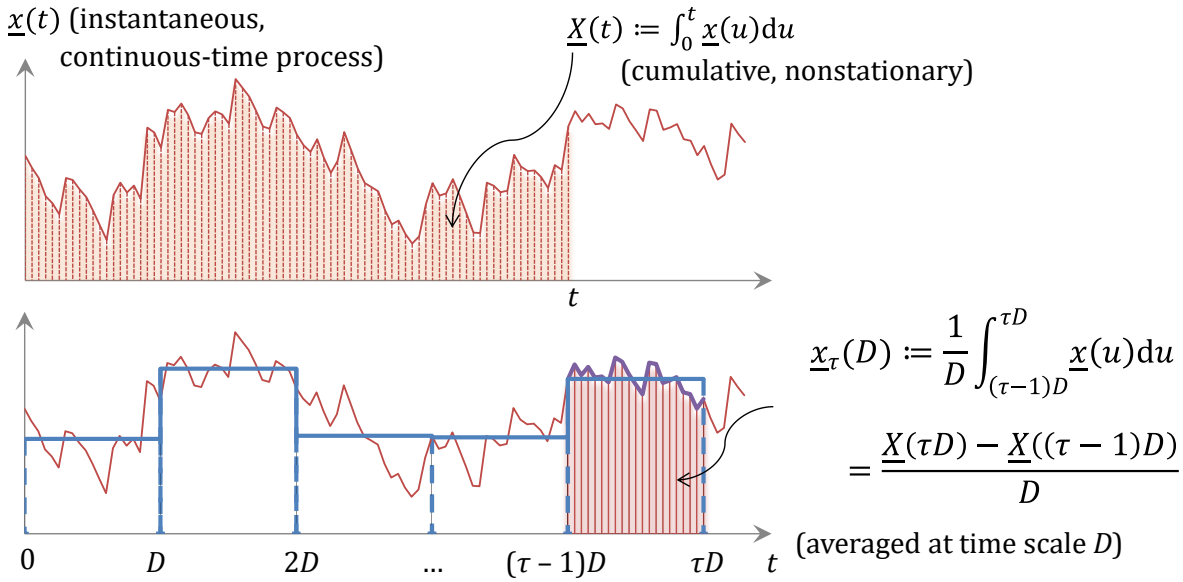

Figure 3.1 Explanatory sketch for a stochastic process in continuous time and its representation in discrete time. Note that the graphs display a realization of the process (it is impossible to display the process per se) while the notation is for the process per se.

The variance of $\underline{X}(t)$ at time t, as a function of t, i.e.:

$$\Gamma(t) := \mathrm{var}\big[\underline{X}(t)\big] \tag{3.15}$$

is known as the *cumulative climacogram*. The variance of the time averaged process $\underline{X}(k)/k$ at a time scale k, as a function of the time scale k, is the continuous-time variant of the *climacogram*, already discussed in section 2.17:

$$\gamma(k) := \mathrm{var}\left[\frac{\underline{X}(k)}{k}\right] = \frac{\Gamma(k)}{k^2} \tag{3.16}$$

The square root $\sqrt{\gamma(k)} =: \sigma(k)$ (standard deviation) is referred to as the σ-climacogram.

The autocovariance function $c(h)$ of the continuous-time process $\underline{x}(t)$ for time lag h, already defined in equation (3.7), is related to the climacogram by (Koutsoyiannis 2016):

$$c(h) \coloneqq \mathrm{cov}[x(t), x(t\ +\ h)] = \frac{1}{2}\,\frac{\mathrm{d}^2\Gamma(h)}{\mathrm{d}h^2} \tag{3.17}$$

If we deal with two processes $\underline{x}(t)$ and $\underline{y}(t)$ we can define the cross-covariance:

$$c_{xy}(h) \coloneqq \mathrm{cov}\left[\underline{x}(t), \underline{y}(t+h)\right] \tag{3.18}$$

This is a continuous-time metric. To also involve the time scale k of the averaged process, we can define the *cross-climacogram* (Koutsoyiannis, 2019b):

$$\gamma_{xy}(k;\eta) \coloneqq \sigma_x\sigma_y\ \mathrm{var}\left[\frac{\underline{X}(k)}{k\sigma_x} + \frac{\underline{Y}\big((\eta+1)k\big) - \underline{Y}(\eta k)}{k\sigma_y}\right] \tag{3.19}$$

where $\underline{Y}(k) \coloneqq \int_0^k \underline{y}(t)\mathrm{d}t$ and η is lag.

The *structure function* (also known as *semivariogram* or *variogram*), $\upsilon(h)$, is another second-order tool, defined as:

$$\upsilon(h) \coloneqq \frac{1}{2}\mathrm{var}\big[\underline{x}(t) - \underline{x}(t+h)\big] = c(0) - c(h) \tag{3.20}$$

The *power spectrum* (also known as *spectral density*), $s(w)$, where w denotes frequency is defined as the Fourier transform of the autocovariance function, i.e.:

$$s(w) \coloneqq 4\int_0^\infty c(h)\cos(2\pi w h)\,\mathrm{d}h \tag{3.21}$$

The power spectrum should necessarily be nonnegative at all w ($s(w) \geq 0$), and this entails that the autocovariance $c(h)$ should be a positive definite function. Also, the climacogram $\gamma(k)$ should be a positive definite function (Koutsoyiannis, 2017).

The power spectrum has some analogies with another stochastic tool, the so-called climacospectrum (Koutsoyiannis, 2017), which is directly given as a transformation of the climacogram. Specifically, it is proportional to the difference of the variances of the averaged process at time scales k and $2k$:

$$\psi(k) \coloneqq \frac{k\big(\gamma(k) - \gamma(2k)\big)}{\ln 2} \tag{3.22}$$

The climacospectrum can also be written in an alternative manner in terms of frequency $w = 1/k$:

$$\tilde{\psi}(w) \coloneqq \psi\left(\frac{1}{w}\right) = \frac{\gamma(1/w) - \gamma(2/w)}{(\ln 2)\,w} \tag{3.23}$$

It is useful to note that the entire area under the power spectrum $s(w)$, as well as that under the curve $\tilde{\psi}(w)$, are precisely equal to each other and to the variance γ_0.

All definitions of second-order characteristics in continuous time are gathered together in Table 3.1. Once any one of these characteristics is known in the continuous-time representation, we can calculate all the others in continuous time as well as those in discrete time, as shown in Table 3.2.

Table 3.1 Summary of notation and second-order characteristics of a stationary stochastic process in continuous time.

Name	Symbol and definition	Remarks	Eqn. no.
Stochastic process of interest	$\underline{x}(t)$	Assumed stationary	
Time, continuous	t	Dimensional	
Cumulative process	$\underline{X}(t) \coloneqq \int_0^t \underline{x}(\xi)\mathrm{d}\xi$	Nonstationary	(3.11)
Variance, instantaneous	$\gamma_0 \coloneqq \mathrm{var}\left[\underline{x}(t)\right]$	Constant (not a function of t)	(3.6)
Cumulative climacogram	$\Gamma(t) \coloneqq \mathrm{var}\left[\underline{X}(t)\right]$	A function of t; $\Gamma(0) \equiv 0$	(3.15)
Climacogram	$\gamma(k) \coloneqq \mathrm{var}\left[\dfrac{\underline{X}(k)}{k}\right] = \dfrac{\Gamma(k)}{k^2}$	A function of time scale; $\gamma(0) = \gamma_0$ as $k \to 0$	(3.16)
Time scale, continuous	k	Units of time	
Climacospectrum	$\psi(k) \coloneqq \dfrac{k\left(\gamma(k) - \gamma(2k)\right)}{\ln 2}$		
Autocovariance function	$c(h) \coloneqq \mathrm{cov}[x(t), x(t\ +\ h)]$	$c(0) = \gamma_0$	(3.17)
Time lag, continuous	h	Units of time	
Structure function (semivariogram, variogram)	$\upsilon(h) \coloneqq \dfrac{1}{2}\mathrm{var}\left[\underline{x}(t) - \underline{x}(t+h)\right]$	$\upsilon(h) = \gamma_0 - c(h)$	(3.20)
Power spectrum (spectral density)	$s(w) \coloneqq 4\int_0^\infty c(h)\cos(2\pi w h)\,\mathrm{d}h$	$\int_0^\infty s(w)\mathrm{d}w = \gamma_0$	(3.21)
Frequency, continuous	$w\ =\ 1/k$	Units of inverse time	

Table 3.2 Summary of notation and second-order characteristics of a stationary stochastic process in discrete time.

Name	Symbol and definition	Remarks	Eqn. no.
Stochastic process, discrete time	$\underline{x}_\tau \coloneqq \dfrac{1}{D}\int_{(\tau-1)D}^{\tau D} \underline{x}(u)\mathrm{d}u = \dfrac{\underline{X}(\tau D) - \underline{X}\left((\tau-1)D\right)}{D}$		(3.12)
Discretization time step	D	Length of time window of averaging	
Time, discrete	$\tau \coloneqq t/D$	Dimensionless	
Averaged stochastic process, discrete time	$\underline{x}_\tau^{(\kappa)} = \dfrac{1}{\kappa}\sum_{j=(\tau-1)\kappa+1}^{\tau\kappa} \underline{x}_j$		(3.14)
Time scale, discrete	$\kappa = k/D$	Dimensionless	(3.24)
Climacogram	$\gamma_\kappa = \mathrm{var}\left[\underline{x}_\tau^{(\kappa)}\right] = \gamma(\kappa D) = \dfrac{\Gamma(\kappa D)}{(\kappa D)^2}$	$\gamma_1 = \mathrm{var}\left[\underline{x}_\tau\right] = \gamma(D)$	(3.25)
Climacospectrum	$\psi_\kappa = \psi(k) = \dfrac{\kappa(\gamma_\kappa - \gamma_{2\kappa})}{\ln 2}$		
Autocovariance function	$c_\eta \coloneqq \mathrm{cov}\left[\underline{x}_\tau, \underline{x}_{\tau+\eta}\right]$	$c_0 = \gamma(D) = \gamma_1$	
Time lag, discrete	$\eta = h/D$	Dimensionless	(3.26)
Structure function	$\upsilon_\eta = \gamma_1 - c_\eta$		(3.27)
Power spectrum	$s_\mathrm{d}(\omega) = \dfrac{1}{D}\sum_{j=-\infty}^{\infty} s\left(\dfrac{\omega + j}{D}\right)\mathrm{sinc}^2\left(\pi(\omega + j)\right)$		(3.28)
Frequency, discrete	$\omega = wD = 1/\kappa$	Dimensionless	(3.29)

Note: In time-related quantities, Latin letters denote dimensional quantities and Greek letters dimensionless ones, as specified above.

The reverse is not true, i.e., from a model formulated in discrete time we cannot precisely infer the characteristics of the continuous-time representation. It may be seen in Table 3.2 that the expressions of the discrete time characteristics may differ substantially from those in continuous time, and thus attention is needed to avoid confusion and misuse.

The rule that continuous- and discrete-time characteristics are different has exceptions: The climacogram and the climacospectrum are not affected by discretization (they admit the same expressions for both continuous and discrete time). They also have some additional advantages, such as simplicity, a close relationship to entropy (see below), and a more stable behaviour than other common tools (Dimitriadis and Koutsoyiannis, 2015; Koutsoyiannis, 2016; 2017). These make them the tool of preference in stochastic modelling—even though they are less popular than other tools. All these tools are transformations of one another, as listed in Table 3.3.

Table 3.3 Relationships between second-order characteristics of a stochastic process.

Related characteristics	Symbol and definition	Inverse relationship	Eqn. no.
$\gamma(k) \leftrightarrow c(h)$	$\gamma(k) = 2\int_0^1 (1-\chi)c(\chi k)\mathrm{d}\chi$	$c(h) = \frac{1}{2}\frac{\mathrm{d}^2\left(h^2\gamma(h)\right)}{\mathrm{d}h^2}$	(3.30)
$s(w) \leftrightarrow c(h)$	$s(w) \coloneqq 4\int_0^\infty c(h)\cos(2\pi w h)\,\mathrm{d}h$	$c(h) = \int_0^\infty s(w)\cos(2\pi w h)\,\mathrm{d}w$	(3.31)
$\gamma(k) \leftrightarrow s(w)$	$\gamma(k) = \int_0^\infty s(w)\ \mathrm{sinc}^2(\pi w k)\,\mathrm{d}w$	$s(w) \coloneqq 2\int_0^\infty \frac{\mathrm{d}^2\left(h^2\gamma(h)\right)}{\mathrm{d}h^2}\cos(2\pi w h)\,\mathrm{d}h$	(3.32)
$\upsilon(h) \leftrightarrow c(h)$	$\upsilon(h) = \gamma_0 - c(h)$	$c(h) = \upsilon(\infty) - \upsilon(h)$ with $\upsilon(\infty) = \gamma_0$	(3.33)
$\psi(k) \leftrightarrow \gamma(k)$	$\psi(k) \coloneqq \frac{k\left(\gamma(k) - \gamma(2k)\right)}{\ln 2}$	$\gamma(k) = \ln 2\sum_{i=0}^\infty \frac{\psi(2^i k)}{2^i k} = \gamma(0) - \ln 2\sum_{i=1}^\infty \frac{\psi(2^{-i}k)}{2^{-i}k}$	(3.34)
$\gamma_\kappa \equiv \gamma(\kappa D) \leftrightarrow c_\eta$	$\gamma_\kappa = \frac{1}{\kappa}\left(c_0 + 2\sum_{\eta=1}^{\kappa-1}\left(1-\frac{\eta}{\kappa}\right)c_\eta\right) = \frac{\Gamma(\kappa D)}{(\kappa D)^2}$ where $\Gamma(0)=0, \Gamma(D)=c_0D^2$ and, recursively, $\Gamma(\kappa D) = 2\Gamma\left((\kappa-1)D\right) - \Gamma\left((\kappa-2)D\right) + 2c_{j-1}D^2$	$c_\eta = \frac{1}{D^2}\left(\frac{\Gamma(\lvert\eta+1\rvert D) + \Gamma(\lvert\eta-1\rvert D)}{2} - \Gamma(\lvert\eta\rvert D)\right)$	(3.35)
$c_\eta \leftrightarrow s_\mathrm{d}(\omega)$	$s_\mathrm{d}(\omega) = 2c_0 + 4\sum_{\eta=1}^\infty c_\eta\cos(2\pi\eta\omega)$	$c_\eta = \int_0^{1/2} s_\mathrm{d}(\omega)\cos(2\pi\omega\eta)\,\mathrm{d}\omega$	(3.36)
$\upsilon_\eta \leftrightarrow c_\eta$	$\upsilon_\eta = \gamma(D) - c_\eta$	$c_\eta \coloneqq \gamma(D) - \upsilon_\eta$	(3.37)

Digression 3.B: What is dependence in time?

Dependence can be simply defined as the absence of independence. With reference to equation (2.5) defining independence and using equations (3.2)–(3.4), we define dependence in a

stochastic process in time (also known as *intertemporal dependence* or simply *time dependence*) by

$$F(x_1, x_2, \dots, x_n; t_1, t_2, \dots, t_n) \neq F(x_1; t_1)F(x_2; t_2) \cdots F(x_n; t_n)$$

It is typically expressed by the autocovariance or the autocorrelation function and its typical (mis)interpretation is *memory*. This has been so common that in many texts the term *memory* has replaced the term *dependence*—even in the titles of several publications, papers and books. Perhaps the scientist who was most influential in establishing this interpretation was Mandelbrot (for example, Mandelbrot and Wallis, 1968, speak about short and long memory, both of which they contrast to independence), though other scientists had used the term before (e.g. Krumbein, 1968). Clearly, in stochastics the term *memory* is metaphorical, while in other disciplines (neuropsychology, computer science) it is literal. In science there is no reason to use a metaphorical term when we have a literal term, particularly when the metaphorical term has another scientific meaning. The use of the metaphorical term *memory* distracts, rather than helps, intuition and the understanding of time dependence in a stochastic process. In particular, the use of its variant *long memory* is totally inappropriate as it stimulates people to imagine a mechanism inducing long memory (e.g. hundreds of years) and of course it is difficult to conceptualize such a mechanism. A better interpretation is a mechanism that produces change, rather than one that recalls information (as is the meaning of memory). And indeed, changes produce dependence—not the other way round. Furthermore, dependence and change need not be interpreted as nonstationarity as many think.

Before discussing how change produces time dependence in a process that is stationary, we will discuss how dependence manifests itself into a time series. In one word, this manifestation is through *patterns*. In pure randomness, without time dependence (like in a sequence of dice outcomes or in the sequence of digits of π) no patterns appear. To better illustrate such patterns, we examine several time series with a small length, $n = 16$. For convenience we make these time series two-valued, with values –1 and 1 and with an average of the 16 values equal to zero, which means that eight values will be –1 and eight 1. The *estimates* of the variance, the lag-one autocovariance and the lag-one autocorrelation coefficient will thus be, respectively:

$$\hat{\gamma}_1 = \frac{1}{16}\sum_{\tau=1}^{16} x_\tau^2 = 1, \qquad \hat{c}_1 = \frac{1}{16}\sum_{\tau=1}^{16} x_\tau x_{\tau+1}\,, \qquad \hat{r}_1 = \frac{\hat{c}_1}{\hat{\gamma}_1} = \hat{c}_1$$

where we set $x_{17} = x_1$ in order to have 16 terms in the sum for $\hat{c}_1$ and thus make possible values up to ±1. (Note, though, that this practice is not being suggested to use in analyses of time series). The formal meaning of the term *estimate* is clarified in section 4.3.

Examples of such time series are shown in Figure 3.2. In the upper left panel, all eight ones are grouped together so that $\sum_{\tau=1}^{16} x_\tau x_{\tau+1} = 7 + 7 - 2 = 12$ and $\hat{r}_1 = 0.75$. This is the highest possible value that a particular arrangement of 16 items, each being ±1, can give. Obviously, there are 16 possible arrangements that will give $\hat{r}_1 = 0.75$. If our time series had length *N*, the highest $\hat{r}_1$ would be $(N-4)/N = 1 - 4/N$ and would approach the value +1 for large *N*. Consequently, a large autocorrelation is caused by grouping together of similar (in our example, the same) values. Such grouping has been termed *persistence*. If the grouping appears but is not that "perfect", such as in the lower left panel, then again, the autocorrelation will be positive but lower ($\hat{r}_1 = 0.5$ in this example).

In contrast, if the patterns appear to be of alternating, rather than grouping, type, then the autocorrelation coefficient is negative. Thus, in the "perfect" alternating shape of the upper middle panel of Figure 3.2 we have $\sum_{\tau=1}^{16} x_\tau x_{\tau+1} = -16$ and $\hat{r}_1 = -1$. In the lower middle panel, alternation is not perfect and $\hat{r}_1 = -0.75$. Finally, the upper right panel is free of patterns and $\hat{r}_1 = 0$.

Now, the effect of change is illustrated in Figure 3.3, where we plot a time series generated from the normal distribution without time dependence. We now assume that the process is affected by a mechanism producing change, namely shifts up and down, at random points in time. As illustrated in Figure 3.3 and detailed in the figure caption, in this case patterns are produced and (positive) autocorrelation is induced.

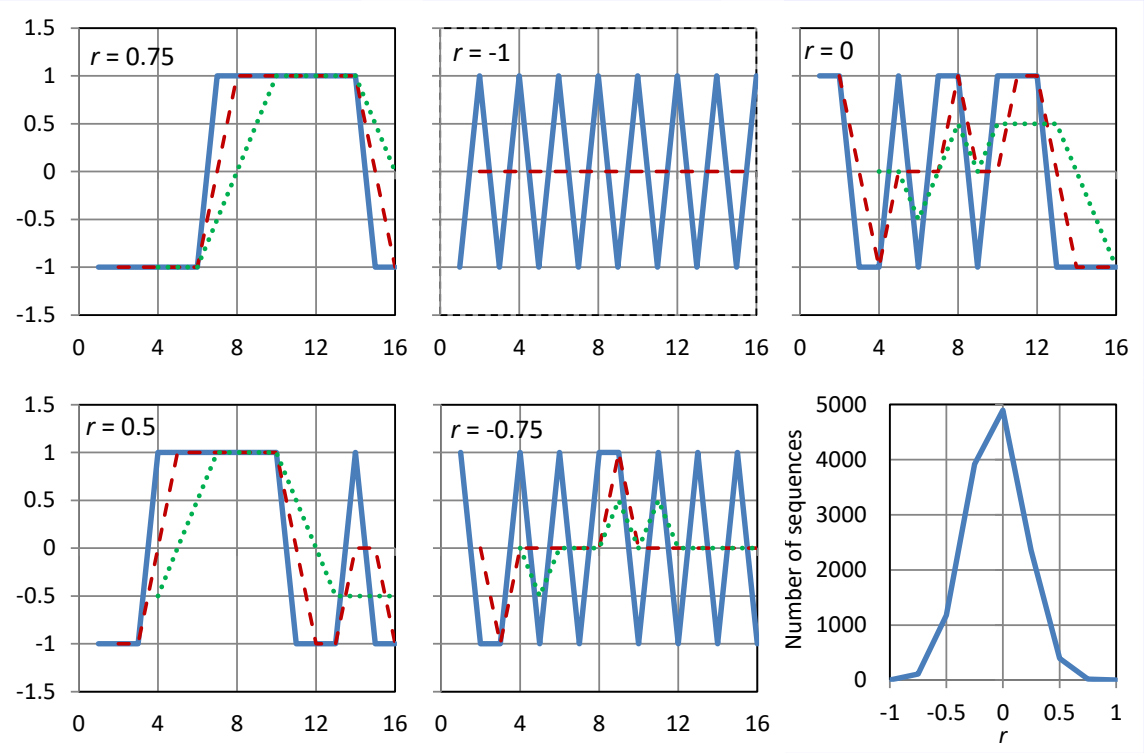

Figure 3.2 Examples of arrangements of eight ones and eight minus ones in the form of time series with length 16, mean zero and unit variance, along with the resulting estimate of the lag-one autocorrelation coefficients *r*. In addition to the original time series (scale 1; continuous line), time-averaged time series are also shown at scales 2 (dashed lines) and 4 (dotted lines). In the bottom right panel, the frequency distribution of *r* for all $16!/(8!)^2 = 12\,870$ possible cases (permutations) are shown.

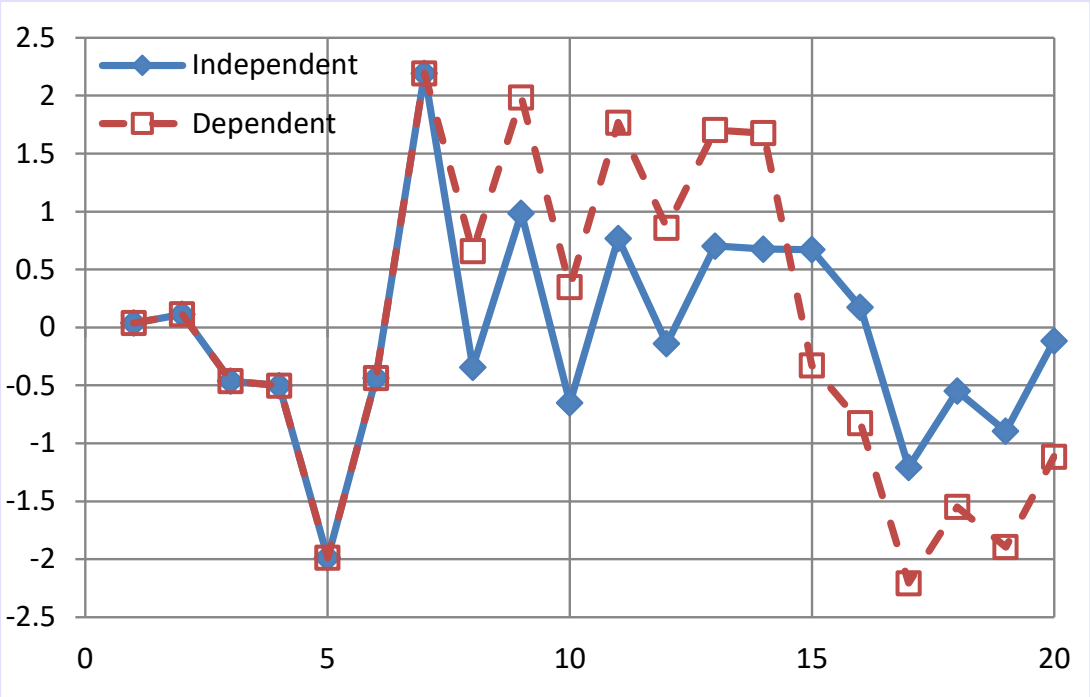

Figure 3.3 Illustration of the fact that change causes autocorrelation using a time series of length 20, generated from the normal distribution N(0,1) without time dependence; the estimates of the statistical characteristics from the time series, plotted as full points connected with continuous lines, are $\hat{\mu} = -0.05, \hat{\gamma}_0 = 0.9^2, \hat{r}_1 = 0.05$. By shifting a time segment up (by +1, items 8-14) and another segment down (by –1, items 15-20) we obtain a new time series (empty points connected with dashed lines) in which the autocorrelation has become $\hat{r}_1 = 0.59$.

Had such change been describable in deterministic terms, as a deterministic function of time, that is, had it been precisely predictable in terms of location of times where it occurs and in terms of magnitude of state shifts, we would speak about nonstationarity. But since, as we said, the points of change are random points in time, they resist a deterministic description and the entire process with the change-producing mechanism is a *stationary stochastic process with dependence*. Unfortunately, this simple truth is not widely understood and therefore the inconsistent interpretations of change as nonstationarity abound in geophysics literature.

3.6 Asymptotic power laws and the log-log derivative

It is quite common for nonnegative functions $f(t)$ defined in $[0, \infty)$ to be associated with asymptotic *power laws* as $t \to 0$ and ∞ (Koutsoyiannis, 2014b, 2017). *Power laws* are functions of the form

$$f(t) \propto t^b \tag{3.38}$$

A power law is visualized in a graph of $f(t)$ plotted against t with logarithmic axes, so that the plot forms a straight line with slope b. Formally, the slope b is expressed by the *log-log derivative* (LLD):

$$f^{\#}(t) := \frac{\mathrm{d}(\ln f(t))}{\mathrm{d}\,(\ln t)} = \frac{t f'(t)}{f(t)} \tag{3.39}$$

We notice that $f^{\#}(t)$ is a dimensionless quantity, irrespective of the dimensions of $f(t)$. If the power law holds for the entire domain, then $f^{\#}(t) = b =$ constant. In this case we speak about a *simple scaling* behaviour. Usually, however, $f^{\#}(t)$ is not constant. Of particular interest are the *asymptotic values* for $t \to 0$ and ∞, symbolically $f^{\#}(0)$ and $f^{\#}(\infty)$, which define *two asymptotic power laws*. We note that, if $0 < f(0) < \infty$, then $f^{\#}(0) = 0$, which means that $f(0)$ has to be either 0 or ∞ in order for $f^{\#}(0) \neq 0$. Basic properties of LLD are given in Table 3.4.

Table 3.4 Basic properties of LLD (from Koutsoyiannis, 2017).

Description	Mathematical formula
Multiplication and addition by constants	$(\lambda f(t) + \mu)^{\#} = f^{\#}(t)$
Sum of two functions	$\big(f_1(t) + f_2(t)\big)^{\#} = \dfrac{f_1(t) f_1^{\#}(t) + f_2(t) f_2^{\#}(t)}{f_1(t) + f_2(t)}$
Product of two functions	$\big(f_1(t) f_2(t)\big)^{\#} = f_1^{\#}(t) + f_2^{\#}(t)$
Quotient of two functions	$\big(f_1(t)/f_2(t)\big)^{\#} = f_1^{\#}(t) - f_2^{\#}(t)$
Raise to a power	$\big(f(t)^{\lambda}\big)^{\#} = \lambda f^{\#}(t)$
Function composition	$\big((f \circ g)(t)\big)^{\#} = \big(f(g(t)\big)^{\#} = f^{\#}\big(g(t)\big)\, g^{\#}(t)$

In particular, the asymptotic properties of the second-order characteristics of a stochastic process for $t \to 0$, where t now denotes time, characterize the *local behaviour* of a process, while those for $t \to \infty$ characterize the *global behaviour*. We will discuss these properties in section 3.8, after introducing the related concept of entropy production in section 3.7.

3.7 Entropy production in stochastic processes

In a stochastic process the change of uncertainty in time can be quantified by the *entropy production,* i.e. the time derivative of the entropy $\Phi\big[\underline{X}(t)\big]$ of the cumulative process $\underline{X}(t)$ (Koutsoyiannis, 2011b):

$$\Phi'\big[\underline{X}(t)\big] \coloneqq \frac{\mathrm{d}\Phi\big[\underline{X}(t)\big]}{\mathrm{d}t} \tag{3.40}$$

A more convenient (and dimensionless) indicator is the *entropy production in logarithmic time* (EPLT):

$$\varphi(t) \equiv \varphi\big[\underline{X}(t)\big] \coloneqq \Phi'\big[\underline{X}(t)\big]t \equiv \frac{\mathrm{d}\Phi\big[\underline{X}(t)\big]}{\mathrm{d}(\ln t)} \tag{3.41}$$

For a Gaussian process, the entropy depends on its variance $\Gamma(t)$ only (see Table 2.4) and is given as:

$$\Phi\big[\underline{X}(t)\big] = \frac{1}{2}\ln\big(2\pi\mathrm{e}\beta^2\Gamma(t)\big) \tag{3.42}$$

where β is the background measure density, assumed to be constant (Lebesgue). The EPLT of a Gaussian process is thus easily shown to be:

$$\varphi(t) = \frac{\Gamma'(t)t}{2\Gamma(t)} = 1 + \frac{\gamma'(t)t}{2\gamma(t)} = \frac{\Gamma^{\#}(t)}{2} = 1 + \frac{\gamma^{\#}(t)}{2} \tag{3.43}$$

That is, EPLT is visualized and estimated by the slope of a log-log plot of the climacogram. Note that, if the cumulative process were used and the background measure density were taken as βt insteaf of β, the result would be practically the same (plus a constant 1).

When the past and the present are observed, instead of the unconditional variance $\gamma(t)$ we should use a variance $\gamma_{\mathrm{C}}(t)$ conditional on the known past and present. This can be expressed in terms of the differenced climacogram (Koutsoyiannis, 2017):

$$\gamma_{\mathrm{C}}(k) = \varepsilon\big(\gamma(k) - \gamma(2k)\big), \qquad \varepsilon = \frac{1}{1 - 2^{\gamma^{\#}(\infty)}} \tag{3.44}$$

We can subsequently define the *conditional entropy production in logarithmic time* (CEPLT) in a manner analogous to (3.43). By also considering the definition of the climacospectrum in (3.22) and (3.23), CEPLT can be written as:

$$\varphi_{\mathrm{C}}(t) = 1 + \frac{\gamma_{\mathrm{C}}^{\#}(t)}{2} = \frac{1 + \psi^{\#}(t)}{2} = \frac{1 - \tilde{\psi}^{\#}(1/t)}{2} \tag{3.45}$$

Thus, for a Gaussian process the conditional entropy production is given in terms of log-log slope of the process climacospectrum. We will use the same result as an approximation for non-Gaussian processes too, even though in a non-Gaussian process the entropy expression becomes more complicated than (3.42) with other terms additional to variance.

3.8 Asymptotic scaling of second-order properties

EPLT and the CEPLT are related to LLDs of second-order tools such as the climacogram, the climacospectrum, the power spectrum, etc. With a few exceptions, these slopes are nonzero asymptotically, hence entailing asymptotic scaling or asymptotic power laws with the LLDs being the scaling exponents. Intuitively, one would expect that an emerging asymptotic scaling law would provide a good approximation of the true law for a range of scales.

If the scaling law were appropriate for the entire range of scales, then we would have a simple scaling law. Such simple scaling sounds attractive from a mathematical point of view, but in physical processes it turns out to be impossible (Koutsoyiannis, 2017; Koutsoyiannis et al., 2018; see also below). It is thus physically more realistic to expect two different types of asymptotic scaling laws, one in each end of the continuum of scales. The respective scaling exponents are given in terms of two parameters, M (to give credit to Mandelbrot) and H (to give credit to Hurst) according to the following relationships:

- The parameter M characterizes the *local scaling* or *smoothness* or *fractal behaviour*, when $k \to 0$ or $w \to \infty$:

$$M \coloneqq \varphi_{\mathrm{C}}(0) - 1 = \frac{\gamma_{\mathrm{C}}^{\#}(0)}{2} = \frac{\upsilon^{\#}(0)}{2} = \frac{\psi^{\#}(0) - 1}{2} = \frac{-s^{\#}(\infty) - 1}{2} \tag{3.46}$$

- The parameter H characterizes the *global scaling* or *persistence* or *Hurst-Kolmogorov* behaviour, when $k \to \infty$ or $w \to 0$:

$$H \coloneqq \varphi_{\mathrm{C}}(\infty) = 1 + \frac{\gamma_{\mathrm{C}}^{\#}(\infty)}{2} = 1 + \frac{\gamma^{\#}(\infty)}{2} = 1 + \frac{c^{\#}(\infty)}{2} = \frac{\psi^{\#}(\infty) + 1}{2} = \frac{-s^{\#}(0) + 1}{2} \tag{3.47}$$

These scaling behaviours have emerged from maximum entropy considerations, and this may provide the theoretical background to the modelling of complex natural processes using such scaling laws. Generally, scaling laws are a mathematical necessity and could be constructed for virtually any continuous function defined in $[0, \infty)$. In other words, there is no magic in power laws, except that they are, logically and mathematically, a necessity (Koutsoyiannis, 2014b).

3.9 Bounds of scaling: The global map of stochastic processes

Both parameters M and H take on values in the interval (0,1) (with the limiting cases $M = 1$ and $H = 0$ being possible). This fact, combined with equations (3.46) and (3.47), defines limits of the possible scaling laws in natural processes. These limits are not well known, and several studies have reported values outside these limits. (See Digression 3.C for an example about how to avoid such mistakes.)

For the global behaviour, it has been shown (Koutsoyiannis et al., 2018) that a process with $-s^{\#}(0) > 1$ is nonergodic. As already explained, inference from data is only possible when the process is ergodic and thus, claiming that $-s^{\#}(0) > 1$ based on data is self-

contradictory. Steep slopes ($-s^{\#}(w) > 1$) are mathematically and physically possible for medium and large w and indeed they are quite frequent in geophysical and other processes. Because of the equality of slopes of power spectrum and climacospectrum, the ergodicity limitation holds also for the slope of the climacospectrum, i.e., $\psi^{\#}(\infty) = -\tilde{\psi}^{\#}(0) < 1$ (cf. Digression 3.C). On the other hand, negative asymptotic slopes of the climacospectrum that are too steep are also impossible. Indeed (because of (52)), $\psi^{\#}(k) = -\tilde{\psi}^{\#}(1/k) < -1$ would entail $\varphi_{\mathrm{C}}(k) < 0$ and $\Gamma_{\mathrm{C}}'(k) < 0$ (Koutsoyiannis, 2017). This means that the variance of the cumulative process would be a decreasing function of time, which is absurd. This holds both for the global case ($k \to \infty$, where the conditional variance $\Gamma_C(\infty)$ equals the unconditional $\Gamma(\infty)$) and for the local case ($k \to 0$, for the conditional variance $\Gamma_{\mathrm{C}}(0)$).

For the local behaviour, there is another severe limitation imposed by physical reasoning. The case $\psi^{\#}(0) = -s^{\#}(\infty) < 1$ would entail infinite variance. Infinite variance would require infinite energy to emerge, which is physically inconsistent (see also section 2.18). Therefore, the physical lower limit for $\psi^{\#}(0) = -s^{\#}(\infty)$ is 1. A final—and quite severe—limitation is an upper bound of the local scaling exponent, which is 3 for $\psi^{\#}(0) = -s^{\#}(\infty)$ (Koutsoyiannis, 2017). If this limitation was violated, then the resulting autocovariance function would not be positive definite or, equivalently, the resulting power spectrum would not be positive for any frequency w but would take on negative values for some w. However, by definition, the power spectrum should be positive for any w. Likewise, in violation, the Fourier transform of the climacogram would take on negative values for some w. Proof is provided in Koutsoyiannis (2017).

The above limits define the "green square" of admissible values of φ_{C}, M and H in Figure 3.4, which is also depicted in terms of admissible values of slopes $\psi^{\#}$ and $s^{\#}$ (noting that $s^{\#}$ can, by exception, take on values out of the square when $\varphi_C(0) = 2$ or $\varphi_C(\infty) = 0$). The reasons why a process out of the square would be impossible or inconsistent, as discussed above, are also marked in the figure.

The centre of the square, with coordinates $\varphi_C(0) = \frac{3}{2}$, $\varphi_C(\infty) = \frac{1}{2}$ represents a neutral process, whose typical representative is the Markov process (to be examined in section 3.11). Larger values of $\varphi_C(0)$ (where $M > 1/2$) indicate a smooth process and smaller ones (where $M < 1/2$) a rough process. Also, larger values of $\varphi_C(\infty)$ (where $H > 1/2$) indicate a persistent process and smaller ones (where $H < 1/2$) an antipersistent process.

A useful observation in Figure 3.4 is that the entire "green square" lies below the equality line. This means that the same scaling exponent is not possible for both local and global behaviour, or else, it is impossible to have a physically realistic simple scaling process. There is one exception, the upper-left corner of the "green square", which corresponds to the so-called "pink noise" or "$1/f$ noise". This will be discussed further in Digression 3.H.

On the left of the "green square" in Figure 3.4 another square is formed, which represents processes that are mathematically feasible but physically unrealistic, because they entail infinite variance. In particular, the centre of this square represents the *white noise* characterized by independence in time, which is discussed in section 3.10. One of

the diagonals of this square represents the Hurst-Kolmogorov process, discussed in section 3.12.

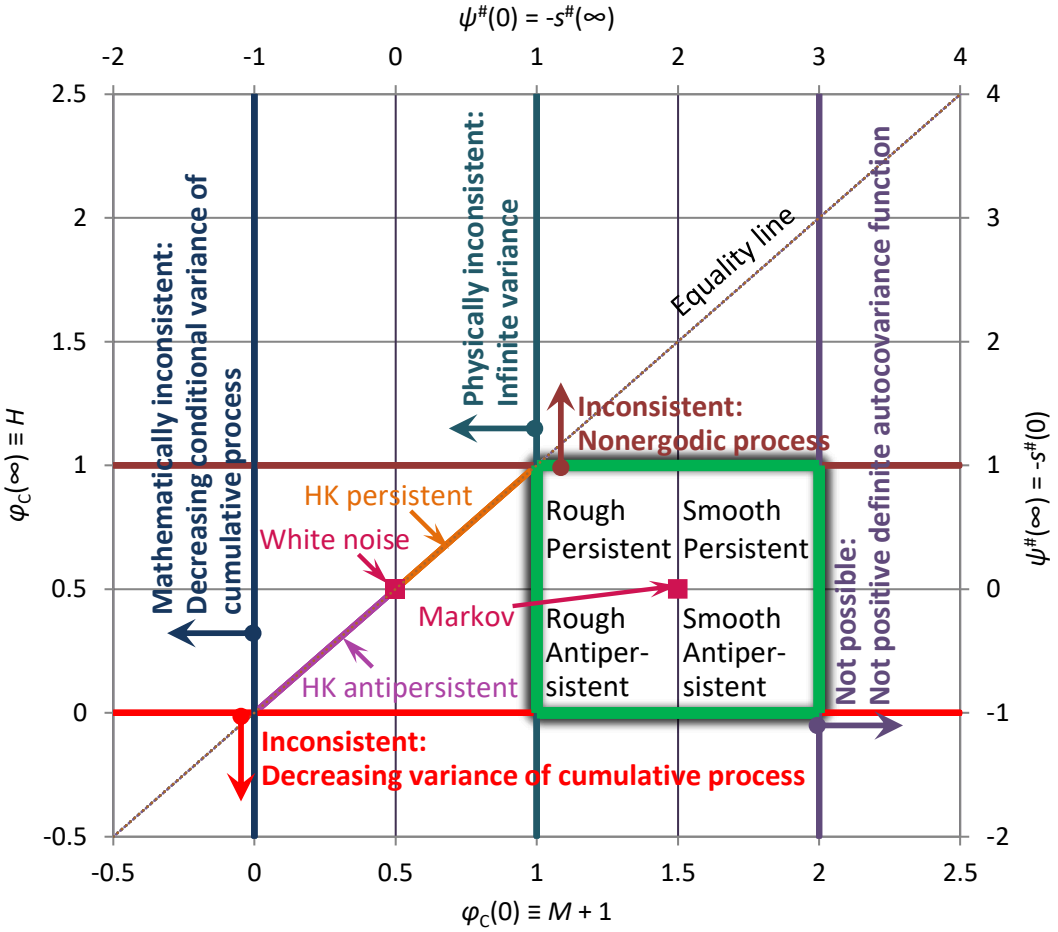

Figure 3.4 Bounds of asymptotic values of CEPLT, $\varphi_C(0)$ and $\varphi_C(\infty)$, and corresponding bounds of the log-log slopes of power spectrum and climacospectrum. The "green square" represents the admissible region; note that $s^{\#}$ can, by exception, take on values out of the square when $\varphi_C(0) = 2$ $(M = 1)$ or $\varphi_C(\infty) = 0$ $(H = 0)$. The reasons why a process out of the square would be impossible or inconsistent are also marked. The lines $\varphi_C(0) = 3/2$ $(M = 1/2)$ and $\varphi_C(\infty) = 1/2$ $(H = 1/2)$ define *neutrality* (which is represented by a Markov process) and support the classification of stochastic processes into the indicated four categories (smaller squares within the "green square"). (Source: Koutsoyiannis, 2017.)

Digression 3.C: Misuses of stationarity and ergodicity (2)

Continuing the examples of the misuse of the concepts of stationarity and ergodicity in Digression 3.A, we refer here to another example, whose standard formulation could be: "From the time series x_τ, we calculated the power spectrum and found that its slope for low frequencies is steeper than –1, which means that the process is nonstationary." We note that a large number of studies exploring several data sets have reported steep constant slopes of power spectrum, i.e. $\beta < -1$, which are thought to confirm the nonstationarity of the process. The fact of the matter is, however, that this entire line of thought is theoretically inconsistent. Such reported numerical results are artefacts caused by insufficient data or inadequate estimation algorithms. Once we construct the power spectrum of a process as a function of frequency only, we have tacitly assumed a stationary process. In a nonstationary process, both the autocovariance and the spectral density, i.e. the Fourier transform of the autocovariance, are functions of two variables, one being related to

"absolute" time (see e.g. Dechant and Lutz, 2015). Thus, it is meaningless to use a stationary representation (setting the power spectrum as a function of frequency only) and simultaneously claim nonstationarity. Furthermore, once we use the power spectrum of a process for inference, as we always do, we should be aware that inference from data is only possible when the process is ergodic. As shown in Koutsoyiannis et al. (2018), in an ergodic process, the asymptotic slope on the lower tail of the power spectrum cannot be steeper than –1. Thus, it is meaningless to report slopes in empirical power spectra < -1 and at the same time make any claims about the process properties (e.g. of nonstationarity) based on the power spectrum. Actually, such a steep slope, when emerging from processing of data, does not suggest that a process is non-ergodic. Rather it signifies inconsistent estimation. Nonetheless, we should be aware, that steep slopes (< -1) are mathematically and physically possible for medium and large frequencies, as has already been discussed.

Consequently, possible remedies for the above inconsistent statement could be the following:

- We cursorily interpreted a slope steeper than –1 in the power spectrum as evidence of nonstationarity, while a simple explanation would be that the frequencies on which our data enable calculation of the power spectrum values are too high.
- We cursorily applied the concept of the power spectrum of a stationary stochastic process, forgetting that the empirical power spectrum of a stationary stochastic process is a (nonstationary) stochastic process per se (see section 4.10). The high variability of the latter (or the inconsistent numerical algorithm we used) resulted in a slope for low frequencies steeper than –1, which is absurd. Such a slope would suggest a non-ergodic process while our calculations were based on the hypothesis of a stationary and ergodic process.
- We cursorily applied the concept of the power spectrum of a stationary stochastic process using a time series which is a realization of a nonstationary stochastic process, and we found an inconsistent result; therefore, we will repeat the calculations recognizing that the power spectrum of a nonstationary stochastic process is a function of two variables, frequency and "absolute" time.

3.10 White noise: how natural and how white is it?

We are all familiar with the notion of independent events at discrete time, such as coin, dice and roulette wheel experiments. If such an experiment is performed sequentially in time, we can model it as a stochastic process $\underline{v}'_\tau, \tau = 1,2 \dots$ with mean μ and variance γ_1. For convenience we subtract its mean, defining the process $\underline{v}_\tau := \underline{v}'_\tau - \mu$ for which:

$$\mathrm{E}[\underline{v}_\tau] = 0, \qquad \mathrm{var}[\underline{v}_\tau] = \mathrm{E}[\underline{v}_\tau^2] = \sigma^2, \qquad c_\eta = \mathrm{cov}[\underline{v}_\tau, \underline{v}_{\tau+\eta}] = \begin{cases} \sigma^2, & \eta = 0 \\ 0, & \eta \neq 0 \end{cases} \tag{3.48}$$

It is easy to show that the time-averaged process:

$$\underline{v}_\tau^{(\kappa)} := \frac{1}{\kappa} \sum_{i=(\tau-1)\kappa}^{\tau\kappa} \underline{v}_i \tag{3.49}$$

has the following properties:

$$\mathrm{E}\left[\underline{v}_\tau^{(\kappa)}\right] = 0, \quad \gamma_1^{(\kappa)} = \mathrm{var}\left[\underline{v}_\tau^{(\kappa)}\right] = \frac{\sigma^2}{\kappa}, \quad c_\eta^{(\kappa)} = \mathrm{cov}\left[\underline{v}_\tau^{(\kappa)}, \underline{v}_{\tau+\eta}^{(\kappa)}\right] = \begin{cases} \dfrac{\sigma^2}{\kappa}, & \eta = 0 \\ 0, & \eta \neq 0 \end{cases} \tag{3.50}$$

Is it legitimate to say that the discrete-time process $\underline{v}_\tau$ originates from a continuous time process $\underline{v}(t)$? And if it is, what are the properties of the latter? The mathematical answer to the former question is positive. To materialize the continuous-time variant it

suffices to generalize the climacogram in (3.50) changing the time scale from an integer κ to a real number $k \coloneqq \kappa D$:

$$\gamma(k) = \mathrm{var}\left[\underline{v}(t)\right] = \frac{\sigma^2 D}{k} \tag{3.51}$$

It is easily seen that if $k \to 0$, the process variance tends to infinity. Thus, to express the properties of the continuous-time process, we need to involve the Dirac delta function $\delta(t)$, whose properties are:

$$\delta(t) = \begin{cases} \infty, & t = 0 \\ 0, & t \neq 0 \end{cases}, \qquad \int_a^b \delta(t)\mathrm{d}t = 1 \tag{3.52}$$

where $[a, b]$ is any interval that contains the 0. To connect the discrete-time process $\underline{v}_\tau$ to the continuous-time process $\underline{v}(t)$, we assume that the former is the time-average of the latter on the time interval of length D, as in equation (3.12). If we define $\underline{v}(t)$ as a stationary stochastic process which has the following properties:

$$\mathrm{E}\left[\underline{v}(t)\right] = 0, \qquad \mathrm{cov}\left[\underline{v}(t), \underline{v}(t')\right] = \mathrm{E}\left[\underline{v}(t)\underline{v}(t')\right] = \sigma^2 D\ \delta(t - t') \tag{3.53}$$

then it results in a discrete-time process with the properties of equation (3.50). Indeed, the variance of $\underline{v}_\tau$ will be:

$$\begin{aligned} \mathrm{var}\left[\underline{v}_\tau\right] = \mathrm{var}\left[\underline{v}_1\right] &= \mathrm{E}\left[\left(\frac{1}{D}\int_0^D \underline{v}(t)\mathrm{d}t\right)^2\right] = \frac{1}{D^2}\mathrm{E}\left[\int_0^D \underline{v}(t)\mathrm{d}u \int_0^D \underline{v}(s)\mathrm{d}s\right] \\ &= \frac{1}{D^2}\int_0^D\int_0^D \mathrm{E}\left[\underline{v}(t)\underline{v}(s)\right]\mathrm{d}t\mathrm{d}s = \frac{1}{D^2}\int_0^D\int_0^D \sigma^2 D\delta(t-s)\,\mathrm{d}t\mathrm{d}s = \frac{\sigma^2 D}{D^2}\int_0^D 1\mathrm{d}s = \sigma^2 \end{aligned} \tag{3.54}$$

The power spectrum of process $\underline{v}(t)$ is found (from equation (3.21)) to be constant:

$$s(w) = \sigma^2 D \tag{3.55}$$

Because all frequencies w are present in the power spectrum with equal density ($\sigma^2 D$), the process $\underline{v}(t)$ has been called *white noise*. This name has been given by analogy to white light, which is a mixture of all visible frequencies. We note though that this is a misnomer as the power spectrum of the white light is very different from flat.

While mathematically the white noise is a well-founded concept and useful for many theoretical analyses, it is not physically realistic for several reasons, such as the following:

- Its variance is infinite: $\mathrm{var}\left[\underline{v}(t)\right] = \mathrm{E}\left[\left(\underline{v}(t)\right)^2\right] = \sigma^2 D\ \delta(0) = \infty$. If this represented a natural process, this process would have infinity energy.
- Its autocorrelation for lags however small is zero. In a natural process, the autocorrelation would be close to 1 for lags close to zero.
- Its spectral density is nonzero as frequency tends to infinity.

In brief, white noise is great for math and terrible for physics. Its problems can be remedied by applying some kind of filtering to the process $\underline{v}(t)$. An example would be to set an upper limit w_c to the frequency, beyond which the spectral density becomes zero (a so-called *low-pass* or *high-cut* filter). The second-order characteristics of the thus obtained stochastic process $\underline{\tilde{v}}(t)$ are:

$$\tilde{\gamma}_0 = \sigma^2 D\, w_c, \qquad \tilde{c}(h) = \sigma^2 D\, w_c\, \mathrm{sinc}(2\pi w_c h), \qquad \tilde{s}(w) = \begin{cases} \sigma^2 D, & w \le w_c \\ 0, & w > w_c \end{cases} \tag{3.56}$$

It may be readily seen that the above three inconsistencies have been remedied. On the other hand, the process $\underline{\tilde{v}}(t)$ does not precisely yield the process $\underline{v}_\tau$ in discrete time. However, if we choose $w_c \gg 1/D$, we can obtain a good approximation.

Digression 3.D: Random walk, Wiener process and Brownian motion

Assuming that the discrete-time white noise process $\underline{v}_\tau$ is two-valued, e.g. taking on the values +1 and –1 with equal probabilities of 1/2 (so that $\mathrm{E}[\underline{v}_\tau] = 0$), the cumulative process $\underline{V}_\tau := \sum_{i=1}^{\tau} \underline{v}_i$, which takes on values in the interval $[-\tau, \tau]$, is called a *random walk*. This is a nonstationary process with its variance being proportional (actually equal in this simple case) to the time τ that has passed from the beginning of the walk, i.e. $\mathrm{var}[\underline{V}_\tau] = \tau$. Its mean is zero at all times.

If both the time t and the state $\underline{v}(t)$ of the white noise are continuous, then the resulting cumulative process $\underline{V}(t) := \int_0^t \underline{v}(s)\mathrm{d}s$ is called the *Wiener process*. This is again a nonstationary process with mean zero and variance proportional to the time t, i.e. $\mathrm{var}[\underline{V}(t)] = \sigma^2 t$, where σ^2 has been defined above.

The quantity $\sigma^2/2$ is known as the *diffusion constant*. The Wiener process is used to model diffusion phenomena and the Brownian motion under free conditions, i.e., when there are no bounds in the motion, nor a restoring force (e.g. gravity in atmospheric motion). However, in real world systems the motion is not free (these conditions do not hold true) and the Brownian motion is bound. In such systems the resulting process is not Wiener but a stationary process.

More information on these processes can be found in Papoulis (1991). An application of the random walk process to derive a classical physical law is discussed in Digression 3.E.

Digression 3.E: Random walk, diffusion, Fick's first law and Fourier's law

Here again we consider the random walk process with some modifications. We assume that the process $\underline{v}_\tau$ can take on three values, 0, +1 and –1, and represents the motion in time $\Delta\tau$ of a particle sitting at a box at a position x to adjacent boxes in a unidimensional lattice with spacing Δx. The value $\underline{v}_\tau = 0$ represents the case that after time $\Delta\tau$ the particle remained at the box x, while the values $\underline{v}_\tau = +1$ and –1 represent the cases where the particle moved to the sites $x + \Delta x$ and $x - \Delta x$, respectively. We assign a probability p to the event that the particle moved and assume symmetry in the direction of motion (a consequence of entropy maximization). Hence the probabilities of $\underline{v}_\tau = 0, +1$ and –1 are $1 - p, p/2$ and $p/2$, respectively.

We assume that at time τ many particles are sited at x and we denote $c(x)$ the respective concentration (number of particles per unit volume) at position x. The number of particles in the volume defined by the increment Δx and unit vertical size is $N(x) = c(x)\Delta x$. We wish to find the net flux $q(x)$ between sites at x and $x + \Delta x$. The net flux q is the net number of particles crossing a unit area per unit time and equals the algebraic sum of two quantities, $q(x) = q_R(x) - q_L(x)$, where $q_R(x)$ is the flux to the right direction from x to $x + \Delta x$ and $q_L(x)$ the reverse flux, directed to the left. We have

$$q_R(x) = (p/2)N(x)/\Delta\tau, \qquad q_R(x) = (p/2)N(x + \Delta x)/\Delta\tau$$

and thus

$$q(x) = (p/2)\big(c(x) - c(x\ +\ \Delta x)\big)\Delta x/\Delta\tau$$

Assuming continuity, we take the Taylor expansion of $c(x\ +\ \Delta x)$:

$$c(x\ +\ \Delta x) =\ c(x) + \left(\frac{\partial c}{\partial x}\right)\Delta x\ + O(\Delta x)^2$$

Furthermore, assuming that Δx is small and omitting the high order terms, we get

$$q(x) = -\frac{p}{2}\frac{\Delta x^2}{\Delta\tau}\left(\frac{\partial c}{\partial x}\right) = -D\left(\frac{\partial c}{\partial x}\right), \qquad D \coloneqq \frac{p}{2}\frac{\Delta x^2}{\Delta\tau} = \frac{\sigma^2}{2} \tag{3.57}$$

where D is the diffusion constant. Notice that when $\Delta x \to 0, \Delta\tau \to 0$, the quantity $\Delta x^2/\Delta\tau$ equals the variance σ^2 of the continuous time process in Digression 3.D, but now it is multiplied by p to take account of the assumption that the particle is possible to stay in the same position ($\underline{v}_\tau = 0$ with probability $1 - p$).

Equation (3.57) is known as the Fick's first law of diffusion. We can also apply to gas molecules' motion, which transfers energy among particles, denoting $q(x)$ the heat (temporal) rate, and substituting temperature T for concentration c, and thermal conductivity κ for the diffusion constant. The resulting relationship, $q(x) = -\kappa(\partial T/\partial x)$, is the Fourier's law.

3.11 The linear Markov process

We will now discuss a more interesting case of filtering of the white noise by means of a stochastic version of a linear differential equation. To establish such an equation, we use a simple hydrological system, a linear reservoir with inflow $v(t)$ and outfow $x(t)$. The reservoir state is characterised by its storage $S(t)$ and the change in outflow (reservoir spill) is assumed (as an approximation) to be proportional to the change in storage, $\mathrm{d}x = \mathrm{d}S/\alpha$, where $\alpha > 0$ is a constant with units of time. The continuity equation (corresponding to mass conservation for an incompressible fluid) is $\mathrm{d}S/\mathrm{d}t = v - x$ and if we make the substitution $\mathrm{d}S = \alpha\mathrm{d}x$ we find that the system dynamics is the first-order linear differential equation:

$$\alpha\frac{\mathrm{d}x(t)}{\mathrm{d}t} + x(t) = v(t) \tag{3.58}$$

Now, let us assume that the inflow is a stochastic process and specifically a white noise process. For convenience we subtract its mean so that $\underline{v}(t)$ has the characteristics given in equation (3.53). The output $\underline{x}(t)$ will be a stochastic process as well. Thus, we can write the stochastic version of equation (3.58) as:

$$\alpha\frac{\mathrm{d}\underline{x}(t)}{\mathrm{d}t} + \underline{x}(t) = \underline{v}(t) \tag{3.59}$$

As simple as may it seem, the transition from the deterministic version in equation (3.58) to the stochastic version in equation (3.59) involves mathematical troubles. In fact, the process $\underline{x}(t)$ is hardly differentiable and the derivative $\mathrm{d}\underline{x}(t)/\mathrm{d}t$ does not generally exist. Thus, stochastic differential equations require their own rules of calculus. Here we use the following simple rule: We solve the differential equation as if it were deterministic with well-defined derivatives. Naturally, the mathematical expression of the solution will not contain derivatives. In that expression we replace the deterministic functions with stochastic processes, thus bypassing the differentiability problem.

In this manner, the linear differential equation (3.59) is easily solved to give:

$$\underline{x}(t) = \underline{x}(0)\mathrm{e}^{-t/\alpha} + \frac{\mathrm{e}^{-t/\alpha}}{\alpha}\int_0^t \underline{v}(u)\mathrm{e}^{u/\alpha}\mathrm{d}u \tag{3.60}$$

We observe in equation (3.60) that:

1. The two additive terms on the right-hand side are independent as the outflow of the present, $\underline{x}(0)$, cannot depend on the future inflows $\underline{v}(u), 0 < u \leq t$.
2. The outflow does depend on the outflow of the present, $\underline{x}(0)$, but not on other $\underline{x}(t)$ of the past ($t < 0$).

A stochastic process that has the latter property is called a Markov process. More generally, a Markov process is one in which the future does not depend on the past once the present is known; symbolically:

$$P\{\underline{x}(t) \leq x | \underline{x}(s) = x(s), s \leq 0 < t\} = P\{\underline{x}(t) \leq c | \underline{x}(0) = x(0)\} \tag{3.61}$$

The particular Markov process $\underline{x}(t)$ of equation (3.60) can be called the linear Markov process and it is also known as *Ornstein–Uhlenbeck* process, while the stochastic differential equation (3.59) is known as the *Langevin equation* (Papoulis, 1991). The mean of the process is:

$$\mathrm{E}[\underline{x}(t)] = \mathrm{E}[\underline{x}(0)]\mathrm{e}^{-t/\alpha} \tag{3.62}$$

Subtracting equation (3.62) from (3.60), squaring and taking expected values we get:

$$\begin{aligned}\mathrm{var}[\underline{x}(t)] &= \mathrm{var}[\underline{x}(0)]\,\mathrm{e}^{-2t/\alpha} + \frac{\sigma^2 D}{\alpha^2}\mathrm{e}^{-2t/\alpha}\int_0^t \mathrm{e}^{2u/\alpha}\mathrm{d}u \\ &= \frac{\sigma^2 D}{2\alpha} + \left(\mathrm{var}[\underline{x}(0)] - \frac{\sigma^2 D}{2\alpha}\right)\mathrm{e}^{-2t/\alpha}\end{aligned} \tag{3.63}$$

From (3.62) and (3.63) we conclude that $\mathrm{E}[\underline{x}(t)]$ and $\mathrm{var}[\underline{x}(t)]$ tend fast (exponentially) to 0 and $\lambda^2 \coloneqq \sigma^2 D/2\alpha$, respectively, regardless of the values $\mathrm{E}[\underline{x}(0)]$ and $\mathrm{var}[\underline{x}(0)]$. In particular, if $\mathrm{E}[\underline{x}(0)] = 0$ and $\mathrm{var}[\underline{x}(0)] = \gamma_0 = \lambda^2$, then the process has constant mean (0) and variance (λ^2) at all times. Interestingly, if we take $D = a$, then $\lambda^2 = \sigma^2/2$, that is it equal the diffusion constant as in Digression 3.D.

It is easily seen that the following equation is a consequence of (3.60):

$$\underline{x}(t+h) = \underline{x}(t)\mathrm{e}^{-h/\alpha} + \frac{\mathrm{e}^{-h/\alpha}}{\alpha}\int_t^{t+h} \underline{v}(u)\mathrm{e}^{u/\alpha}\mathrm{d}u \tag{3.64}$$

Multiplying this equation by $\underline{x}(t)$ and taking expected values we get:

$$c(t,h) = \mathrm{E}[\underline{x}(t+h)\underline{x}(t)] = \mathrm{E}[\underline{x}(t)^2]\mathrm{e}^{-h/\alpha} \tag{3.65}$$

and in the case where $\mathrm{E}[\underline{x}(0)] = 0$, $\mathrm{var}[\underline{x}(0)] = \lambda^2$ this becomes:

$$c(h) = \lambda^2 \mathrm{e}^{-h/\alpha} \tag{3.66}$$

In other words, the autocovariance is a function of the lag h only and the process is wide-sense stationary. The other second-order characteristics of the process in continuous and discrete time, derived through the generic equations contained in Table 3.3, are summarized in Table 3.5 and illustrated in Figure 3.5.

The celebrated linear Markov process is nothing more than filtered white noise through a linear differential equation. The filtering eliminates the problems related to the appearance of infinities discussed in section 3.10 and, thus, it is physically consistent. Furthermore, the simplicity of the equations of its second-order properties makes it attractive and easy to use. On the other hand, its Markovian property, i.e. the independence of the future from the past once the present is known, may contradict our perception that history always influences future developments. We may thus regard it as too simplistic a model of natural reality. Furthermore, the fact that it minimizes entropy production for large times ($t \to \infty$) (Koutsoyiannis, 2011b; see also Digression 3.H) may be another obstacle in accepting it as a good model to represent natural processes.

Table 3.5 Second-order characteristics of the Markov process at continuous and discrete time.

Property	Formula	Eqn. no.
Variance		
Continuous-time process (instantaneous)	$\gamma_0 = \gamma(0) = c(0) = \lambda^2$	(3.67)
Averaged process at scale k (climacogram)	$\gamma(k) = \dfrac{2\lambda^2}{k/\alpha}\left(1 - \dfrac{1-\mathrm{e}^{-k/\alpha}}{k/\alpha}\right)$	(3.68)
Autocovariance function		
Continuous-time, lag h	$c(h) = \lambda^2 \mathrm{e}^{-\lvert h\rvert/\alpha}$	(3.66)
Discrete time, lag $\eta = h/D$	$c_0 = \gamma(D), \qquad c_\eta = \dfrac{\lambda^2\left(1-\mathrm{e}^{-D/\alpha}\right)^2}{(D/\alpha)^2}\,\mathrm{e}^{-(\eta-1)D/\alpha}, \qquad \eta \geq 1$	(3.69)
Power spectrum		
Continuous time, frequency w	$s(w) = \dfrac{4\alpha\lambda^2}{1+(2\pi\alpha w)^2}$	(3.70)
Discrete time, frequency $\omega = wD$	$s_\mathrm{d}(\omega) = 4\alpha\lambda^2\left(1 - \dfrac{\sinh(D/\alpha)}{D/\alpha}\dfrac{1-\cos(2\pi\omega)}{\cosh(D/\alpha)-\cos(2\pi\omega)}\right)$	(3.71)

A discretized Markov process at time step D tends to be uncorrelated in time as D increases. Therefore, at large time scales the Markov model is indistinguishable from white noise: indeed, from equation (3.68) we conclude that for large k (or small α) the variance is inversely proportional to the time scale, as in the white noise. Thus, even though sometimes it is said that the Markov model reflects short-term persistence, it is better not to use the term persistence in this case. Certainly, it entails short-range dependence in time. However, its asymptotic properties (cf. equations (3.46) and (3.47)) are (Koutsoyiannis, 2017):

$$M = \frac{1}{2}, \quad \varphi_\mathrm{C}(0) = \frac{3}{2}, \quad \gamma^\#(0) = \; c^\#(0) = \; 0, \quad \psi^\#(0) = \; 2, \quad s^\#(\infty) = -2 \tag{3.72}$$

$$H = \frac{1}{2},\quad \varphi_C(\infty) = \frac{1}{2},\quad \gamma^{\#}(\infty) = -1,\quad c^{\#}(\infty) = -\infty,\quad \psi^{\#}(\infty) = s^{\#}(0) = 0$$

Thus, according to the classification of section 3.9, the process is neutral: neither antipersistent nor persistent and neither rough nor smooth.

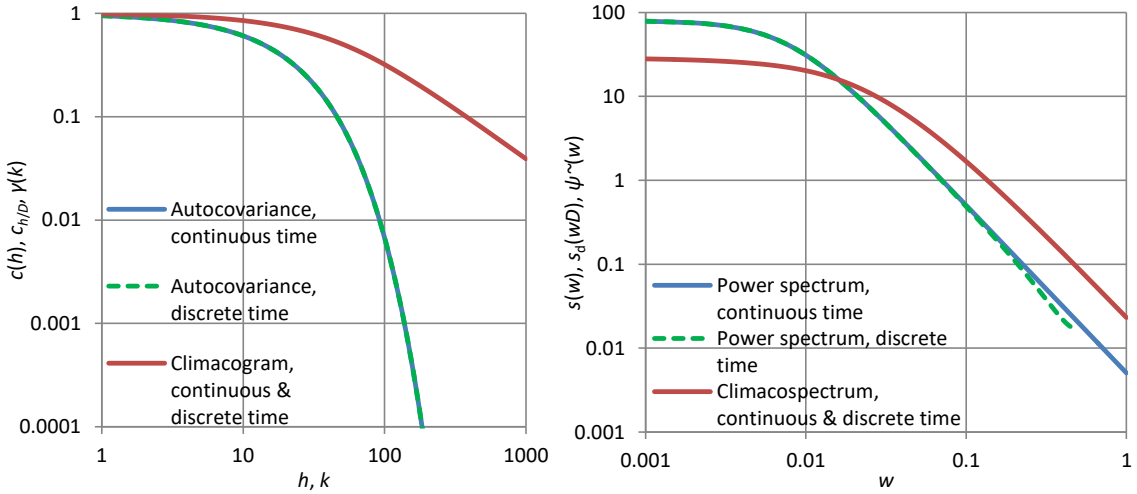

Figure 3.5 Second-order characteristics of a linear Markov process with parameters $\lambda = 1, \alpha = 20$ and discretization time step $D = 1$. The climacogram and climacospectrum are precisely the same for the continuous- and discrete-time representations. The autocovariance and the power spectrum have some differences between the two representations which are invisible in the former case and visible in the latter.

While the linear differential equation, on which the introduction of the Markov model has been based, has some physical basis, the assumption that the inflow is white noise is physically problematic, as we clarified in section 3.10. This is another reason making the simple Markov model inappropriate for natural systems, except in simple systems where independence is justified. Even though it is rarely noticed, this problem is also encountered in most cases of stochastic differential equations that are deterministic equations perturbed by white noise.

Related to the Markov process in continuous time is the discrete-time process:

$$\underline{x}_\tau = a\underline{x}_{\tau-1} + \underline{v}_\tau + b\underline{v}_{\tau-1} \tag{3.73}$$

commonly known as ARMA(1,1), which stands for autoregressive–moving-average process of orders (1,1). Here $\underline{v}_\tau$ is discrete-time white noise with variance σ_v^2, and a and b are parameters. It can be easily shown (homework) that its second-order characteristics are interrelated by:

$$c_0 = \left(1 + \frac{(a+b)^2}{1-a^2}\right)\sigma_v^2, \qquad c_1 = ac_0 + b\sigma_v^2, \qquad c_\eta = a^{\eta-1}c_1,\ \ \eta \geq 1 \tag{3.74}$$

Comparing with equation (3.69) we see that the ARMA(1,1) process is identical to the discrete-time representation of the Markov process if we choose:

$$a = \mathrm{e}^{-D/\alpha}, \qquad c_0 = \gamma_1 = \frac{2\lambda^2}{D/\alpha}\left(1 - \frac{1-\mathrm{e}^{-D/\alpha}}{D/\alpha}\right), \qquad c_1 = \frac{\lambda^2\left(1-\mathrm{e}^{-D/\alpha}\right)^2}{(D/\alpha)^2} \tag{3.75}$$

Alternatively, if we know the first three terms of the autocovariance function in discrete time, then, without referring to the continuous time formulation, the parameter a can be found as the ratio

$$a = c_2/c_1 \tag{3.76}$$

The remaining parameters b and σ_v^2 can be found from the first two equations in (3.74) in terms of $c_0 \equiv \gamma_1$ and c_1. A rather involved but explicit solution can also be found (homework).

The special case in which:

$$b = 0 \Leftrightarrow c_1/c_0 = a \tag{3.77}$$

is known as the AR(1) process, standing for autoregressive process of order 1. This is the limiting case as $D/\alpha \to 0$. It can also appear in a discrete-time representation of the Markov process for finite time step D if, instead of time averages, we use instantaneous quantities, thus obtaining the so-called *sampled process* defined in discrete time as:

$$\underline{x}_\tau \coloneqq \underline{x}(\tau D) \tag{3.78}$$

(compare this with (3.12)). The AR(1) process is thus:

$$\underline{x}_\tau = a\underline{x}_{\tau-1} + \underline{v}_\tau \tag{3.79}$$

and its second-order characteristics are:

$$c_0 = \frac{\sigma_v^2}{1-a^2}, \qquad c_\eta = a^{|\eta|}c_0, \qquad \gamma^{(\kappa)} = \frac{\lambda}{\kappa\,(1-a)^2}\left(1 - a^2 - \frac{2a(1-a^\kappa)}{\kappa}\right) \tag{3.80}$$

It is worth mentioning one more process of similar type, the AR(2) process, which is:

$$\underline{x}_\tau = a_1\underline{x}_{\tau-1} + a_2\underline{x}_{\tau-2} + \underline{v}_\tau \tag{3.81}$$

and has second-order characteristics interrelated by:

$$c_0 = a_1c_1 + a_2c_2 + \sigma_v^2, \qquad c_1 = a_1c_0 + a_2c_1, \qquad c_\eta = a_1c_{\eta-1} + a_2c_{\eta-2}, \ \ \eta \geq 1 \tag{3.82}$$

Once the covariances c_0, c_1, c_2 are known (estimated from data or derived theoretically) the three parameters a_1, a_2, σ_v^2 can be easily found as the system of equations is linear. Additional information about similar discrete time processes is given in Digression 3.F.

Digression 3.F: The Time Series School and its processes

The AR(1), AR(2) and ARMA(1,1) processes discussed in section 3.11 are representatives of bigger families of models developed within the *Time Series School*. Obviously, higher order AR and ARMA models can be formulated, and actually are in common use, along with additional families such as ARIMA(*p*,*d*,*q*) (standing for autoregressive integrated moving average models) and ARFIMA(*p*,*d*,*q*) (with the additional 'F' standing for fractional). Equations such as (3.82) are called *Yule – Walker equations* as they were introduced by Yule (1927) and Walker (1931), both British

statisticians who, starting from an analysis of sunspot numbers, studied autoregressive processes and in particular their periodogram and autocorrelation properties. However, we will not refer to this type of processes here, preferring to base our analyses on the *Stochastic School*, pioneered by A. Kolmogorov, which provides more solid grounds, both for foundation and application, than the Time Series School. As discussed in Koutsoyiannis (2025), useful tasks such as the application of stochastics in simulation can be undertaken in a generic and simple manner without any reference to the non-parsimonious models of the Time Series School.

We should note, however, that the Time Series School and its models are much more popular than the Stochastic School in many disciplines, including geophysics. It appears that the former was initiated by the American economist W.M. Persons. In studying the problem "*When to buy or sell*", Persons (1919) introduced the study of time series, which he called *statistical series*, and asserted that they "*result from the combination of four elements: secular trend, seasonal variation, cyclical fluctuation, and a residual factor.*" He also proposed methods for "*Eliminating secular trends*" and "*Eliminating seasonal variation*". Interestingly, the Ukrainian/Russian/Soviet mathematical statistician and economist Slutsky (1927) demonstrated that what Persons (and other economists) regarded as cyclical component is only a meaningless statistical artefact (see e.g. Kyun and Kim 2006; Barnett, 2006). Subsequently, the notion of a cyclical component was abandoned but the decomposition of a time series into the three other components, trends, seasonal variation and residuals remains popular.

Perhaps the first definition of a time series was given by the American statistician Bailey (1929):

> *A time series is a series of observations taken at different times and recorded with the time at which they were taken.*

The biggest progress in the Time Series School was made in Uppsala by the Norwegian-born (with career in Sweden) econometrician and statistician H.O.A. Wold and the New-Zealand-born mathematician and statistician P. Whittle, who in their doctoral theses provided the stochastic foundation for time series analysis. Wold (1938, 1948) proved that a stochastic process (even though he referred to it as a time series) can be decomposed into a *regular process* (i.e., a process linearly equivalent to a white noise process) and a *predictable process* (i.e., a process that can be expressed in terms of its past values). This has been known as *Wold's decomposition*. Whittle (1951, 1952, 1953) laid the mathematical foundation for autoregressive and moving average models in univariate and multivariate settings. Later, in their influential book, Box and Jenkins (1970) named these models with the above acronyms and they became popular with these names and also with the name *Box – Jenkins models* (cf. Stigler's law of eponymy, which states that no scientific discovery is named after its original discoverer; Stigler, 2002).

Despite the wider influence of the Time Series School over the Stochastic School, there are several problems with the former. First, the term *time series* is ambiguous, sometimes denoting a series of observations as in the original definition of Bailey (1929) (or, equivalently, a realization of a stochastic process), and at other times denoting the stochastic process per se (as in the aforementioned use by Wold). As we have already emphasized, here the term *time series* is used with the first meaning, a series of numbers, while for a series of stochastic variables we use the term stochastic process. Second, with the exception of the simplest models of these families, such as the AR(1) and ARMA(1,1), time series models are too artificial because, being complicated discrete-time models, they do not necessarily correspond to a continuous time process, while natural processes evolve in continuous time. Furthermore, their identification, typically based on the estimation of the autocorrelation function from data (see Digression 4.C for an explanation), usually neglects estimation bias and uncertainly, which in stochastic processes (as opposed to purely random processes) are often tremendous (Lombardo et al., 2014).

Indeed, from the outset (Whittle, 1952), time series models have been closely associated with a large number of parameters, and they usually become over-parameterized and thus not parsimonious. These parameters are estimated from data, which usually are too few to support a reliable estimation. The decomposition of a time series to components, trends, seasonal variation and residuals, is fundamentally problematic, despite being popular. Remarkably, a meaningful

definition of a trend has never been given. It is also hard to fathom how *time per se* could be regarded as an explanatory variable for a complex process and on what logical basis the statistics of a physical process could be expressed as a deterministic function of time. Accumulation of data series with long time spans (cf. Koutsoyiannis, 2025) has shown that, what have been regarded as trends, are mostly parts of long-term fluctuations (and in accord with Slutsky's work, they could also be regarded as statistical artefacts). Finally, "deseasonalization" (or in Persons's original terminology "*Eliminating seasonal variation*") is a delusion. We can hardly remove seasonality in the multivariate distribution of a stochastic process. What we typically do is in the marginal distribution only and so there is no elimination.

3.12 The Hurst-Kolmogorov process

The Hurst-Kolmogorov (HK) process in its continuous-time version is defined through its climacogram:

$$\gamma(k) = \lambda^2 \left(\frac{\alpha}{k}\right)^{2-2H} \tag{3.83}$$

By setting $H = 1/2$ we recover equation (3.51), which means that the HK process is a generalization of the white noise. Its other second-order characteristics are given in Table 3.6 and illustrated in Figure 3.6. Their LLDs are constant for all time lags and scales and all frequencies:

$$\varphi(k) = \varphi_{\mathrm{C}}(k) = H, \;\; \gamma^{\#}(k) = c^{\#}(h) = 2H - 2, \;\; \psi^{\#}(k) = -s^{\#}(w) = 2H - 1 \tag{3.84}$$

including their asymptotic values at 0 and ∞. Accordingly, $M = H - 1$.

Table 3.6 Second-order characteristics of the HK process at continuous and discrete time.

Property	Formula	Eqn. no.
Variance		
Continuous-time process (instantaneous)	$\gamma_0 = \gamma(0) = c(0) = +\infty$	(3.85)
Averaged process at scale k (climacogram)	$\gamma(k) = \lambda^2(\alpha/k)^{2-2H}$	(3.83)
Autocovariance function		
Continuous-time, lag h	$c(h) = \begin{cases} \lambda^2 H(2H-1)\left(\frac{\alpha}{h}\right)^{2-2H}, & H > \frac{1}{2} \\ \lambda^2\,\delta\left(\frac{h}{\alpha}\right), & H = \frac{1}{2} \\ \lambda^2 H(2H-1)\left(\frac{\alpha}{h}\right)^{2-2H} + \delta''\left(\frac{h}{\alpha}\right), & H < \frac{1}{2} \end{cases}$	(3.86)
Discrete time, lag $\eta = h/D$	$c_\eta = \lambda^2(\alpha/D)^{2-2H}\left(\frac{\lvert j-1\rvert^{2H} + \lvert j+1\rvert^{2H}}{2} - \lvert j\rvert^{2H}\right)$ (for $\eta > 2,\; c_\eta \approx H(2H-1)j^{2H-2}\gamma_1,\; \gamma_1 = \lambda^2(\alpha/D)^{2-2H}$)	(3.87)
Power spectrum[1]		
Continuous time, frequency w	$s(w) = \frac{2\alpha\lambda^2\Gamma(2H+1)\sin(\pi H)}{(2\pi\alpha w)^{2H-1}}$	(3.88)

[1] The power spectrum of the discrete-time (averaged) process exists (it is finite for $w > 0$) but it does not have a closed expression. However, for small frequencies ($\omega = wD < 0.1$), the continuous-time expression is a very good approximation for the discrete-time process, i.e. $s_{\mathrm{d}}(\omega) \approx s(\omega/D)$.

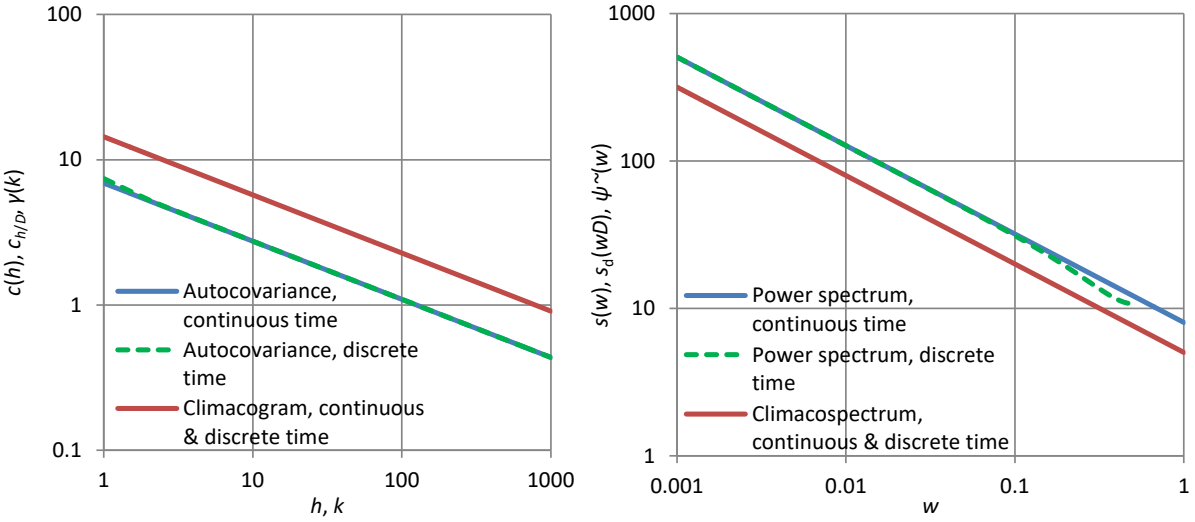

Figure 3.6 Second-order characteristics of a HK process with parameters $\lambda = 1, \alpha = 20, H = 0.8$ and discretization time step $D = 1$. The climacogram and climacospectrum are precisely the same for the continuous- and discrete-time representations. The autocovariance and the power spectrum have some differences between the two representations, which are visible in both cases.

The Gaussian version of the process is also known as *fractional Gaussian noise* (FGN) due to Mandelbrot and van Ness (1968), although these authors used a more complicated approach to define it. Here we do not use the term FGN as the adjective *fractional* is not quite informative (even white noise, in which $H = 0.5$, is fractional too), the adjective *Gaussian* is too restrictive (non-Gaussian HK are also common; see Koutsoyiannis, 2025) and the noun *noise* is too negative and perhaps misleading when we try to describe Nature's processes. As already mentioned, a variant of that mathematical process had been proposed earlier by Kolmogorov (1940), while Hurst (1951) pioneered the detection in geophysical time series of the behaviour described by this process. Hence the name HK, which we use for this process.

Because this process has infinite instantaneous variance, the sampled process in discrete time is not meaningful (many characteristics take infinite values). However, the averaged process is well behaving with all its characteristics (including its variance) finite. This makes it quite useful in applications, once we exclude the very small scales.

The HK process is almost as simple and parsimonious as the Markov process; again, it contains only one parameter, H, in addition to those describing its marginal distribution. Notice that the process variance is controlled by the product $\lambda^2\ \alpha^{2-2H}$, so that λ and α, are not in fact separate parameters. Despite that, we prefer the formulation shown in Table 3.6 with three nominal parameters for dimensional consistency: α and λ are scale parameters with dimensions $[t]$ and $[x]$, respectively, while H, the Hurst coefficient, is dimensionless in the interval (0, 1).

For $H = 1/2$ the process reduces to pure white noise. For $1/2 < H < 1$ the process is persistent and for $0 < H < 1/2$ antipersistent. Most of the expressions shown in Table 3.6 hold in all three cases. However, the autocovariance $c(h)$ has different expressions in the three cases, as shown in Table 3.6. Specifically, for $H < 1/2$, the autocovariance $c(h)$is

negative for any lag $h > 0$, tending to $-\infty$ as $h \to 0$. However, at $h = 0, c(0) = +\infty$, because this is the variance of the process which cannot be negative. In other words, there is an infinite discontinuity at $h = 0$. Consequently, the averaged process has positive variance and all covariances negative. Such a process is not physically realistic because real-world events at near times are always positively correlated, which means that for small h, $c(h)$ should be positive. Also, the infinite variance cannot appear in Nature. Thus, the HK process can describe natural phenomena only for $1/2 < H < 1$ and for time scales that are not too small. Furthermore, values $H > 1$ that are sometimes reported in the literature are mathematically invalid (Koutsoyiannis, 2014b, 2017; Koutsoyiannis et al., 2018; see also Figure 3.4). They are the results of inconsistent algorithms. In terms of entropy production, the process maximizes it for large times ($t \to \infty$) but minimizes it for small times ($t \to 0$).

Digression 3.G: Developments in stochastic modelling in geophysics before and after Hurst

Hurst's (1951) discovery of the natural behaviour named after him was triggered by a real-world problem of engineering hydrology, the design of reservoirs. This gave hydrology a central role in understanding such behaviour and subsequently in the dissemination process to other disciplines. Hydrology has mostly been an importer of stochastic methods from other fields, but Hurst's research has been an exception marking a large-scale "export" to other fields (O'Connell et al., 2016).

The understanding that hydrological processes could not be modelled effectively by deterministic techniques did precede Hurst's research. Techniques that could be classified as applications of the Monte Carlo method had appeared in the hydrological literature much earlier than the "official start" of the Monte Carlo method in 1949 and of Hurst's (1951) paper. Hazen (1914) did a pioneering study in which he introduced the reservoir storage-yield-reliability relationship, a concept that would remain unexploited in Western hydrological literature, even though it offered a scientific basis of modern reservoir design (Klemeš, 1987). In that study he proposed an empirical simulation technique and formed a synthetic time series by combining historical flow records of different rivers 'spliced' sequentially together. Sudler (1927) extended the work of Hazen by resampling from a sequence of historical river flows using cards, which he shuffled to form new sequences of data. Obviously, this method heavily distorts the time dependence of river flows whose importance was not known at that time.

For it was Hurst (1951) who understood that importance along with the omnipresence in natural processes of a clustering behaviour of similar events in time, a behaviour that is now understood as (long-term) persistence, long-range dependence (LRD) or Hurst-Kolmogorov dynamics. In his attempt to compare natural and random events, Hurst performed physical experiments to generate random numbers. Specifically, he tossed 10 coins (sixpences) simultaneously and repeated this 1025 times (note that 10 binary digits are equivalent to about 3 decimal digits). As he notes, his rate was 100 random numbers per 35 minutes (while that would be of the order of a microsecond in modern computer environments, even slow ones). He also used another method, shuffling and cutting a pack of 52 cards, in which he improved the rate to 100 random numbers per 20 min.

The behaviour discovered by Hurst is now known to many disciplines, most prominently in information sciences, biological and medical sciences, economics and finance, and geophysical sciences—except the so-called climate science where it is hardly known. Even within the hydrological community it took decades before Hurst's discovery of persistence was assimilated (O'Connell et al., 2016). Thus, the initial studies implementing primitive variants of stochastic simulation did not reproduce LRD. Barnes (1954), in designing a reservoir in Australia, used a table of random numbers from normal distribution to generate a 1000-year sequence of synthetic

annual data. Thomas and Fiering (1962) generated flows correlated in time, but using only the lag-one autocorrelation, obviously neglecting LRD. Beard (1965) and Matalas (1967) generated concurrent flows at several sites. Chow (1969), and Chow and Kareliotis (1970) systematized the use of time series models (in particular—and using their terminology—moving average models, sum of harmonics models and autoregression models) and highlighted their value in the economic planning of water supply and irrigation projects. It is evident from the above pioneering studies, as well as from the myriads of subsequent studies, that geophysicists have followed (and today still do) the Time Series School rather than the more rigorous Stochastic School.

3.13 The Filtered Hurst-Kolmogorov process

The HK process should not be regarded as a model of general validity, but one that it is valid for large scales—and indeed, we shall use it since it is more physically plausible than processes with exponential decrease of autocovariance (e.g. the Markov process). To this end, we can filter HK appropriately to make it a physically consistent process for all scales. This is the same as what we did to the white noise to make it physically consistent by removing infinities.

Similar to the white noise process, if we filter an input $\underline{v}(t)$ that is now an HK process, either by a moving average filter or by a linear differential equation system, then it is easy to see that the filtered output is a physically realistic process with finite variance $\gamma(0)$, practically an unaffected climacogram $\gamma(k)$ at large scales with $\gamma^{\#}(\infty) = 2H - 2$ (as in the original HK process), but a highly modified climacogram at small scales, thus having a valid structure with $M = \varphi_C(0) - 1 = (\psi^{\#}(0) - 1)/2 = H$.

However, to enrich the process we can make the parameter M independent of H, thus making it more flexible in order to model real world data. For the model application it is not necessary to specify the linear filter needed to convert the HK process into a filtered Hurst-Kolmogorov (FHK) process. (In some cases this would be too involved). It suffices to specify a convenient expression of the climacogram. Below we provide three such expressions (from Koutsoyiannis, 2017). All expressions contain the dimensionless parameters M and H with the meaning and values discussed in section 3.8.

1. The generalized Cauchy-type (FHK-C) climacogram:

$$\gamma(k) = \lambda^2(1 + (k/\alpha)^{2M})^{\frac{H-1}{M}} \tag{3.89}$$

2. The generalized Dagum-type (FHK-D) climacogram:

$$\gamma(k) = \lambda^2\left(1 - \left(1 + (k/\alpha)^{2(H-1)}\right)^{\frac{M}{H-1}}\right) \tag{3.90}$$

3. The composite Cauchy-Dagum-type (FHK-CD) climacogram, derived by summing an FHK-C with $M = 1$ and an FHK-D with $H = 0$:

$$\gamma(k) = \lambda_1^2(1 + (k/\alpha_1)^2)^{H-1} + \lambda_2^2(1 - (1 + (k/\alpha_2)^{-2})^{-M}) \tag{3.91}$$

4. A second form of FHK-CD (FHK-CD2), derived by summing an FHK-C with $M = 1/2$ and an FHK-D with $H = 1/2$:

$$\gamma(k) = \lambda_1^2(1 + k/\alpha_1)^{2H-2} + \lambda_2^2(1 - (1 + \alpha_2/k)^{-2M}) \tag{3.92}$$

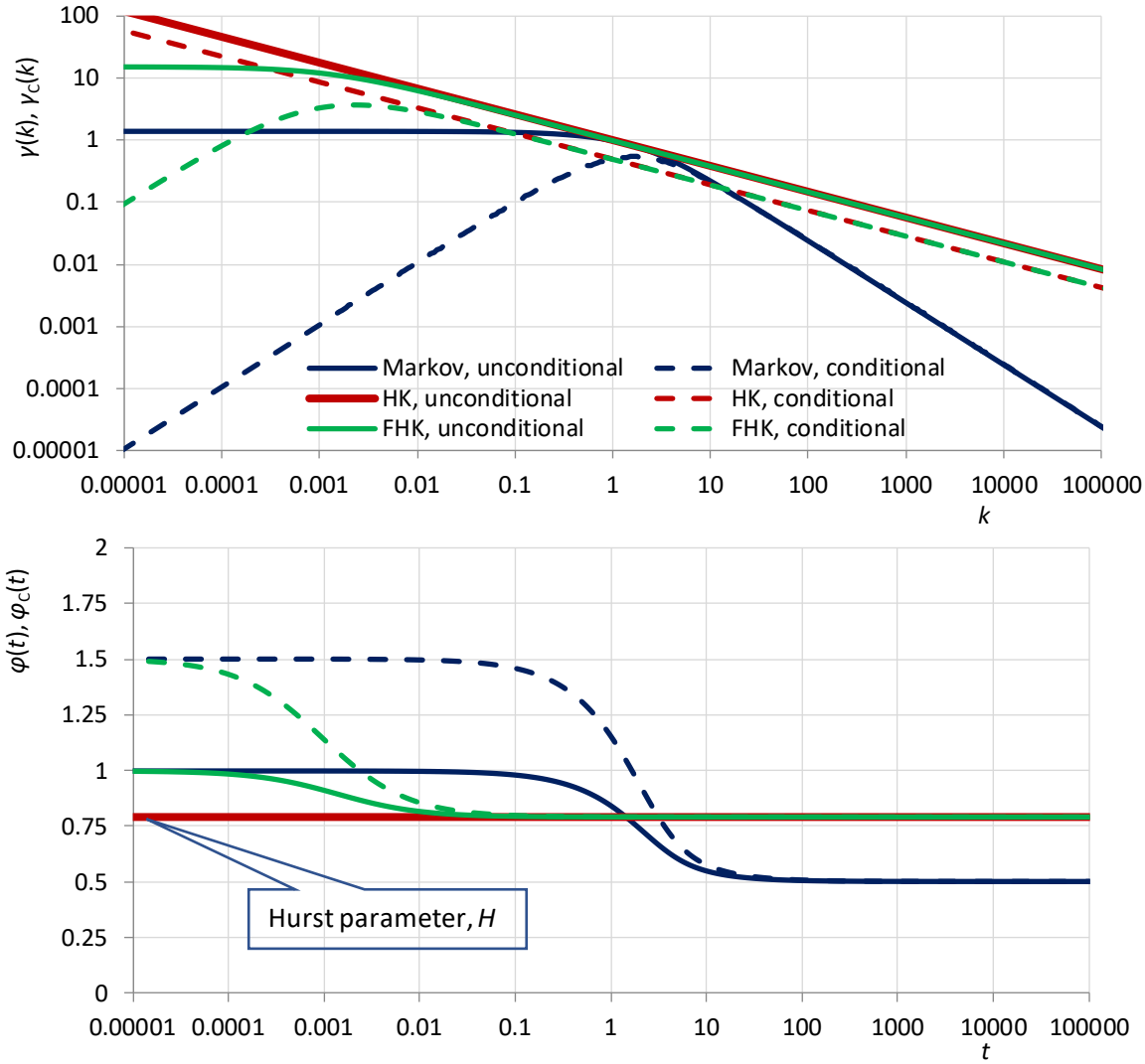

Figure 3.7 (**upper**) Climacograms and (**lower**) EPLT ($\varphi(t)$) and CEPLT ($\varphi_C(t)$) of the three indicated example processes for neutral smoothness ($M = 0.5$). At time scale $D = 1$ all three processes have the same variance $\gamma(1) = 1$ and the same autocovariance for lag 1, $c_1^{(1)} = 0.5$. Their parameters are: for the linear Markov process $a = 0.8686$, $\lambda = 1.4176$; for the HK process $a = 0.0013539$, $\lambda = 15.5032$, $H = 0.7925$ (equivalently, $a = \lambda = 1$ but the former parameter set was preferred in order to be comparable to the FHK); for the FHK process $a = 0.0013539$, $\lambda = 15.5093$, $M = 0.5$, $H = 0.7925$. In the lower graph conditional and unconditional HK coincide (adapted from Koutsoyiannis, 2016).

FHK-CD in either of the variants (3.91) and (3.92), is most convenient, as the first additive term determines only the persistence of the process and the second one the smoothness of the process. In addition, it is more flexible and richer than its constituents, as it contains two couples of scale parameters. However, if parsimony is sought, then it can take the same number of parameters as each of the constituents by setting $\alpha_1 = \alpha_2 = \alpha$ and $\lambda_1 = \lambda_2 = \lambda$ (note that, for dimensional consistency, λ and α are minimal parameter requirements).

In the special case $M = 1 - H$ both FHK-C and FHK-D result in the same expression:

$$\gamma(k) = \frac{\lambda^2}{1 + (k/\alpha)^{2(1-H)}} \tag{3.93}$$

For large k/a the FHK process tends to the HK one. This is illustrated in Figure 3.7, where, in addition, the linear Markov model (for the same value of the lag-one autocovariance) is plotted for comparison. We notice that, as time tends to zero, the Markov and the FHK models have the same entropy production while the HK model is associated with minimal entropy production. For intermediate times the Markov model gives higher entropy production than the other two models, but this is done at the "expense" of giving too low entropy production at large time scales, at which both the HK and the FHK give precisely the same high entropy production.

Digression 3.H: Entropy production and time series patterns

The different patterns in time series generated by different M and H (specifically for the Cauchy-type climacogram) are illustrated in the plots of Figure 3.8, also in comparison with two other models, the white noise (panel (a)) and the linear Markov model (panel (b)). These two serve as useful benchmark models for comparisons: the former is free of patterns as it reflects pure randomness, and the latter is fully neutral (neither rough nor smooth as $\varphi_C(0) = 3/2$, and neither antipersistent nor persistent as $\varphi_C(\infty) = 1/2$).

The time series plotted in Figure 3.8 were generated by the symmetric moving average (SMA) scheme described in Koutsoyiannis (2025) with 1024 coefficients (weights) a. In all cases the discretization time scale is $D = 1$, the characteristic time scale $\alpha = 10$, and the characteristic variance scale λ is chosen so that for time scale D, $\gamma(D) = 1$. The mean is 0 in all cases and the marginal distribution is normal. The FHK is implemented using the Cauchy-type climacogram. Each of the panels shows the first fifty terms of time series produced by each of the model implementations at time scales $k = 1$ and 20. In addition, each panel contains a "stamp" of the specific model represented by the plot of CEPLT, $\varphi_C(k)$. In this way the time series patterns can be connected to the entropy production of the generating mechanism.

In panel (c) the CEPLT is close to the absolute maximum both for small and large scales ($H = M = 0.97$ so as to obtain $\varphi_C(0) = 1.97 \approx 2$ and $\varphi_C(\infty) = 0.97 \approx 1$); notable is the very smooth shape at scale 1 and the large departures from the mean (which is 0) at scale 20. On the contrary, in panel (d) the CEPLT is close to the absolute minimum for all scales ($H = M = 0.05$, so as to obtain $\varphi_C(0) = 1.05 \approx 1$ and $\varphi_C(\infty) = 0.05 \approx 0$—for better visualization it was preferred not to use values of H and $M < 0.05$). Furthermore, in panel (e) the CEPLT is close to the absolute maximum for large scales ($H = \varphi_C(\infty) = 0.99 \approx 1$) and close to the absolute minimum for small scales ($M = 0.01$ resulting in $\varphi_C(0) = 1.01 \approx 1$). Finally, in panel (f) the conditions are opposite to those in (e) i.e., the CEPLT is equal to the absolute minimum for large scales ($H = \varphi_C(\infty) = 0.01 \approx 0$) and to the absolute maximum for small scales ($M = 0.99$ resulting in $\varphi_C(0) = 1.99 \approx 2$).

The particular case of panel (e) is close to what is usually called "pink noise" or "$1/f$ noise", as the power spectrum has almost constant slope −1 for the entire frequency domain (which is the same in the climacospectrum). This means that using the FHK model we can theoretically represent and practically produce even "pink noise" in a consistent stationary setting without linking it to a nonstationary process (Keshner, 1982; Wornell, 1993), which involves several theoretical inconsistencies. Indeed, the small change of slope from 0.99 to 1.01 is not actually visible, especially considering the very rough shape of the empirical periodogram, which certainly cannot support differentiation between 0.99 and 1. The FHK model can be used also in other ways to produce "pink noise", that is, by selecting a very large (small) parameter α so as to expel from our field of vision the asymptotic behaviour on large (small) scales. And we can imagine that in several cases of empirical exploration using observations of natural processes, the observation resolution and length, compared to characteristic scale(s) of the process, would be such as to hide the asymptotic behaviour of the process. We can use this as a trick to obtain virtually constant

power spectrum slopes much steeper than −1. Specifically, we can use a large α that does not allow viewing of the asymptotic behaviour at low frequencies or large scales and the slope (see example in Koutsoyiannis, 2017). But this should not mislead us into interpreting the steep slopes as indicators of nonstationarity (see Digression 3.C).

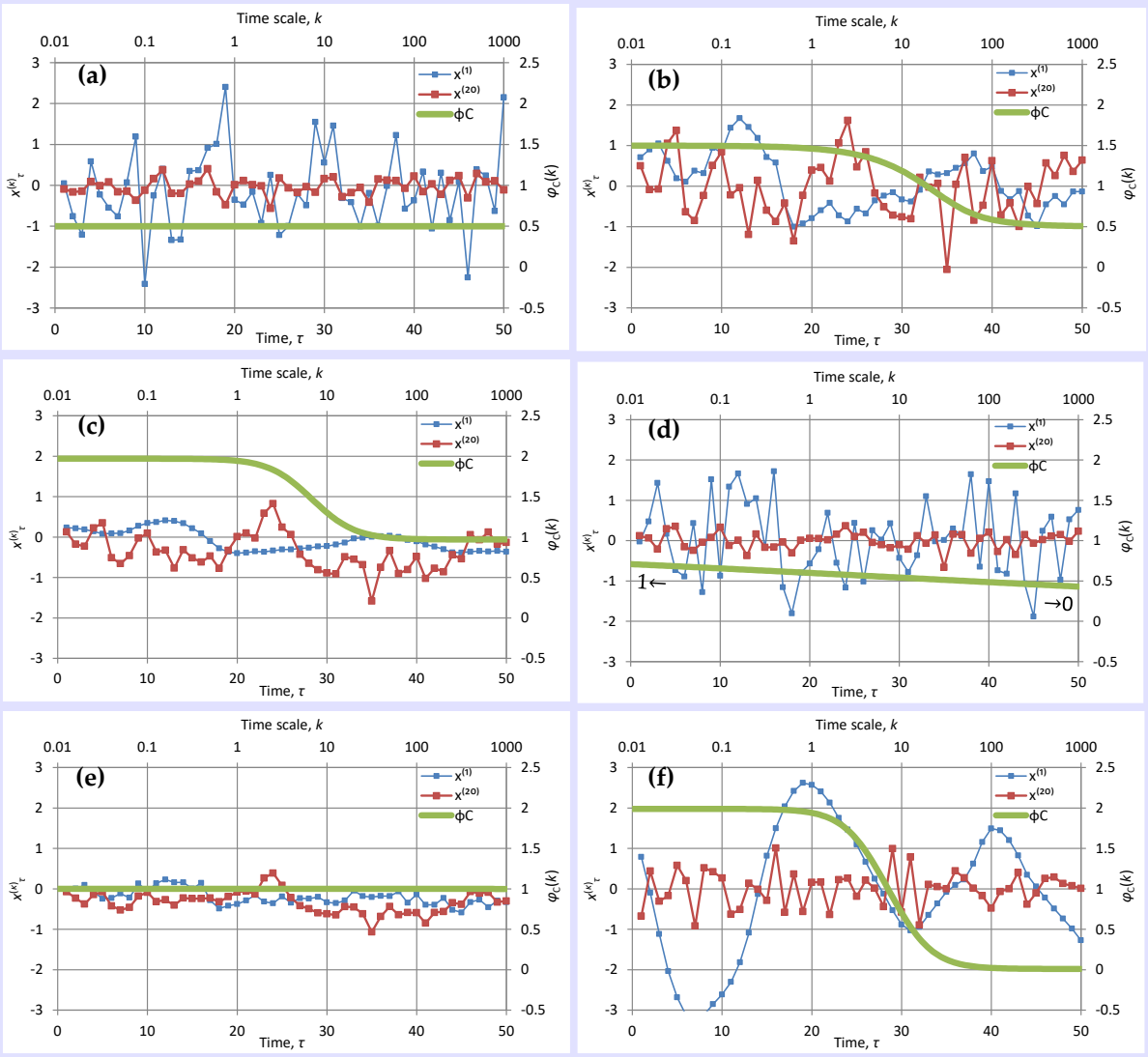

Figure 3.8 The first fifty terms at time scales $k = 1$ and 20 of time series produced by various models, along with "stamps" of the models (thick lines plotted with respect to the right vertical axes) represented by the CEPLT, $\varphi_C(k)$. The different models are **(a)** white noise; **(b)** Markov; **(c)** FHK, with CEPLT close to the absolute maximum ($H = M = 0.97$); **(d)** FHK, with CEPLT close to the absolute minimum ($H = M = 0.05$; notice the slow convergence of $\varphi_C(k)$ to the limiting values 0 and 1); **(e)** FHK, with CEPLT close to the absolute maximum for large scales ($H = 0.99$) and close to the absolute minimum for small scales ($M = 0.01$); **(f)** FHK with CEPLT close to the absolute minimum for large scales ($H = 0.01$) and to the absolute maximum ($M = 0.99$) for small scales. (Source: Koutsoyiannis, 2025.)

3.14 Dependence and behaviour of extremes

When we study extremes, we are usually satisfied by specifying the marginal distribution. As analysed in Chapter 2, this is generally sufficient for design purposes, where the design is based upon the concept of return period (see Digression 2.I). In this respect, the dependence structure of the process of interest may not affect the design procedure per se. However, the dependence in a stochastic process substantially alters the temporal distribution of extremes. In a process with dependence there are patterns. Specifically,

there are periods with clusters of extremes and periods without extremes or with infrequent extremes. We should thus adapt our perception of the behaviour of extremes to become consistent with this reality. Without such adaptation our perception will probably continue to be guided by the "roulette-wheel" paradigm, in which there are no patterns.

There is an additional, more severe, consequence of the presence of dependence. Geophysical studies necessarily rely on data to make inferences. Data records are usually inadequate, becoming worse when there are extremes. The latter problem also affects the specification of the marginal distribution. This is illustrated by a simulation experiment in Digression 3.I.

Digression 3.I: Relationship of persistence and distribution upper tail

To illustrate whether or not (and how) the persistence (or long-range dependence or just change) affects the estimation of the marginal distribution of a discrete-time stationary process $\underline{x}_\tau$ we perform a simulation experiment. We assume that the marginal distribution of $\underline{x}_\tau$ is exponential: $f_{\underline{x}}(x|\lambda) = \lambda\, \mathrm{e}^{-\lambda x}$. Further, we make two alternative assumptions:

(a) that the parameter λ is constant, $\lambda = 5$, and
(b) that λ varies slowly with mean $\mu_\lambda = 5$ and standard gamma distribution, $f_{\underline{\lambda}}(\lambda) = \lambda^{\zeta-1}\mathrm{e}^{-\lambda}/\Gamma(\zeta)$ with $\zeta = \mu_\lambda = 5$.

To simulate a slowly varying λ we initially generate a time series of a stochastic process $\underline{\lambda}'$ with the same distribution as $\underline{\lambda}$ from the HK process with a high $H = 0.95$. Then we form a time series of $\underline{\lambda}$ with the rule $\lambda_i = \lambda_i'$ with probability $1/100$, otherwise $\lambda_i = \lambda_{i-1}$. The latter rule guarantees that each value λ_i lasts on average for 100 time units. The HK process used for λ_i' guarantees that there is change on all scales, not just at scale 100. Koutsoyiannis (2004a) has shown that the unconditional distribution of x in this case is Pareto rather than exponential, i.e. $f_x(x) = \zeta\,(1+x)^{\zeta-1}$.

With either of the two alternatives, once λ is known at time step τ, we generate x_τ from the exponential distribution independently of previous and next x_τ. In alternative (a), the resulting process will be white noise. However, in alternative (b), the change of the parameter induces dependence, while the process $\underline{x}_\tau$ remains stationary (because the change is stochastic, resisting a deterministic description).

Figure 3.9 (upper row) depicts two time series x_τ, each with length 10 000, generated with alternatives (a) (left panel) and (b) (right panel). Moving averages for a time scale of 500, also plotted in the two panels, indicate the absence of patterns (pure randomness, white noise) in alternative (a) and the long-range dependence (not nonstationarity) in alternative (b).

Now let us assume that this time series represents a hypothetical geophysical process on an annual scale. Let us further assume that a researcher has a record of fewer than 100 observations. Most probably all of these refer to the same value of the parameter λ_i. Consequently, the researcher would diagnose that:

- the process behaves like white noise—and indeed, the slope of the climacogram (Figure 3.9, lower right) for scales < 10 (one tenth of the sample size) is –1;
- the marginal distribution is exponential—because it indeed is exponential conditionally on a single value of λ.

The two distributions for constant and varying λ (cases (a) and (b)) are shown in the bottom-left panel of Figure 3.9, along with the distribution of λ in case (b), as empirically derived from the simulations. The adoption of the former underestimates the design quantities for large return periods. Furthermore, the bottom-right panel shows the dramatic differences in the climacograms of the two cases. The climacogram in case (b) starts with a slope –1 for scales < 10, but for large

scales this becomes −0.33, suggesting $H = 0.84$. The varying slope is consistent with the findings of Markonis and Koutsoyiannis (2016) for the rainfall process. Overall, this simulation experiment shows two things.

- Long series are needed to diagnose natural behaviours and in particular the multi-scale change in natural processes.
- The mechanisms producing change may also lead to thickening of the distribution upper tail, thus enhancing the occurrence probability or the intensity of extremes.

These effects are particularly important when we study maxima, neglecting the small values (below a high threshold), a practice that tends to hide the existence of long-range dependence even in long records (see Iliopoulou and Koutsoyiannis, 2019).

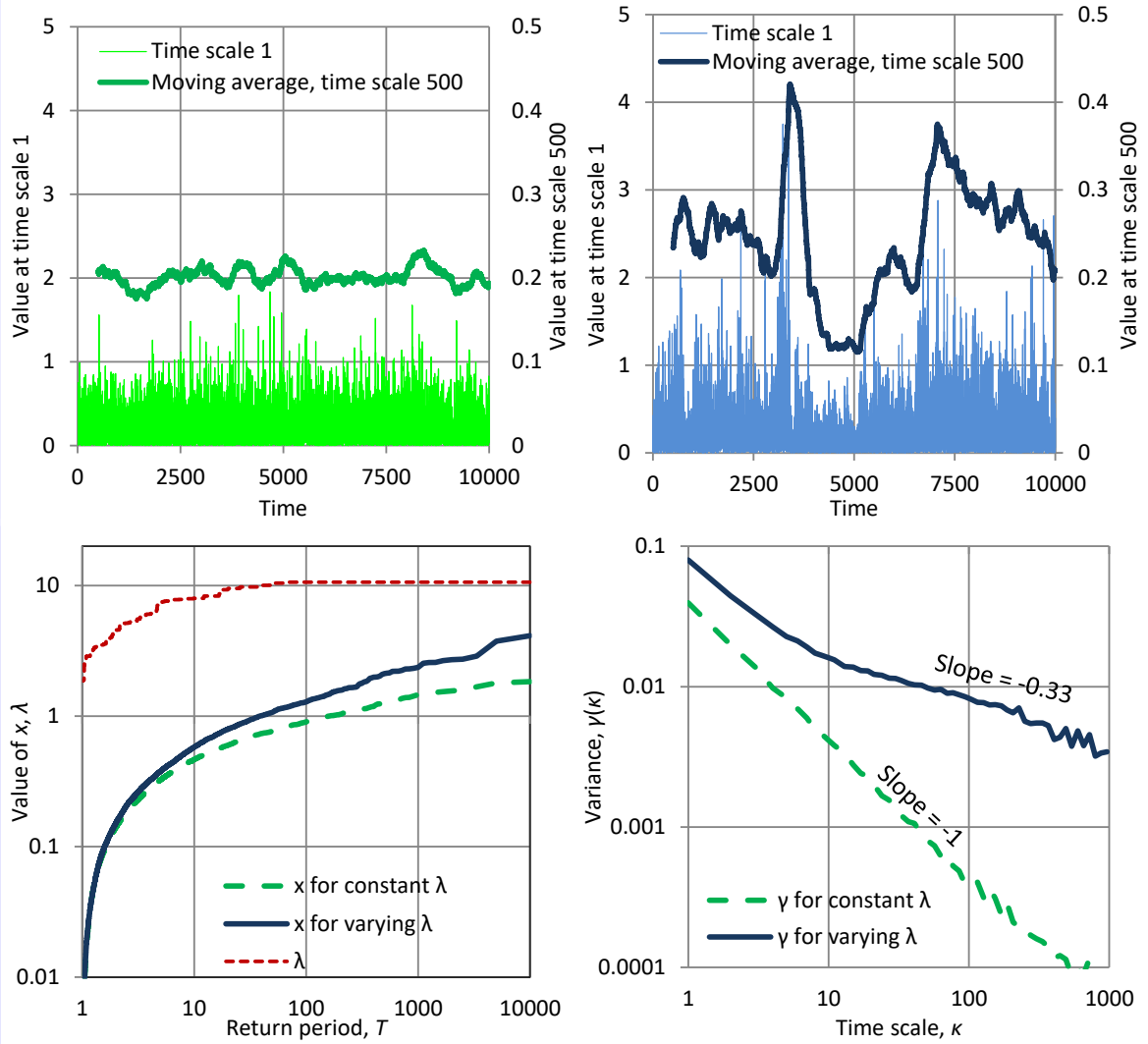

Figure 3.9 Graphs for the hypothetical example studied in Digression 3.I: (**upper left**) for constant λ; (**upper right**) for varying λ; (**lower left**) plots of distribution functions; (**lower right**) plots of climacograms (see text for further explanation).

Chapter 4. Fundamental concepts of statistics and their adaptation to stochastic processes

4.1 Introductory comments

The first aim of this chapter is to serve as a synopsis (rather than a systematic and complete presentation) of fundamental statistical concepts. It is well known that the aim of statistics per se is to provide a methodology for drawing conclusions based on observations. The conclusions are only inferences based on induction, not deductive mathematical proofs (see Digression 4.A); however, if the associated probabilities approach 1, they almost become certainties.

Classical statistical theory is entirely based on the assumption that observations are from a *sample*, a concept (formally defined in section 4.2) whose very definition relies on the independence of observations. However, when we deal with geophysical processes, independence is infrequent at a macroscopic level. Instead of samples we have time series and there is dependence in time. Even when we are interested in the spatial behaviour of processes, again we have to deal with dependence in space. Hence, the second aim of this chapter is to adapt and extend the classical statistical concepts and methodologies to make them applicable to a universe in which there is dependence.

Two important tasks in statistics are *estimation* and *hypothesis testing*. Statistical estimation can be distinguished in *parameter estimation* and *prediction* and can be performed either on a *point basis* (resulting in a single value, typically the expectation; cf. the Aristotelian *mesotes*), or on an *interval basis* (resulting in an interval where the quantity sought lies, associated with a certain probability or confidence). The results of an estimation procedure are called *estimates*. Uses of statistical estimation in geophysical applications include the estimation of parameters of marginal probability distributions or of the stochastic model describing the dependence in time, and of distributions quantiles. All these concepts are briefly discussed both at a theoretical level, to clarify the concepts and avoid misuses, and at a more practical level to illustrate the application of the concepts.

Statistical hypothesis testing is also an important tool that constitutes the basis of decision theory. In geophysical studies, it is useful not only in decision making, but also in exploratory tasks, such as in detecting relationships between different processes. Hypothesis testing is typically performed by the classical framework known as *statistical significance* (related to a *null hypothesis*) or, alternatively, within a *Bayesian framework*. These topics have been mostly developed on the basis of (independent) samples and, therefore, are not covered in this text. On the other hand, we emphasize the concept of *order statistics* (section 4.12), and the related to them *knowable moments* (*K-moments*; section 4.13). Both of these are particularly useful in estimating distribution quantiles, which in turn are useful in estimating model parameters and exploring the appropriateness of models.

Digression 4.A: Deduction and induction

The theory of probability has provided solid scientific grounds for philosophical concepts such as indeterminism and causality. In many scientific and technological applications, probability has provided the tools to quantify uncertainty, rationalize decisions under uncertainty, and make predictions of future events under uncertainty, in lieu of unsuccessful deterministic predictions (see Koutsoyiannis, 2010).

Probability has also provided the basis for extending the typical mathematical logic, offering the mathematical foundation of induction. Thus, probability made it possible to incorporate into mathematics the entire Aristotelian logic, which in addition to *deductive reasoning* or *deduction* (the Aristotelian *apodeixis*) also includes *induction* (the Aristotelian *epagoge*).

In classical mathematical logic, determinism can be paralleled to the premise that all truth can be revealed by deductive reasoning. This type of reasoning consists of repeated application of strong syllogisms concerning the logical propositions *A* and *B*, such as:

(Premise)	If *A* is true, then *B* is true;	If *A* is true, then *B* is true;
(Evidence)	*A* is true;	*B* is false;
(Conclusion)	*B* is true.	*A* is false.

Deduction uses a set of axioms to prove propositions known as theorems, which, given the premises (based on axioms), are irrefutable, absolutely true statements. It is also irrefutable that deduction is the preferred route to truth. The question is, however, does deduction have any limits?

David Hilbert's famous aphorism "*Wir müssen wissen, wir werden wissen*" ("We must know, we will know"; see section 1.1), expressed his belief that there were no limits to deduction. According to this belief, more formally known as *completeness*, any mathematical statement could be proved or disproved by deduction from axioms. However, developments in mathematical logic, and particularly Gödel's *incompleteness theorem*, challenged the omnipotence of deduction suggesting the usefulness and necessity of induction.

Induction uses weaker inference rules of the type:

(Premise)	If *A* is true, then *B* is true;	If *A* is true, then *B* is true;
(Evidence)	*B* is true;	*A* is false;
(Conclusion)	*A* becomes more plausible.	*B* becomes less plausible.

Induction offers no proof as to whether a proposition is true or false and may lead to errors. However, it is very useful in decision making, when deduction is not possible, which is the case quite frequently in the real world and everyday life (see Jaynes, 2003).

The important achievement of probability is that it quantifies (expresses in the form of a number between 0 and 1) the degree of plausibility of a certain proposition or statement. The formal probability framework uses both deduction, for proving theorems, and induction, for inference with incomplete information or data. For the latter we use the branch of stochastics called statistics.

4.2 Samples vs. time series

Loosely speaking, statistics draws conclusions for a *population* based on a *sample*. Although the content of *population* is not strictly defined in the statistical literature, the term describes any collection of objects whose measurable attributes are of interest. The population can refer to the real world and be finite (e.g., the inhabitants of Europe or the mean annual flows of year 2000 at the outlets of all river basins on Earth with size greater than 100 km^2). Population can also be an abstraction from a real-world entity referring to the possible (typically infinite) outcomes of a real or a hypothetical experiment (e.g., the population of all possible annual flows at a river cross-section). Here we deal with populations of the latter type and, because of this, it is not necessary to use the term

population at all—and hence to define it. Rather, the notions of a stochastic variable and a stochastic process suffice. Therefore, we will not use terms like *population mean* to distinguish from *sample mean*. Instead, we will refer to the former concept by the terms *true mean*, *ensemble mean* or simply *mean*, where the term *ensemble* suggests all possible outcomes of repeated experiments.

Unlike the term *population*, the term *sample* has a clear definition. Specifically, a sample of size (or length) n of a stochastic variable $\underline{x}$, defined on a ground set Ω, with probability distribution function $F(x)$, is a sequence of n *independent identically distributed* (IID) stochastic variables $(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n)$ defined on the sample space $\Omega_n = \Omega \times \cdots \times \Omega$, each having distribution $F(x)$ (adapted from Papoulis, 1990, p. 238). After observation of the variables $\underline{x}_i$, to each variable there corresponds *one* numerical value. Consequently, we will have a numerical sequence $x_1, x_2, \dots, x_n$, called the *observed sample*. It is clear from this definition that a sample is not a subset of the population, as some may think, but a concept related to the Cartesian product of the population.

Thus, the concept of a sample is related to sequences of two types: an abstract sequence of stochastic variables and the corresponding sequence of their numerical values. It has been common practice to use the term *sample* indistinguishably for both sequences, omitting the term *observed* from the latter. However, the two concepts are fundamentally different and each time we should be attentive to distinguish to which of the two cases the term sample refers.

The above definition (and in particular the IID specification) suggests that the construction of a sample of size n, or the *sampling*, is done by performing n repetitions of an experiment. The repetitions should be independent of each other and be performed under virtually the same conditions. However, in dealing with natural phenomena (outside the laboratory) it is not possible to repeat the same experiment, and thus literally no sampling can be done. Instead, what is actually done is measurement of the natural process at different times. As a consequence, it is not possible to ensure that independence and same conditions hold. Actually, in most cases we can be sure of the opposite. Then using classical statistics can become dangerous as the estimates and inferences may be completely wrong.

Still, however, we can do our job in a reliable manner if, instead of using classical statistics, we rely on stochastics and use the following correspondence between classical statistical concepts the stochastic concepts:

Classical statistics (independence) →	Statistics within stochastics (dependence)
Sample →	Stochastic process (discrete or discretized)
Observed sample →	Time series

Typically, the use of stochastics assuming dependence makes the mathematical derivations and calculations more complicated, while the resulting uncertainty is greater when there is dependence.

4.3 Expectation and its estimation

As we have stressed in Chapter 2, functions of stochastic variables, e.g. $\underline{z} \coloneqq g(\underline{x})$ are stochastic variables and expected values of stochastic variables are common variables. For example $\mathrm{E}[\underline{x}]$ and $\mathrm{E}[g(\underline{x})]$ are constants and not functions of x or $\underline{x}$, i.e.:

$$\mathrm{E}[\underline{x}] \coloneqq \int_{-\infty}^{\infty} x f(x)\mathrm{d}x =: \mu, \qquad \mathrm{E}[g(\underline{x})] \coloneqq \int_{-\infty}^{\infty} g(x) f(x)\mathrm{d}x, \tag{4.1}$$

where $f(x)$ is the probability density function. It should be stressed that these expectations are not time averages. Sometimes to make it clearer we call them true or ensemble means, variances, covariances, etc. For an ergodic process, true expectations are related to time averages through the following asymptotic relationship (section 3.4):

$$\underline{\hat{G}}^{(\infty)} \coloneqq \lim_{T\to\infty} \frac{1}{T}\int_0^T g\left(\underline{x}(t)\right)\mathrm{d}t = \mathrm{E}\left[g\left(\underline{x}(t)\right)\right] =: G \tag{4.2}$$

We notice that the left-hand side, $\underline{\hat{G}}^{(\infty)}$, is a stochastic variable while the right-hand side, G, is a common variable. Their equality implies that the variance of $\underline{\hat{G}}^{(\infty)}$ is zero.

When dealing with data from a process $\underline{x}(t)$ with a joint distribution function that is unknown, neither the left- nor the right-hand side of (4.2) can be known a priori. Assuming that we have a time series, at a time step D, with observations $x_\tau \coloneqq (1/D)\int_{(\tau-1)D}^{\tau D} x(u)\mathrm{d}u\,, \tau = 1, \dots, n$ (see equation (3.1)) we can approximate the left-hand side by:

$$\hat{G} \coloneqq \frac{1}{n}\sum_{\tau=1}^{n} g(x_\tau) \tag{4.3}$$

The common variable $\hat{G}$ is called an *estimate* of the true expectation G. Replacing in equation (4.3) the values x_τ with the stochastic variables $\underline{x}_\tau$ we define:

$$\underline{\hat{G}} \coloneqq \frac{1}{n}\sum_{\tau=1}^{n} g(\underline{x}_\tau) \tag{4.4}$$

The stochastic variable $\underline{\hat{G}}$ is called an *estimator* of the true expectation G. In classical statistics $\underline{\hat{G}}$ is also called a *statistic*, where the latter term denotes a (scalar) function of the *sample vector* $\underline{\boldsymbol{x}} \coloneqq [\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n]^{\mathrm{T}}$.

While the above procedure to form an estimator $\underline{\hat{G}}$ of the true expectation G is useful in many cases, we should bear in mind that many different estimators can be formulated for a certain parameter G. An estimator is typically biased (with some exceptions, the most notable being the estimator of the mean; see below), meaning that:

$$\mathrm{E}[\underline{\hat{G}}\,] \neq G \tag{4.5}$$

A formal definition of bias is:

$$b \coloneqq \mathrm{E}\left[\underline{\hat{G}}\right] - G \tag{4.6}$$

An estimator is also characterized by its variance and its mean square error, i.e.

$$\gamma_G \coloneqq \mathrm{var}\left[\underline{\hat{G}}\right], \qquad e_G \coloneqq \mathrm{E}\left[\left(\underline{\hat{G}} - G\right)^2\right] = \gamma_G + b^2 \tag{4.7}$$

An estimator is called:

- *unbiased* if $b = 0$.
- *consistent* if, with probability 1, $\underline{\hat{G}} - G \to 0$ as $n \to \infty$;
- *best* if e_G is minimum.
- *most efficient* if it is unbiased and best.

The main takeaway and central point of the above discussion is this. When dealing with quantification of uncertainty, for each parameter there are four different concepts, with slightly different names but very different meanings and content. These are often confounded in the literature with the same symbol and name being used for all of them, which causes confusion and may result in wrong conclusions. Table 4.1 clarifies the four different concepts using the variance as an example.

Table 4.1 Different variants of the variance of a stationary process in discrete time, $\underline{x}_\tau$, as an example for clarifying the four different concepts.

Name	Symbol and definition	Type of variable	Type of determination
Variance (true)	$\gamma \coloneqq \int_{-\infty}^{\infty} (x-\mu)^2 f_{\underline{x}_\tau}(x)\mathrm{d}x$	Common variable (not depending on τ)	Theoretical calculation from model (by integration)
Variance estimate	$\hat{\gamma} \coloneqq \frac{1}{n}\sum_{\tau=1}^{n}(x_\tau - \hat{\mu})^2$ where: $\hat{\mu} \coloneqq \frac{1}{n}\sum_{\tau=1}^{n} x_\tau$	Common variable	Estimation from data—but model is also necessary (e.g. to calculate the estimation bias and uncertainty)
Variance estimator	$\underline{\hat{\gamma}} \coloneqq \frac{1}{n}\sum_{\tau=1}^{n}\left(\underline{x}_\tau - \underline{\hat{\mu}}\right)^2$ where: $\underline{\hat{\mu}} \coloneqq \frac{1}{n}\sum_{\tau=1}^{n}\underline{x}_\tau$	Stochastic variable	Theoretical calculation from model
Variance estimator limit	$\underline{\hat{\gamma}}^{(\infty)} = \lim_{T\to\infty}\frac{1}{T}\int_0^T \left(\underline{x}(t) - \underline{\hat{\mu}}^{(\infty)}\right)^2 \mathrm{d}t$ where $\underline{\hat{\mu}}^{(\infty)} \coloneqq \lim_{T\to\infty}\frac{1}{T}\int_0^T \underline{x}(t)\,\mathrm{d}t$	Stochastic variable, which for an ergodic process has zero variance and becomes a common variable, equal to γ	Theoretical calculation from model

From Table 4.1 we notice that the data can be used only with one of the variance variants, namely the variance estimate, while a theoretical model is necessary to

determine any of them. Even for the variance estimate, a model is necessary to estimate the estimation bias and uncertainty (in classical statistics, that model is the IID assumption). And before specifying that model, it is fundamental to ensure that the assumptions of stationarity and ergodicity are valid for the process and the data we are dealing with. If they are valid, then the four concepts become three because the variance estimator limit becomes identical to the true variance. But if stationarity and ergodicity do not hold, then one may use the data again, do the calculations and find a result. However, this result is meaningless and cannot be called the variance estimate.

4.4 Moment estimators

The estimator of the noncentral moment (moment about the origin) of order q, μ'_q, of a stochastic variable $\underline{x}$, formed according to the method described in section 4.3, is:

$$\underline{\hat{\mu}}'_q \coloneqq \frac{1}{n}\sum_{i=1}^{n} \underline{x}_i^q \tag{4.8}$$

It can be proved (Kendall and Stewart, 1963, p. 229) that:

$$\mathrm{E}\left[\underline{\hat{\mu}}'_q\right] = \mu'_q \tag{4.9}$$

Consequently, the noncentral moment estimators are unbiased. If $\underline{x}_i$ is a (IID) sample of size n then the variance of the estimator is:

$$\mathrm{var}\left[\underline{\hat{\mu}}'_q\right] = \frac{1}{n}\left(\mu'_{2q} - \mu'^2_q\right) \tag{4.10}$$

It can be observed that if the moments are finite, then the variance tends to zero as $n \to \infty$; therefore, the estimator is consistent. However, if $\underline{x}_i$ is a stochastic process (with time dependence) then (4.10) does not hold, even for q as low as 1.

The estimator of the central moment μ_q, is:

$$\underline{\hat{\mu}}_q \coloneqq \frac{1}{n}\sum_{\tau=1}^{n} \left(\underline{x}_\tau - \underline{\hat{\mu}}\right)^q \tag{4.11}$$

where $\underline{\hat{\mu}} \equiv \underline{\hat{\mu}}'_1$ is the estimator of the mean. This is a biased estimator for any $q > 1$. Even for relatively low q (e.g. 2-4), the bias can be substantial in the case that the process exhibits long-range dependence (see section 4.6 about the variance). In the case of (IID) samples and low q, the bias is much smaller and can be easily quantified (see e.g. Koutsoyiannis, 1997). For higher q the estimation of moments becomes almost impossible. This applies not only to the biased estimators of central moments, but also to the unbiased estimators of noncentral moments. The reasons are the high variance and the extraordinarily high skewness of the estimators, which means that their expectation can be different from the mode (the most probable value) by orders of magnitude. Because of that, classical moments have been called *unknowable* (see Digression 4.B) and their estimation from data is not recommended.

In the framework developed and followed in this text, we avoid estimation of classical moments of order higher than 2. For this reason, in the following sections we will only study the estimators of classical moments of orders 1 and 2. A different type of moments, the so-called *knowable moments*, whose estimates are relatively reliable up to an order p comparable to the sample size n will be discussed in section 4.13.

Digression 4.B: Are classical moments knowable?

The estimators of the noncentral moments $\underline{\hat{\mu}}'_q$ (or even the central ones if μ is known a priori, which however is almost never the case) are in theory unbiased, but it is impractical to use them in estimation if $q > 2$ (cf. Lombardo et al., 2014).

It is well known that for large q and positive x_i the following relationship holds as an approximation:

$$\left(\sum_{i=1}^{n} x_i^q\right)^{1/q} \approx \max_{1\le i\le n}(x_i)$$

This is related to the well-known mathematical fact that the maximum norm is the limit of the q-norm as $q \to \infty$. This result can be generalized for x_i that are not necessarily positive but satisfy the condition $\max_{1\le i\le n}(x_i) > |\min_{1\le i\le n}(x_i)|$. A numerical illustration of how fast the convergence of the left-hand side to the right-hand side of the above equation is provided in Table 4.2.

Table 4.2 Illustration of the fact that raising to a power and adding converges fast to the maximum value.

Linear, $q = 1$	Pythagorean, $q = 2$	Cubic, $q = 3$	High order, $q = 8$
$3 + 4 = 7$	$3^2 + 4^2 = 5^2$	$3^3 + 4^3 = 4.5^3$	$3^8 + \mathbf{4}^8 \approx \mathbf{4}^8$
$3 + 4 + 12 = 19$	$3^2 + 4^2 + 12^2 = 13^2$	$3^3 + 4^3 + 12^3 = 12.2^3$	$3^8 + 4^8 + \mathbf{12}^8 \approx \mathbf{12}^8$

Therefore, for relatively large q the estimate of μ'_q will be:

$$\hat{\mu}'_q = \frac{1}{n}\sum_{i=1}^{n} x_i^q \approx \frac{1}{n}\left(\max_{1\le i\le n}(x_i)\right)^q$$

(Note that for large q the term $(1/n)$ on the right-hand side can be omitted with a negligible error). Thus, for an unbounded variable $\underline{x}$ and for large q, we can conclude that $\underline{\hat{\mu}}'_q$, while theoretically an unbiased estimator of μ'_q, is in practice more an estimator of an extreme quantity than an estimator of μ'_q. (As seen in Koutsoyiannis, 2025, section 6.11, the estimated quantity is related to K-moments). This happens because the convergence of $\underline{\hat{\mu}}'_q$ to μ'_q is very slow, while the convergence to the maximum value is fast.

This is further illustrated in Figure 4.1 for the eighth moment of a process specified in the figure caption. Even for n as large as 64 000 the sample moment estimate continues to be smaller than the theoretical value by several orders of magnitude. However, the proximity of the moment estimate to the maximum value is evident even for n as small as 10. The jagged shapes of the curves are a clear indication of the dominance of maxima in the moment estimation: The steps occur when a new higher maximum value enters the sample, while the gradual decreases before these steps are due to the increase of the sample size without a higher maximum value. The ensemble simulation results in the right panel show that the 99% prediction limits (see their definition in section 4.11) from 1000 simulations are unable even to envelop the true value.

As a result, unless q is very small, μ'_q is not a knowable quantity: we cannot infer its value from a sample. This is the case even if n is very large as in Figure 4.1. Also, the various $\underline{\hat{\mu}}'_q$ are not independent of each other as they only differ in the power to which the maximum value is raised.

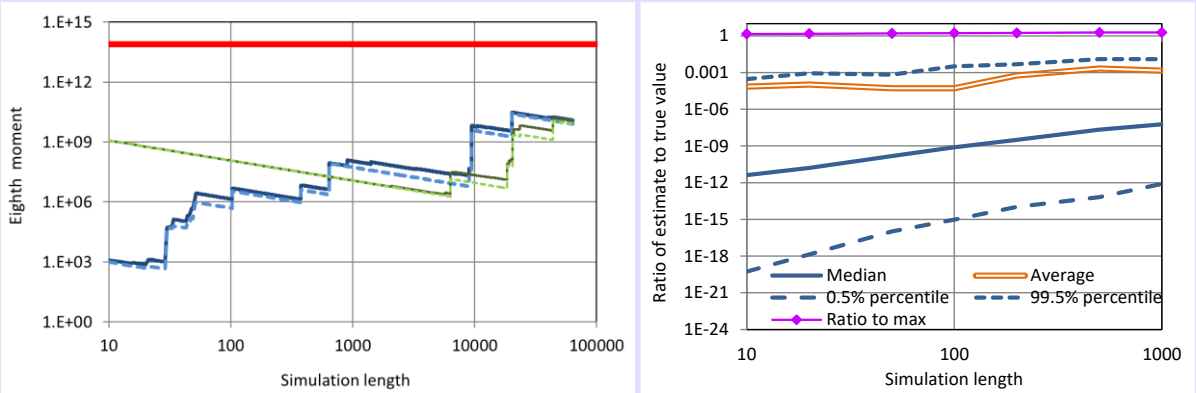

Figure 4.1 Illustration of the slow convergence of the sample estimate of the eighth noncentral moment to its true value, which is depicted as a thick horizontal line and corresponds to a lognormal distribution LN(0,1) where the process is an exponentiated Hurst-Kolmogorov process with Hurst parameter $H = 0.9$. **(left)** The sample moments are estimated from a single simulation of that process with length 64 000, where parts of this time series with sample size n from 10 to 64 000 are used for the estimation. Subsetting of the time series to sample size n was done either from the beginning to the end (thicker lines) or from the end to the beginning (finer lines). Continuous lines in the two cases represent the eighth moment estimates, $\sum_{i=1}^{n} x_i^8 \,/n$, and dashed lines represent maximum values, $\left(\max_{1\le i\le n}(x_i)\right)^8/n$. **(right)** Sampling distribution of the eighth moment estimator $\sum_{i=1}^{n} \underline{x}_i^8 \,/n$ estimated from 1000 simulated series of length 1000 each and visualized by the 99% prediction limits (percentiles), the median and the average, plotted as ratios to the true value. Theoretically, the ratio should be 1, but it is smaller by many orders of magnitude, and the convergence to 1 is very slow. The ratio to $\left(\max_{1\le i\le n}(x_i)\right)^8/n$, also plotted, is close to 1. (Source: Koutsoyiannis, 2019a.)

4.5 Sample mean estimator and effective sample size

According to equation (4.12), the estimator of the true mean $\mu = \mu_1'$ is:

$$\underline{\hat{\mu}} := \frac{1}{n}\sum_{i=1}^{n} \underline{x}_i \tag{4.12}$$

Another common notation of the mean estimator is $\overline{\underline{x}}$. The estimator is unbiased ($\mathrm{E}\left[\underline{\hat{\mu}}\right] = \mathrm{E}\left[\underline{x}\right] = \mu$). Its numerical value $\hat{\mu} := (1/n)\sum_{i=1}^{n} x_i$, else denoted as $\overline{x}$, is called the *observed mean* or the *average*. If $\underline{x}_i$ is a (IID) sample of size n then the variance of the estimator is:

$$\mathrm{var}\left[\underline{\hat{\mu}}\right] = \frac{\mathrm{var}\left[\underline{x}\right]}{n} = \frac{\gamma_1}{n} \tag{4.13}$$

regardless of the distribution function of $\underline{x}$. However, if $\underline{x}_i$ is a stochastic process (with dependence) then combining (3.12) and (4.12) we conclude that:

$$\underline{\hat{\mu}} = \underline{x}_1^{(n)} = \frac{\underline{X}(nD)}{nD} \tag{4.14}$$

where the superscript in parenthesis indicates that the discretization scale is nD (see equation (3.14)). Consequently:

$$\mathrm{var}\left[\underline{\hat{\mu}}\right] = \mathrm{var}\left[\frac{\underline{X}(nD)}{nD}\right] = \gamma(nD) = \gamma_n \tag{4.15}$$

Both equations (4.13) and (4.15) suggest that the estimator is consistent (assuming ergodicity) but they may result in quite different values of the variance. By means of these two equations we can define the notion of the ''equivalent'' (or ''effective'') sample size n' in the classical statistics (IID) sense (Koutsoyiannis and Montanari, 2007). This is the sample size of a hypothetical IID sample of a variable $\underline{x}$ with variance γ_1 whose variance of the mean equals γ_n; symbolically:

$$\frac{\gamma_1}{n'} = \gamma_n \Leftrightarrow n' = \frac{\gamma_1}{\gamma_n} \tag{4.16}$$

As an example, in an HK process, (equation (3.83); $\gamma_n = \lambda^2(\alpha/nD)^{2-2H}$), we will have:

$$n' = n^{2-2H} \tag{4.17}$$

In white noise ($H = 0.5$), clearly $n' = n$. However, if $H = 0.9$ and $n = 1000$ then $n' = 4$ (a big difference from 1000, highlighting uncertainty amplification). Thus, a time series of 1000 terms of that HK process is equivalent to a (classical, IID) sample of only 4 terms. This example shows the dramatic increase of uncertainty where there is dependence.

4.6 Climacogram estimator and its bias

The *typical variance estimator*:

$$\underline{\hat{\mu}}_2 \equiv \underline{\hat{\gamma}}_1 \coloneqq \frac{1}{n}\sum_{\tau=1}^{n}\left(\underline{x}_\tau - \underline{\hat{\mu}}\right)^2 \tag{4.18}$$

is well known to be biased. It is also well known from elementary classical statistics books that the replacement of n with $n-1$ in the denominator on the right-hand side makes the estimator unbiased. Thus, the *classical variance estimator* is:

$$\underline{\hat{\gamma}}_1^* \coloneqq \frac{1}{n-1}\sum_{\tau=1}^{n}\left(\underline{x}_\tau - \underline{\hat{\mu}}\right)^2 = \frac{n}{n-1}\underline{\hat{\gamma}}_1 \tag{4.19}$$

This is also known as *sample variance* or *unbiased variance estimator*. However, the latter term is incorrect: In stochastic processes describing natural phenomena, this slight change does not make the estimator unbiased. Here we use the term *typical* when we divide the sum by n (equation (4.18)) and *classical* when we divide by $n-1$ (equation (4.19)). We will use the same terminology for covariances below and we will explain the reasons why we prefer the typical over the classical.

In stochastic processes the bias can be determined analytically in terms of the climacogram as follows (see also Koutsoyiannis 2003, 2011a, 2016):

$$\begin{aligned}\mathrm{E}\left[\underline{\hat{\gamma}}_1\right] &= \frac{1}{n}\,\mathrm{E}\left[\sum_{\tau=1}^{n}\left(\left(\underline{x}_\tau - \mu\right) - \left(\underline{x}_1^{(n)} - \mu\right)\right)^2\right] \\ &= \frac{1}{n}\mathrm{E}\left[\sum_{\tau=1}^{n}\left(\underline{x}_\tau - \mu\right)^2\right] - 2\frac{1}{n}\,\mathrm{E}\left[\left(\underline{x}_1^{(n)} - \mu\right)\sum_{\tau=1}^{n}\left(\underline{x}_\tau - \mu\right)\right] + \mathrm{E}\left[\left(\underline{x}_1^{(n)} - \mu\right)^2\right]\end{aligned} \tag{4.20}$$

Since $\sum_{\tau=1}^{n}\left(\underline{x}_\tau - \mu\right) = n\left(\underline{x}_1^{(n)} - \mu\right)$ we find after the algebraic manipulations:

$$\mathrm{E}\left[\underline{\hat{\gamma}}_1\right] = \gamma_1 - \gamma_n = \left(1 - \frac{\gamma_n}{\gamma_1}\right)\gamma_1 = \left(1 - \frac{1}{n'}\right)\gamma_1 \tag{4.21}$$

Note that in the IID case, $\gamma_n = \gamma_1/n$, so the bias is $(1 - 1/n)\gamma_1$. On the other hand,

$$\mathrm{E}\left[\underline{\hat{\gamma}}_1^*\right] = \frac{n}{n-1}(\gamma_1 - \gamma_n) = \frac{1 - \gamma_n/\gamma_1}{1 - 1/n}\gamma_1 = \frac{1 - 1/n'}{1 - 1/n}\gamma_1 \tag{4.22}$$

Likewise, for the climacogram at scale $k = \kappa D$, if the observation period is $L = nk$, the estimators become:

$$\underline{\hat{\gamma}}_\kappa \equiv \underline{\hat{\gamma}}(k) \coloneqq \frac{1}{n}\sum_{\tau=1}^{n}\left(\underline{x}_\tau^{(\kappa)} - \underline{\hat{\mu}}\right)^2, \qquad \underline{\hat{\gamma}}_\kappa^* \equiv \underline{\hat{\gamma}}^*(k) \coloneqq \frac{n}{n-1}\underline{\hat{\gamma}}(k) \tag{4.23}$$

and their expectations are:

$$\mathrm{E}\left[\underline{\hat{\gamma}}(k)\right] = \gamma(k) - \gamma(L) = \left(1 - \frac{\gamma(L)}{\gamma(k)}\right)\gamma(k), \qquad \mathrm{E}\left[\underline{\hat{\gamma}}^*(k)\right] = \frac{1 - \gamma(L)/\gamma(k)}{1 - k/L}\gamma(k) \tag{4.24}$$

The above equations show that there is no gain in using the *classical* estimator (dividing by $n-1$) of variance $\underline{\hat{\gamma}}_1^*$ (or $\underline{\hat{\gamma}}^*(k)$). The equations are simpler if we use the *typical* estimator $\underline{\hat{\gamma}}_1$ (or $\underline{\hat{\gamma}}(k)$) (dividing by n). As we will see below, the *typical* estimator is also preferable when fitting distributional parameters. Whatever estimator we use, there is estimation bias which should be taken into account in model fitting.

4.7 Covariance and autocovariance estimators

The *typical* and *classical* estimators of covariance, i.e.:

$$\underline{\hat{c}}_{xy} \coloneqq \frac{1}{n}\sum_{\tau=1}^{n}\left(\underline{x}_\tau - \underline{\hat{\mu}}\right)\left(\underline{y}_\tau - \underline{\hat{\mu}}\right), \qquad \underline{\hat{c}}_{xy}^* \coloneqq \frac{1}{n-1}\sum_{\tau=1}^{n}\left(\underline{x}_\tau - \underline{\hat{\mu}}\right)\left(\underline{y}_\tau - \underline{\hat{\mu}}\right) \tag{4.25}$$

respectively, are both biased if $\underline{x}_\tau$ and $\underline{y}_\tau$ are stochastic processes non identical to white noise. For example, if they are HK processes with common Hurst parameter H, then the expectation of $\underline{\hat{c}}_{xy}$ is (Koutsoyiannis, 2003):

$$\mathrm{E}\left[\underline{\hat{c}}_{xy}\right] = \left(1 - \frac{1}{n^{2-2H}}\right)c_{xy} = \left(1 - \frac{1}{n'}\right)c_{xy} \tag{4.26}$$

In the case of autocovariance estimation, it is common knowledge that there is downward bias (Wallis and O'Connell, 1972; Salas, 1993, p. 19.10). The *typical* estimator of the lag η autocovariance is:

$$\underline{\hat{c}}_\eta \coloneqq \frac{1}{n}\sum_{\tau=1}^{n-\eta}\left(\underline{x}_\tau - \underline{\hat{\mu}}\right)\left(\underline{x}_{\tau+\eta} - \underline{\hat{\mu}}\right) \tag{4.27}$$

and it has been common practice to prefer it over the classical estimator (with division by $n-1$ or $n-\eta$), particularly when we use autocovariance to estimate the power spectrum. The expectation of $\underline{\hat{c}}_\eta$ is (see also Koutsoyiannis, 2003):

$$\begin{aligned}\mathrm{E}\left[\underline{\hat{c}}_\eta\right] &= \frac{1}{n}\,\mathrm{E}\left[\sum_{\tau=1}^{n-\eta}\left(\left(\underline{x}_\tau-\mu\right)-\left(\underline{x}_1^{(n)}-\mu\right)\right)\left(\left(\underline{x}_{\tau+\eta}-\mu\right)-\left(\underline{x}_1^{(n)}-\mu\right)\right)\right] \\ &= \frac{1}{n}\mathrm{E}\left[\sum_{\tau=1}^{n-\eta}\left(\underline{x}_\tau-\mu\right)\left(\underline{x}_{\tau+\eta}-\mu\right)\right] \\ &\quad -\frac{1}{n}\,\mathrm{E}\left[\left(\underline{x}_1^{(n)}-\mu\right)\sum_{\tau=1}^{n-\eta}\left(\left(\underline{x}_\tau-\mu\right)+\left(\underline{x}_{\tau+\eta}-\mu\right)\right)\right]+\mathrm{E}\left[\left(\underline{x}_1^{(n)}-\mu\right)^2\right]\end{aligned} \tag{4.28}$$

Since $\sum_{\tau=1}^{n-\eta}\left(\underline{x}_\tau-\mu\right)=(n-\eta)\left(\underline{x}_1^{(n-\eta)}-\mu\right)$, assuming that η is small in comparison to n so that we can interchange $n-\eta$ and n, and also expanding the corresponding sums, after the algebraic manipulations we obtain:

$$\mathrm{E}\left[\underline{\hat{c}}_\eta\right]\approx c_\eta-\gamma_n=\left(1-\frac{\gamma_n}{c_\eta}\right)c_\eta=\left(1-\frac{\gamma_n}{r_\eta\gamma_1}\right)c_\eta=\left(1-\frac{1}{r_\eta n'}\right)c_\eta \tag{4.29}$$

For positive lag-η cross-correlation $\left(0<r_\eta<1\right)$, the relative bias $(-1/r_\eta n')$ is higher than that of the climacogram γ_η (i.e. $-1/n'$). An exact equation has been derived in Dimitriadis and Koutsoyiannis (2015; Table 2).

If we estimate the autocorrelation coefficient by:

$$\underline{\hat{r}}_\eta \coloneqq \frac{\underline{\hat{c}}_\eta}{\underline{\hat{\gamma}}_1} \tag{4.30}$$

then this will be biased again. An approximately unbiased estimator would be:

$$\underline{\tilde{r}}_\eta \coloneqq \frac{\underline{\hat{c}}_\eta+\gamma_n}{\underline{\hat{\gamma}}_1+\gamma_n}=\frac{\underline{\hat{r}}_\eta\underline{\hat{\gamma}}_1+\gamma_n}{\underline{\hat{\gamma}}_1+\gamma_n}\approx\left(1-\frac{1}{n'}\right)\underline{\hat{r}}_\eta+\frac{1}{n'} \tag{4.31}$$

It is stressed that the use of autocovariance and (even more so) of the autocorrelation estimates should be avoided when identifying and fitting a stochastic model. Identification and fitting are better served by the climacogram (see Digression 4.C).

Digression 4.C: The climacogram and the climacogram-based metrics compared to standard metrics

The most popular procedure in time series modelling, is to construct the empirical autocorrelogram of the time series using equation (4.27) and assess which stochastic process (e.g., of AR or ARMA type) is suitable and how many autocorrelation terms should be preserved. It is easy to illustrate how this technique can completely distort the underlying process. Figure 4.2(a) depicts the autocorrelogram of a time series with length 100, which does not seem to have any relationship to the theoretical autocorrelation function of the model from which it was constructed. Namely, the model is the FHK with parameters as in the caption of Figure 3.7. Clearly, the empirical autocorrelation does not give any hint that the time series stems from a process with persistence. With that autocorrelogram one would conclude that an AR(1) model with a lag-1 autocorrelation of about 0.4 would be appropriate.

The reasons for the failure of the autocorrelogram to capture the real behaviour of the process are two. First is the bias, as analysed in section 4.7. Second, from equation (3.30) it is seen that the autocorrelation is by nature the second derivative of the climacogram standardized by

variance. Estimation of the second derivative from data is too uncertain and makes a very rough graph.

The alternative to using the periodogram (the estimate of the power spectrum, which is the Fourier transform of the autocovariance; see section 4.10) is even worse as it entails an even rougher shape and more uncertain estimation than in the autocovariance (see also section 4.10 and Dimitriadis and Koutsoyiannis, 2015).

Thus, for model identification, it is much more preferable to use the climacogram directly instead of the autocorrelogram. For our example time series, this is illustrated in Figure 4.2(b), which indicates that the long-term persistence is well captured by the empirical climacogram, and the parameter H is correctly estimated ($H = 0.79$, based on the method presented in Koutsoyiannis, 2003, and Tyralis and Koutsoyiannis, 2011). Additional advantages to using the climacogram are (a) its intactness on discretization, (b) its close relationship with entropy production and (c) its expandability to high-order moments.

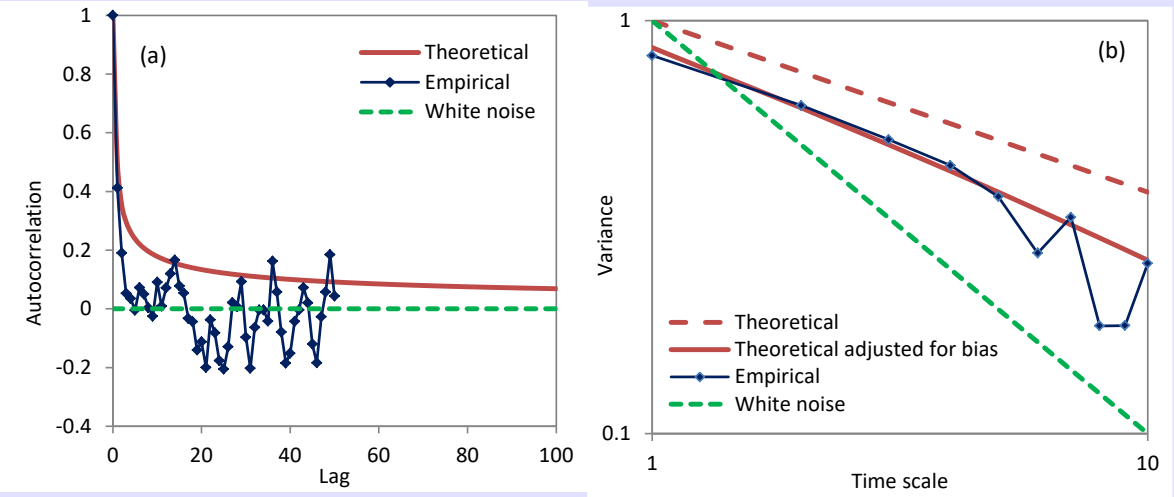

Figure 4.2 (left) Autocorrelogram and **(right)** climacogram of a time series of 100 terms generated from the FHK model with parameters as in the caption of Figure 3.7. (Source: Koutsoyiannis, 2016.)

4.8 Parameter estimation of distribution functions – The method of moments

Assuming a stochastic variable $\underline{x}$ with known distribution function but with unknown parameters $\boldsymbol{\theta} \coloneqq [\theta_1, \theta_2, \ldots, \theta_m]^{\mathrm{T}}$, we can denote the probability density function of $\underline{x}$ as a function $f(x, \boldsymbol{\theta})$. Here, we will examine the problem of the estimation of these parameters based on a sample vector $\underline{\boldsymbol{x}} \coloneqq [\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n]^{\mathrm{T}}$. In this section, we present one of the two most popular methods in statistics, namely the *method of moments.* The other popular method, the *maximum likelihood method*, we present in section 4.9. Several other general methods have been developed in statistics for parameter estimation, e.g. the maximum entropy method (e.g. Singh and Rajagopal, 1986) and the L-moments method (Hosking et al., 1985a,b; Hosking, 1990). Moreover, in practical applications, other types of methods like graphical, tabulated, empirical and semi-empirical, have been devised. Another approach, which is based on K-moments, is preferable, particularly when we do not have information about the true model for the marginal distribution function, and is detailed in Koutsoyiannis (2025).

The method of moments is based on equating the theoretical moments of $\underline{x}$ with the corresponding sample estimates of noncentral moments. Thus, as m is the number of the unknown parameters of the distribution, we can write m equations of the form

$$\mu'_q = \hat{\mu}'_q, \qquad q = 1, \dots, m \tag{4.32}$$

where the theoretical moments μ'_q are functions of the unknown parameters given by:

$$\mu'_q = \int_{-\infty}^{\infty} x^q \, f(x, \boldsymbol{\theta}) \mathrm{d}x \tag{4.33}$$

Thus, the solution to the resulting system of the m equations gives the unknown parameters $(\theta_1, \theta_2, \dots, \theta_m)$. In general, the system of equations may not be linear and may not have an analytical solution, necessitating a numerical solution.

This method is easy to apply. However, for distributions involving more than two parameters ($m > 2$), the problem of knowability of moments intervenes and makes the method unreliable. Certainly K-moments are preferable for this task. Furthermore, when dealing with extremes we must bear in mind that they are closely linked to high-order moments and, thus, it is not the best practice to rely on the lowest-order moments.

Digression 4.D: Illustration of the method of moments

As an example of the implementation of the method of moments, we will determine the parameters of the normal distribution. The probability density function:

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

has two parameters, μ and σ. Thus, we need two equations. Based on Table 2.3, these equations are:

$$\mu = \hat{\mu}, \qquad \sigma^2 + \mu^2 = \hat{\mu}_2 + \hat{\mu}^2 \Rightarrow \sigma^2 = \hat{\mu}_2$$

where we have used the identity $\mu'_2 = \mu_2 + \mu^2$. Consequently, the final estimates are

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \sigma^2 := \frac{1}{n}\sum_{\tau=1}^{n} (x_\tau - \hat{\mu})^2$$

This estimation μ is unbiased but that of σ^2 (and σ) is biased even in IID statistics (notice in the latter equation that the result contains the typical, rather the classical estimate).

As we have seen in this example, the application of the method of moments is very simple and this extends to many distribution functions.

4.9 Parameter estimation of distribution functions – The maximum likelihood method

While the method of moments is an ad hoc method and has several weaknesses described in section 4.8, the method of maximum likelihood has a strong logical background. We will initially present the method in a Bayesian framework and then we will see that it also stands outside that framework.

The problem that we have to solve is to find the parameter vector $\boldsymbol{\theta}$ from the known observations $\underline{\boldsymbol{x}} = \boldsymbol{x}$. Since the observations $\boldsymbol{x}$ are known while the parameters $\boldsymbol{\theta}$ are

unknown, we can regard the latter as stochastic variables $\underline{\boldsymbol{\theta}}$. This allows us to assign $\underline{\boldsymbol{\theta}}$ a probability density function $f_{\underline{\boldsymbol{\theta}}}(\boldsymbol{\theta})$ and also express conditional densities using the Bayes theorem (equation (2.3)). This can be written in terms of densities as:

$$f_{\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{x}}}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f_{\underline{\boldsymbol{x}}|\underline{\boldsymbol{\theta}}}(\boldsymbol{x}|\boldsymbol{\theta})}{f_{\underline{\boldsymbol{x}}}(\boldsymbol{x})}\, f_{\underline{\boldsymbol{\theta}}}(\boldsymbol{\theta}) \tag{4.34}$$

where we have replaced the events A and B with the vectors $\boldsymbol{x}$ and $\boldsymbol{\theta}$, respectively. The terminology used in the Bayesian framework is:

- Prior (before observation) probability density for $f_{\underline{\boldsymbol{\theta}}}(\boldsymbol{\theta})$
- Posterior (after observation) probability density for $f_{\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{x}}}(\boldsymbol{\theta}|\boldsymbol{x})$
- Likelihood for the conditional density $f_{\underline{\boldsymbol{x}}|\underline{\boldsymbol{\theta}}}(\boldsymbol{x}|\boldsymbol{\theta})$; this is the hypothesized model (i.e. distribution for $\underline{\boldsymbol{x}}$) given the parameters $\boldsymbol{\theta}$.

According to this terminology, we can write (4.34) in the following form:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior} \tag{4.35}$$

Since we have to assign $\underline{\boldsymbol{\theta}}$ a single value $\boldsymbol{\theta}$, the most rational choice for that value is the mode of its distribution conditional on $\underline{\boldsymbol{x}} = \boldsymbol{x}$, i.e., the value that maximizes the posterior $f_{\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{x}}}(\boldsymbol{\theta}|\boldsymbol{x})$. To find the mode we equate the derivative of the conditional density to 0, i.e.:

$$\frac{\mathrm{d}f_{\underline{\boldsymbol{\theta}}|\underline{\boldsymbol{x}}}(\boldsymbol{\theta}|\boldsymbol{x})}{\mathrm{d}\boldsymbol{\theta}} = \mathbf{0} \Leftrightarrow \frac{1}{f_{\underline{\boldsymbol{x}}}(\boldsymbol{x})}\left(\frac{\mathrm{d}f_{\underline{\boldsymbol{x}}|\underline{\boldsymbol{\theta}}}(\boldsymbol{x}|\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\, f_{\underline{\boldsymbol{\theta}}}(\boldsymbol{\theta}) + f_{\underline{\boldsymbol{x}}|\underline{\boldsymbol{\theta}}}(\boldsymbol{x}|\boldsymbol{\theta})\frac{\mathrm{d}f_{\underline{\boldsymbol{\theta}}}(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\right) = \mathbf{0} \tag{4.36}$$

Since we know nothing about the prior $f_{\underline{\boldsymbol{\theta}}}(\boldsymbol{\theta})$, we can choose a so-called noninformative prior, which does not change with $\boldsymbol{\theta}$, i.e. $\mathrm{d}f_{\underline{\boldsymbol{\theta}}}(\boldsymbol{\theta})/\mathrm{d}\boldsymbol{\theta} = 0$. In this case from (4.36) we obtain:

$$\frac{\mathrm{d}f_{\underline{\boldsymbol{x}}|\underline{\boldsymbol{\theta}}}(\boldsymbol{x}|\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} = \mathbf{0} \tag{4.37}$$

which demands also that the likelihood be at maximum. In other words, we find $\boldsymbol{\theta}$, demanding that the density $f_{\underline{\boldsymbol{x}}|\underline{\boldsymbol{\theta}}}(\boldsymbol{x}|\boldsymbol{\theta})$ have a value as high as possible at the point $\underline{\boldsymbol{x}} = \boldsymbol{x}$.

If the vector $\boldsymbol{x}$ is part of a stochastic process, determination of $f_{\underline{\boldsymbol{x}}|\underline{\boldsymbol{\theta}}}(\boldsymbol{x}|\boldsymbol{\theta})$ can be laborious. However, in IID statistics, $\boldsymbol{x}$ is a sample vector with independent items and thus the joint probability density function is:

$$f_{\underline{\boldsymbol{x}}|\underline{\boldsymbol{\theta}}}(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_{\underline{x}|\underline{\boldsymbol{\theta}}}(x_i|\boldsymbol{\theta}) \tag{4.38}$$

Thus, we seek a solution of:

$$\frac{\mathrm{d}\prod_{i=1}^{n} f_{\underline{x}|\underline{\boldsymbol{\theta}}}(x_i|\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} = \mathbf{0} \tag{4.39}$$

We can also convert the product to a sum by taking the logarithm of $f_{\underline{\boldsymbol{x}}|\underline{\boldsymbol{\theta}}}(\boldsymbol{x}|\boldsymbol{\theta})$:

$$L(\boldsymbol{x}|\boldsymbol{\theta}) \coloneqq \ln f_{\underline{x}|\underline{\boldsymbol{\theta}}}(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f_{\underline{x}|\underline{\boldsymbol{\theta}}}(x_i|\boldsymbol{\theta}) \tag{4.40}$$

The function $L(\ )$ is called the log-likelihood function. In this case, the condition of maximum is:

$$\frac{\mathrm{d}L(\boldsymbol{x}|\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{1}{f_{\underline{x}|\underline{\boldsymbol{\theta}}}(x_i|\boldsymbol{\theta})} \frac{\mathrm{d}f_{\underline{x}|\underline{\boldsymbol{\theta}}}(x_i|\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} = \mathbf{0} \tag{4.41}$$

Both (4.39) and (4.41) are vector equations equivalent to m scalar equations. Solving either of them will give us the values of the m unknown parameters.

Digression 4.E: Illustration of the maximum likelihood method

We will determine the parameters of the normal distribution from a sample using the maximum likelihood method. The probability density function of the normal distribution is:

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The likelihood function is:

$$f(\boldsymbol{x}|\mu,\sigma) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

The log-likelihood function is:

$$L(\boldsymbol{x}|\mu,\sigma) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Taking the derivatives with respect to the unknown parameters μ and σ and equating them to 0 we find

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0\,, \qquad \frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

and solving the system we obtain the final parameter estimates:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}, \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 = \hat{\mu}_2$$

The results are precisely identical with those of Digression 4.D, despite the fact that the two methods are fundamentally different. The application of the maximum likelihood method is more complex than that of the method of moments. The coincidence of results found here is not the rule for all distribution functions. On the contrary, in most cases the two methods yield different results.

4.10 Estimation of power spectrum and periodogram

We assume that a stochastic process $\underline{x}(t)$ is observed on a time-average basis at equidistant times $\tau D, \tau = 0, \ldots, n-1$, where D is a time step (a total observation time $L = nD$). We have thus a time series with a finite number, n, of observations x_τ of the discrete-

time process $\underline{x}_\tau$. If we study the process in the frequency domain, we have the following characteristic frequencies, dimensional (w) or dimensionless ($\omega = wD$):

Sampling frequency	$w_D = 1/D = n/L$	$\omega_D = w_D D = 1$
Nyquist frequency	$w_\mathrm{N} = 1/2D = n/2L$	$\omega_\mathrm{N} = w_\mathrm{N} D = 0.5$
Frequency resolution	$w_1 = 1/L = w_D/n$	$\omega_1 = w_1 D = D/L = 1/n$
Half frequency resolution	$w_2 = 1/2L = w_D/2n$	$\omega_2 = w_2 D = D/2L = 1/2n$

As we will see, the Nyquist frequency ($\omega_\mathrm{N} = 0.5$) is the maximum frequency on which we can make estimates as beyond that the resulting spectrum estimates are repeated in a cyclic manner.

We are interested in estimators of the power spectrum of the discrete-time process $\underline{x}_\tau$. A first estimator can be established by utilizing the relationship between the power spectrum and the autocovariance function (equation (3.36)). From n observations we can estimate from equation (4.27) up to n autocovariance terms, $\underline{\hat{c}}_0, \underline{\hat{c}}_1, \dots, \underline{\hat{c}}_{n-1}$ (noting that most of them will not be reliably estimated). Then, by truncating equation (3.36) to a finite number of terms we can formulate an estimator of the spectrum in the form:

$$\underline{\hat{s}}_\mathrm{d}(\omega) = 2\underline{\hat{c}}_0 + 4\sum_{\eta=1}^{n-1} \underline{\hat{c}}_\eta \cos(2\pi\eta\omega) + 2\underline{\hat{c}}_n \cos(2\pi n\omega) \tag{4.42}$$

where we have put a last term for $\underline{\hat{c}}_n$ with a weight 2 (instead of 4), which, as we will see facilitates and accelerates calculations. If we have n data values x_τ, then $\underline{\hat{c}}_n \equiv 0$, but the calculation should stand in cases where we use a fewer number of autocorrelations or in cases where we process true values rather than estimates (in the latter case, $c_n \neq 0$). While from first glance we can use this equation to estimate $\underline{\hat{s}}_\mathrm{d}(\omega)$ for any ω, the resulting values are not always consistent and therefore it is advisable to make estimates for a finite number of discrete frequencies $\omega_j = j\omega_0$, where ω_0 is either ω_1 or ω_2 with j taking integer values as we will specify below.

The inversion of the formula to find the autocovariance estimates from the power spectrum estimates is possible through the equation:

$$\underline{\hat{c}}_\eta = \omega_0 \left(\frac{\underline{\hat{s}}_\mathrm{d}(0) + (-1)^n \underline{\hat{s}}_\mathrm{d}(0.5)}{2} + \sum_{0<\omega_j<0.5} \underline{\hat{s}}_\mathrm{d}(\omega_j) \cos(2\pi\eta\omega_j) \right) \tag{4.43}$$

The estimation of $\underline{\hat{s}}_\mathrm{d}(\omega)$ is streamlined and accelerated if we use the *discrete Fourier transform* (DFT) and particularly its variant named *fast Fourier transform* (FFT), for which the required software exists on all computational environments. For a sequence of numbers $x_\tau, \tau = 0, \dots, N-1$, the DFT is defined as a sequence $u_j, j = 0, \dots, N-1$, where:

$$u_j = \frac{1}{N} \sum_{\tau=0}^{N-1} x_\tau \mathrm{e}^{-\mathrm{i}2\pi\tau j/N}, \qquad j = 0, \dots, N-1 \tag{4.44}$$

The sequence x_τ is recovered from the sequence u_j by the inverse DFT, which is:

$$x_\tau = \sum_{j=0}^{N-1} u_j \mathrm{e}^{\mathrm{i}2\pi j\tau/N}, \qquad \tau = 0, \dots, N-1 \tag{4.45}$$

The FFT is the DFT made by a fast computational algorithm; the fastest case is when n is a power of 2.

To utilize DFT and FFT in determining $\underline{\hat{s}}_{\mathrm{d}}(\omega)$ we write equation (4.42) as:

$$\underline{\hat{s}}_{\mathrm{d}}(\omega) = \sum_{\eta=0}^{n} 2\underline{\hat{c}}_\eta \cos(2\pi\eta\omega) + \sum_{\eta=1}^{n-1} 2\underline{\hat{c}}_\eta \cos(2\pi\eta\omega) \tag{4.46}$$

Setting $j = \eta$ for the first sum and $j = 2n - \eta$ for the second sum we have:

$$\underline{\hat{s}}_{\mathrm{d}}(\omega) = \sum_{j=0}^{n} 2\underline{\hat{c}}_j \cos(2\pi j\omega) + \sum_{j=n+1}^{2n-1} 2\underline{\hat{c}}_{2n-j} \cos(2\pi(2n-j)\omega) \tag{4.47}$$

If ω is an integer multiple of $\omega_2 = 1/N$ where $N \coloneqq 2n$, then $2n\omega$ will be an integer and thus $\cos(2\pi(2n-j)\omega) = \cos(2\pi j\omega)$. By setting:

$$\underline{u}_j = \begin{cases} 2\underline{\hat{c}}_j, & 0 \le j \le n \\ 2\underline{\hat{c}}_{2n-j}, & n \le j \le N-1 \end{cases} \tag{4.48}$$

we can simplify (4.47) to:

$$\underline{\hat{s}}_{\mathrm{d}}(\omega) = \sum_{j=0}^{N-1} \underline{u}_j \cos(2\pi j\omega) \tag{4.49}$$

Considering that the imaginary part of $\underline{u}_j$ is zero, setting $\omega_\tau = \tau/N$, and comparing equations (4.45) and (4.49), we conclude that $\underline{\hat{s}}_{\mathrm{d}}(\omega_\tau)$ is the inverse DFT of $\underline{u}_j$. If we have taken care to choose n as a power of 2, N will also be a power of 2 and thus we can use the inverse FFT to calculate estimates $\hat{s}_{\mathrm{d}}(\omega_\tau)$ from estimates $\hat{c}_\eta$ for frequencies ω ranging from 0 to 0.5 with a resolution $\omega_2 = 1/N = 1/2n$. The inverse of (4.49) is:

$$\underline{u}_j = 2\underline{\hat{c}}_j = \frac{1}{N} \sum_{\tau=0}^{N-1} \underline{\hat{s}}_{\mathrm{d}}(\omega_\tau) \cos(2\pi j\omega_\tau), \qquad 0 \le j \le n \tag{4.50}$$

There is an alternative way to produce another estimator of the power spectrum using the DFT in the discrete-time process per se, rather than in its autocovariance. Specifically, the DFT of $\underline{x}_\tau$ is:

$$\underline{u}_j = \frac{1}{n} \sum_{\tau=0}^{n-1} \underline{x}_\tau \mathrm{e}^{-\mathrm{i}2\pi\tau j/n}, \qquad j = 0, \dots, n-1 \tag{4.51}$$

Assuming that $\underline{x}_\tau, \tau = 0, \dots, n-1$, are real-valued stochastic variables, their transformation $u_j, j = 0, \dots, n-1$, will be complex valued stochastic variables, i.e. $\underline{u}_j = \underline{u}_j^{\mathrm{R}} + \mathrm{i}\,\underline{u}_j^{\mathrm{I}}$, where $\underline{u}_j^{\mathrm{R}}$ and $\underline{u}_j^{\mathrm{I}}$ are real-valued. The inverse DFT of $\underline{u}_j$ recovers the real-valued $\underline{x}_\tau$. The sequence of the absolute values of $\underline{u}_j$ multiplied by $2n$:

$$\underline{S}_j := 2n\left|\underline{u}_j\right|^2 = 2n\left(\left(\underline{u}_j^{\mathrm{R}}\right)^2 + \left(\underline{u}_j^{\mathrm{I}}\right)^2\right) \tag{4.52}$$

is real valued and, as a function of $\omega_j = \frac{j}{n}$, is known as the *periodogram* of $\underline{x}_\tau$. It is another estimator of $s_{\mathrm{d}}(\omega)$ with a resolution $\omega_1 = \frac{1}{n}$ (while in the estimator (4.49) this is $\omega_2 = 1/2n$). The two alternatives used for estimating the power spectrum are schematically presented in Figure 4.3.

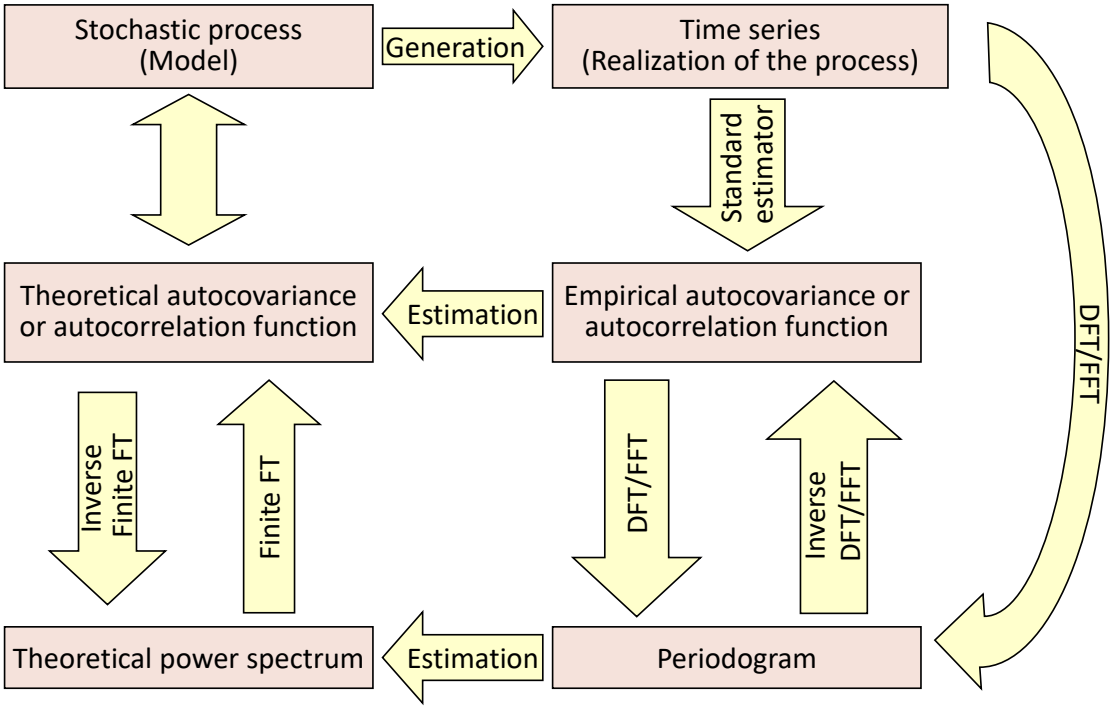

Figure 4.3 Schematic of the different paths to estimate the power spectrum.

For real-valued $\underline{x}_\tau$ the stochastic variables $\underline{u}_j$ and $\underline{S}_j$ have the following properties of symmetry:

$$\begin{gathered} \underline{u}_0 = \underline{u}_0^{\mathrm{R}} = \underline{\hat{\mu}}, \qquad \underline{u}_0^{\mathrm{I}} = 0 \\ \underline{u}_{n-j}^{\mathrm{R}} = \underline{u}_j^{\mathrm{R}}, \qquad \underline{u}_{n-j}^{\mathrm{I}} = -\underline{u}_j^{\mathrm{I}}, \qquad \underline{S}_{n-j} = \underline{S}_j, \qquad 1 \le j \le n-1 \end{gathered} \tag{4.53}$$

In other words, the real component of $\underline{u}_j$ and $\underline{S}_j$ are symmetric with respect to $n/2$, while the imaginary component is antisymmetric. Consequently, if n is even, then $\underline{u}_{n/2}^{\mathrm{I}} = 0$. Because of the symmetries, starting with n real numbers x_τ we end up with $n/2$ pairs of real numbers $\underline{u}_j^{\mathrm{R}}$ and $\underline{u}_j^{\mathrm{I}}$, and $n/2$ real numbers $\underline{S}_j$. The values of $\underline{S}_j$ for frequencies $\omega = j/n \le 0.5$ provide all extractable information while larger frequencies do not add anything of value.

Other interesting properties of the periodogram and the related quantities are:

$$\frac{1}{n}\sum_{\tau=0}^{n-1}\underline{x}_\tau^2 = \sum_{j=0}^{n-1}\left|\underline{u}_j\right|^2, \qquad \underline{\hat{\gamma}}_1 = \frac{1}{n}\sum_{\tau=0}^{n-1}\left(\underline{x}_\tau - \underline{\hat{\mu}}\right)^2 = \sum_{j=1}^{n-1}\left|\underline{u}_j\right|^2 = \frac{1}{n}\sum_{1\le j\le n/2}\underline{S}_j - \frac{\underline{S}_{n/2}}{2n} \tag{4.54}$$

where if n is odd, the last term $\underline{S}_{n/2}$ is set to zero. The latter equation allows us to decompose the variance estimate $\hat{\gamma}_1$ into partial components $\left|u_j\right|^2$, each corresponding to a particular frequency, which ranges from $\omega_1 = w_1 D = 1/n$ to $\omega_{\mathrm{N}} = w_{\mathrm{N}} D = 0.5$. The frequency 0 corresponds to the estimate of the mean and is not related to the variance. Any prominence (peak) in one or more $\left|\underline{u}_j\right|^2$ over the other is very often regarded as evidence of a periodic behaviour of the process with a frequency j/n (period n/j).

However, claims of periodicities without a deterministic explanation are usually meaningless. As evident from the notation in the entire section, all related concepts, including the periodogram, are estimators, i.e. stochastic variables, which produce estimates. Considered as a sequence of stochastic variables, the periodogram $\underline{S}_j$ is a nonstationary stochastic process indexed by $j = 1, \ldots, \lfloor n/2 \rfloor$. The same happens with the estimator $\underline{\hat{s}}_{\mathrm{d}}(\omega_j)$, which is a nonstationary stochastic process indexed by $j = 1, \ldots, n$, as well as with the covariance estimator $\underline{\hat{c}}_\eta$. The produced shapes in graphs of estimates indicate high variability and roughness, and thus possible peaks are most probably random effects. Note that by increasing the number of observations, the variability and roughness do not necessarily decrease (cf. (4.52), where $\left|\underline{u}_j\right|^2$ is multiplied by $2n$).

An illustration is given in Figure 4.4 for a time series generated from the discrete-time HK process, where several peaks appear, all of which are random effects. Simple techniques that help to understand that these are random effects are (a) to slide the starting point by a certain number of time steps forming different sequences with same length (Koutsoyiannis and Iliopoulou, 2024), and (b) to split the time series into two halves, three thirds, etc. (Koutsoyiannis and Georgakakos, 2006). Thes allow us to inspect whether the peaks appear systematically in all cases. Splitting the time series and taking the average of the different parts for the same frequency is a method of smoothing the periodogram (for details and other smoothing methods see Papoulis, 1991).

The least square trend (power law) of the spectrum estimates from autocovariance is also shown in the log-log spectrum plot of Figure 4.4 (bottom-right). The slope is -1.24, an inconsistent value as theoretically the slope cannot be steeper than -1 (the slope of the theoretical curve, also shown in the figure, is $1 - 2H = -0.6 > -1$). This inconsistency is not expected to be resolved by the aforementioned smoothing of the power spectrum. For these reasons, the use of the climacospectrum, instead of the power spectrum, is recommended for estimation of slopes (Koutsoyiannis, 2017).

4.11 Interval estimation and confidence intervals

An *interval estimate* of a parameter λ of a distribution function is an interval of the form (θ_1, θ_2), where θ_1 and θ_2 are functions of the observed sample vector $\boldsymbol{x}$, i.e., $\theta_1 = g_1(\boldsymbol{x})$ and $\theta_2 = g_2(\boldsymbol{x})$. If we replace the observed sample with the sample (or the part of a stochastic process), then the interval's limits become stochastic variables, $\underline{\theta}_1 = g_1(\underline{\boldsymbol{x}})$ and $\underline{\theta}_2 = g_2(\underline{\boldsymbol{x}})$. The interval $(\underline{\theta}_1, \underline{\theta}_2)$ is an interval estimator of the parameter λ.

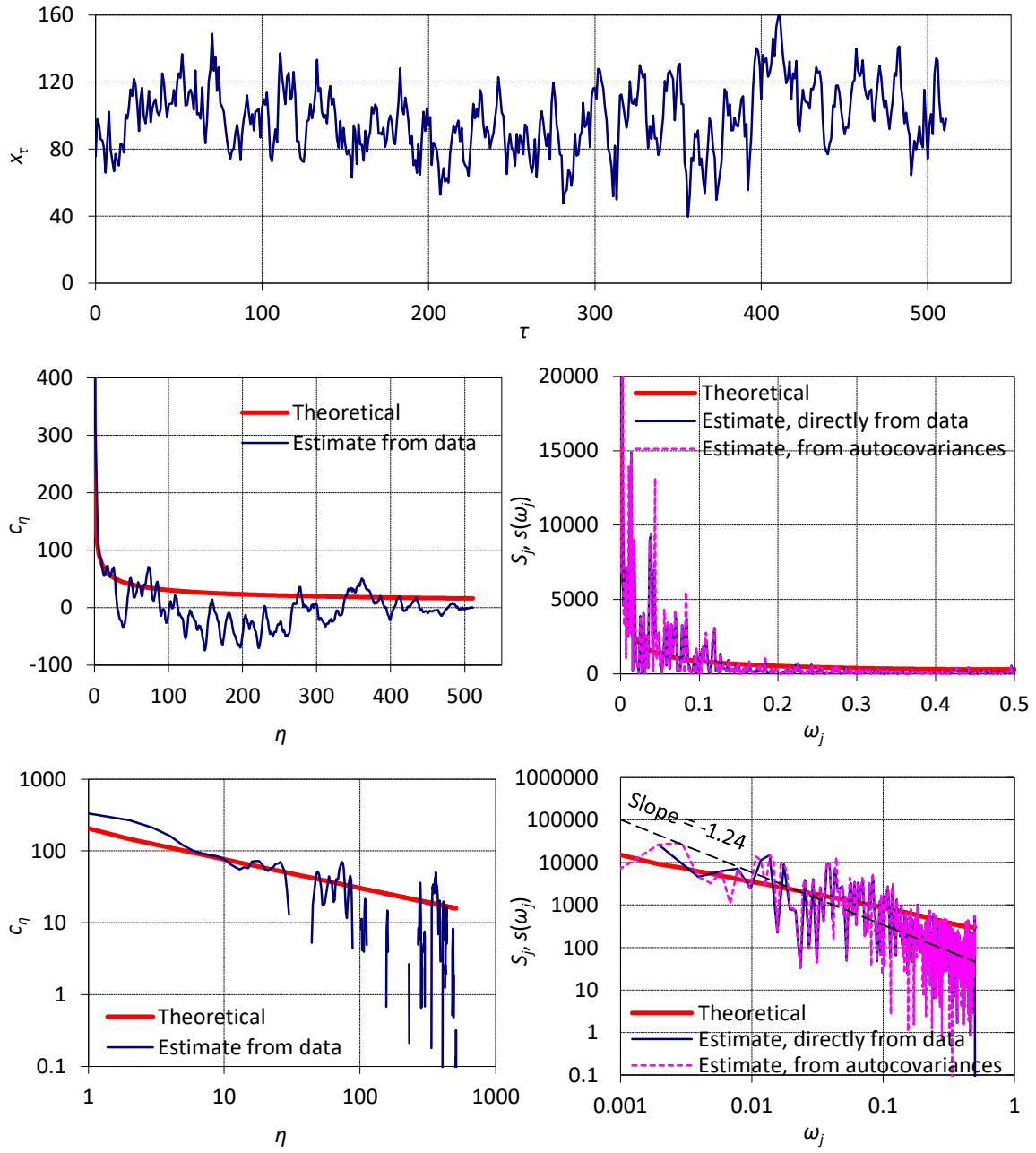

Figure 4.4 (**upper**) A plot of a time series with $n = 512$ terms generated from the Gaussian HK model with $H = 0.8, \mu = 100, \gamma_1 = 400$. (**middle**) The autocovariance and power spectrum of the generating stochastic process and their estimates. (**lower**) Same as middle but with logarithmic axes. The least square trend (power law) of the estimates from autocovariance, with slope $= -1.24$ is also plotted in the spectrum panel.

We say that the interval $(\underline{\theta}_1, \underline{\theta}_2)$ is a *C-confidence interval* of the parameter λ if:

$$P\{\underline{\theta}_1 < \lambda < \underline{\theta}_2\} = C \tag{4.55}$$

where C is a given constant $(0 < C < 1)$ called the *confidence coefficient*, and the limits $\underline{\theta}_1, \underline{\theta}_2$ are called *C-confidence limits*. Usually, we choose values of C near 1 (e.g. 0.9, 0.95, 0.99, so that the probability in (4.55) become near certainty). In practice the term

confidence limits is often (loosely) used to describe the numerical values of the statistics $\underline{\theta}_1, \underline{\theta}_2$, whereas the same happens for the term confidence interval.

In order to provide a method for the calculation of a confidence interval, we will assume that the statistic $\underline{\theta} = g(\underline{\boldsymbol{x}})$ is a point estimator of the parameter λ with distribution function $F_{\underline{\theta}}(\theta)$. Based on this distribution function it is possible to calculate two positive numbers ξ_1 and ξ_2, so that the estimation error $\underline{\theta} - \lambda$ lie in the interval ($-\xi_1$, ξ_2) with probability C, i.e.:

$$P\{\lambda - \xi_1 < \underline{\theta} < \lambda + \xi_2\} = C \tag{4.56}$$

and at the same time the interval ($-\xi_1$, ξ_2) be as small as possible. If the distribution of $\underline{\theta}$ is symmetric then the interval ($-\xi_1$, ξ_2) has minimum length for $\xi_1 = \xi_2$. For asymmetric distributions, it is difficult to calculate the minimum interval and, thus, we simplify the problem by splitting (4.56) into two equations, namely, $P\{\underline{\theta} < \lambda - \xi_1\} = P\{\underline{\theta} > \lambda + \xi_2\} = (1 - C)/2$. Equation (4.56) can be written as:

$$P\{\underline{\theta} - \xi_2 < \lambda < \underline{\theta} + \xi_1\} = C \tag{4.57}$$

Consequently, the confidence limits we are seeking are $\underline{\theta}_1 = \underline{\theta} - \xi_2$ and $\underline{\theta}_2 = \underline{\theta} - \xi_1$.

Although equations (4.56) and (4.57) are equivalent, their statistical interpretations differ. The former is a *prediction*, i.e., it gives the *prediction interval** of the stochastic variable $\underline{\theta}$. The latter is an *interval parameter estimator*, i.e., it gives the confidence limits of the unknown parameter λ, which is not a stochastic variable.

Classical statistical texts provide expressions for interval estimators of some common parameters, such as the mean and variance of the normal distribution of IID samples. However, in most real-world cases we deal with problems much more demanding than such idealized cases. The distributions may be non-normal, the parameter of interest may not be the mean or the variance, and instead of a sample we may have a stochastic process. Then analytical calculation of confidence limits becomes impossible. Naturally, the method of choice for such (that is, most) cases is the Monte Carlo simulation. General methodologies for tackling the problem have been proposed by Tyralis et al. (2013) and Tyralis and Koutsoyiannis (2014).

4.12 Order statistics

Let $\underline{x}$ be a stochastic variable and $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ be IID copies of it, forming a sample. We arrange them in increasing order such that $\underline{x}_{(i:n)}$ be the ith smallest of the n, i.e.:

$$\underline{x}_{(1:n)} \leq \underline{x}_{(2:n)} \leq \cdots \leq \underline{x}_{(n:n)} \tag{4.58}$$

The stochastic variable $\underline{x}_{(i:n)}$ is termed the ith *order statistic*. It may seem puzzling that stochastic variables can be ordered, as they are not numbers (but see also Digression 2.H).

* The terms confidence limits, confidence interval, confidence coefficient etc. are also used loosely for this prediction form of the equation.

To clarify this, we observe that given the numbers $x_1, x_2, \dots, x_n$, we can define the *i*th smallest of them as a deterministic function, $g(x_1, x_2, \dots, x_n)$:

$$x_{(i:n)} \coloneqq g(x_1, x_2, \dots, x_n) \coloneqq \min_{1 \le j \le n} \left(x_j \middle| \sum_{k=1}^{n} \mathrm{I}_{\{x_k \ge x_j\}} \ge i \right) \tag{4.59}$$

where I_A is the indicator function (with value equal to 1 when condition A is satisfied or 0 otherwise). Now if we substitute the stochastic variables $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n$ for the numbers $x_1, x_2, \dots, x_n$, we get $\underline{x}_{(i:n)} = g(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n)$, which, as we have seen, is a stochastic variable. Additional insights on stochastic ordering and its application to order statistics are provided by Shaked and Shanthikumar (2007).

The minimum and maximum order statistics are, respectively,

$$\underline{x}_{(1:n)} = \min(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n), \qquad \underline{x}_{(n)} \coloneqq \underline{x}_{(n:n)} = \max(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n) \tag{4.60}$$

and represent special cases of the order statistics, the lowest and the highest.

For a continuous variable $\underline{x}$, if $f(x)$ and $F(x)$ are respectively its density and its distribution function, then the density function of $\underline{y} \coloneqq \underline{x}_{(i:n)}$ is (Papoulis 1990):

$$f_{\underline{y}}(y) \coloneqq f_{(i:n)}(y) = (n - i + 1) \binom{n}{i - 1} \big(F(y)\big)^{i-1} \big(1 - F(y)\big)^{n-i} f(y) \tag{4.61}$$

Now if we define the stochastic variable $\underline{u} \coloneqq F\left(\underline{y}\right) = F(\underline{x}_{(i:n)})$, then according to (2.53):

$$f_{\underline{u}}(u) \coloneqq \frac{f_{\underline{y}}\big(F^{-1}(u)\big)}{f\big(F^{-1}(u)\big)} = (n - i + 1) \binom{n}{i - 1} u^{i-1}(1 - u)^{n-i} = \frac{u^{i-1}(1 - u)^{n-i}}{\mathrm{B}(i, n - i + 1)} \tag{4.62}$$

This is the density of the *Beta distribution* function and hence:

$$F_{(i:n)}(x) = P\{\underline{x}_{(i:n)} \le x\} = P\{\underline{u} \le F(x)\} = \frac{\mathrm{B}_{F(x)}(i, n - i + 1)}{\mathrm{B}(i, n - i + 1)} \tag{4.63}$$

For the special cases of the minimum and maximum we have, respectively,

$$F_{(1:n)}(x) = \frac{\mathrm{B}_{F(x)}(1, n)}{\mathrm{B}(1, n)} = 1 - \big(1 - F(x)\big)^n, F_{(n:n)}(x) = \frac{\mathrm{B}_{F(x)}(n, 1)}{\mathrm{B}(n, 1)} = \big(F(x)\big)^n \tag{4.64}$$

As we will see in section 4.13, the order statistics form the basis for estimating knowable moments and hence are quite important for studying extremes

4.13 Knowable moments and related estimations

As explained in Digression 4.B, classical moments are unknowable for orders p higher than 2-4, even for very large samples. Koutsoyiannis (2019a) introduced the so-called *knowable moments* (or *K-moment*), defined, after an adaptation by Koutsoyiannis (2025), as follows.

The *upper knowable moment (K-moment) of order p* is the expectation of the largest order statistic $\underline{x}_{(p)}$:

$$K'_p \coloneqq \mathrm{E}[\underline{x}_{(p)}] = \mathrm{E}\big[\max(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p)\big] \tag{4.65}$$

and the *lower knowable moment (K-moment) of order p* is the expectation of the smallest order statistic $\underline{x}_{(1:p)}$:

$$\overline{K}'_p := \mathrm{E}\left[\underline{x}_{(1:p)}\right] = \mathrm{E}\left[\min\left(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p\right)\right] \tag{4.66}$$

An important property, directly resulting from their definition, is that the K-moments are ordered as follows:

$$\overline{K}'_p \le \cdots \le \overline{K}'_2 \le \overline{K}'_1 = K'_1 = \mu \le K'_2 \le \cdots \le K'_p \tag{4.67}$$

These moments are noncentral, and we can also define central moments as

$$K_p := K'_p - K'_1, \qquad \overline{K}_p := \overline{K}'_1 - \overline{K}'_p, \qquad K_p, \overline{K}_p \ge 0 \tag{4.68}$$

As shown in Koutsoyiannis (2025, chapter 6), for a stochastic variable $\underline{x}$ of continuous type, the upper K-moment of order p of $\underline{x}$ is theoretically calculated as follows:

$$K'_p = p\mathrm{E}\left[\left(F\left(\underline{x}\right)\right)^{p-1}\underline{x}\right] = p\int_{-\infty}^{\infty}\left(F(x)\right)^{p-1}x\,f(x)\,\mathrm{d}x = p\int_0^1 x(F)\,F^{p-1}\mathrm{d}F \tag{4.69}$$

Likewise, the lower K-moment of order p is theoretically calculated as follows:

$$\overline{K}'_p = p\mathrm{E}\left[\left(\overline{F}\left(\underline{x}\right)\right)^{p-1}\underline{x}\right] = p\int_{-\infty}^{\infty}\left(\overline{F}(x)\right)^{p-1}x\,f(x)\,\mathrm{d}x = p\int_0^1 x(\overline{F})\,\overline{F}^{p-1}\mathrm{d}F \tag{4.70}$$

The latter two equations allow for the extension of the evaluation of K-moments for non-integer order p for a stochastic variable $\underline{x}$ of continuous type. For discrete-type variables as well as for generalizations of K-moments, the interested reader is referred to Koutsoyiannis (2023, 2025).

The unbiased estimator of the upper K-moment $\underline{K}'_p$ from a sample of size n is

$$\underline{\widehat{K}}'_p = \sum_{i=1}^{n} b_{inp}\,\underline{x}_{(i:n)} \tag{4.71}$$

and that of the lower K-moment is

$$\underline{\widehat{\overline{K}}}'_p = \sum_{i=1}^{n} b_{inp}\,\underline{x}_{(n-i+1:n)} = \sum_{i=1}^{n} b_{n-i+1,n,p}\,\underline{x}_{(i:n)} \tag{4.72}$$

where

$$b_{inp} = \begin{cases} 0, & i < p \\ p\dfrac{\Gamma(n-p+1)}{\Gamma(n+1)}\dfrac{\Gamma(i)}{\Gamma(i-p+1)}, & i \ge p \ge 0 \end{cases} \tag{4.73}$$

while $\Gamma(\,)$ is the gamma function and $\underline{x}_{(i:n)}$ is the ith *order statistic*. Note that for $p = 1$, $b_{in1} = 1/n$ and $\underline{\widehat{K}}'_1$ is identical to the estimator of the mean. At the other end, for $p = n$, $b_{inn} = 0$ for $i < n$ and $b_{nnn} = 1$, so that $\underline{\widehat{K}}'_n = \underline{x}_{(n)} := \underline{x}_{(n:n)} = \max\left(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n\right)$, in full compliance with (4.66). For integer moment order p and $i \ge p \ge 0$, this simplifies to

$$b_{inp} = \binom{i-1}{p-1} \Big/ \binom{n}{p} \tag{4.74}$$

Based on the K-moments, we define the *K-centre of order p*, C_p, measuring central tendency, and the *K-spread of order p*, D_p, measuring dispersion, as:

$$C_p := \frac{K'_p + \overline{K}'_p}{2}, \qquad D_p := \frac{K'_p - \overline{K}'_p}{2} \tag{4.75}$$

where $D_p \geq 0$. Their values for orders 1 and 2 are:

$$C_1 = K'_1 = \overline{K}'_1 = \mu, \;\; C_2 = \frac{K'_2 + \overline{K}'_2}{2} = K'_1 = C_1, \;\; D_1 = 0, \;\; D_2 = \frac{K'_2 - \overline{K}'_2}{2} = K_2 = \overline{K}_2 \tag{4.76}$$

In other words, the first and second order K-centre parameters are equal to each other and equal to the mean, while the second order K-spread is equal to the second order central K-moment. Geometrical depiction of these parameters for orders 1 to 4 is given in Figure 4.5.

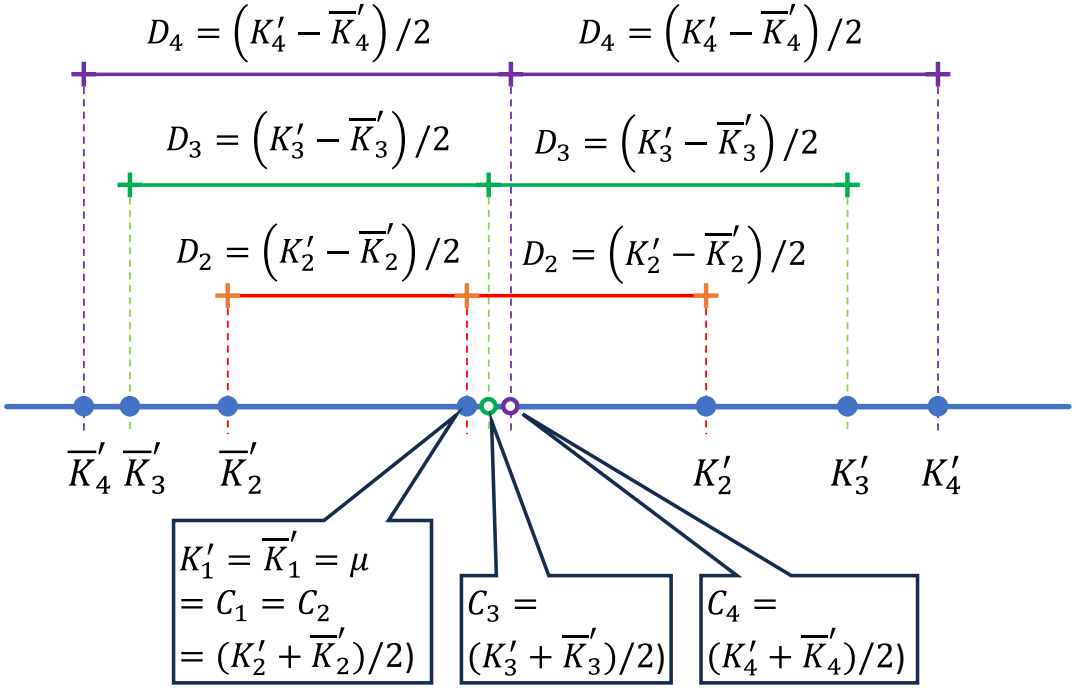

Figure 4.5 Explanatory sketch for the definition of summary statistics based on K-moments. (Source: Koutsoyiannis, 2025.)

Like the order statistics, K-moments offer a means to estimate probabilities from observations in a sample. Koutsoyiannis (2025) offers a detailed framework for this task with several options, from rough approximations to fully accurate derivations. Here, we reproduce a simple method which provides good approximations. This is based on the Λ-coefficient of order *p*, defined as:

$$\Lambda_p := \frac{1}{p\left(1 - F\left(K_p'\right)\right)} \tag{4.77}$$

The coefficients Λ_p happen to vary only slightly with p. Any symmetric distribution will give exactly $\Lambda_1 = 2$ because K_1' is the mean, which will equal the median and thus yield $F\left(K_p'\right) = 1/2$. Thus, a rough approximation is the rule of thumb:

$$\Lambda_p \approx 2 \tag{4.78}$$

Generally, the exact value Λ_1 is easy to determine, as it is directly related to the mean:

$$\Lambda_1 = \frac{1}{1 - F(\mu)} \tag{4.79}$$

The exact value of Λ_∞ depends only on the tail index ξ of the distribution:

$$\Lambda_\infty = \begin{cases} \Gamma(1-\xi)^{1/\xi}, & \xi \neq 0 \\ \mathrm{e}^{\gamma}, & \xi = 0 \end{cases} \tag{4.80}$$

where $\gamma = 0.577$ is the Euler's constant. The two values Λ_1 and Λ_∞ enable the simple approximation of Λ_p and hence of the non-exceedance probability:

$$\Lambda_p \approx \Lambda_\infty + \frac{\Lambda_1 - \Lambda_\infty}{p}, \qquad F\left(K_p'\right) \approx 1 - \frac{1}{\Lambda_\infty p + (\Lambda_1 - \Lambda_\infty)} \tag{4.81}$$

Hence, the estimate of return period of K_p' *is*

$$\frac{T\left(\widehat{K}_p'\right)}{D} = \frac{1}{1 - F\left(\widehat{K}_p'\right)} \approx \Lambda_\infty p + (\Lambda_1 - \Lambda_\infty) \tag{4.82}$$

For example, for the highest possible $p = n$ and for distributions with exponential tail ($\xi = 0$) this yields $T\left(\widehat{K}_n'\right)/D \approx 1.781n + \Lambda_1 - \Lambda_\infty$, which is consistent with the equations in Digression 2.I.

Conversely, for a given non-exceedance probability F, we can calculate the quantile x as the K_p' that corresponds to:

$$p \approx \frac{1}{\Lambda_\infty(1 - F)} + 1 - \frac{\Lambda_1}{\Lambda_\infty} \tag{4.83}$$

The estimate of K_p' based on the typical estimator $\underline{\widehat{K}}_p' = \sum_{i=1}^{n} b_{inp}\, \underline{x}_{(i:n)}$ is more reliable than that based on a single $x_{(i:n)}$ because it is derived from many data points (except when $i = n$, where the two approaches are precisely identical). In addition, equations (4.82)–(4.83) can give quantiles x for arbitrary values of F (not only the values contained in the sample) as the value of p should not necessarily be an integer but can be a real number. The only restriction is $p \leq n$.

Likewise, we can introduce the tail-based Λ-coefficient of order p as:

$$\overline{\Lambda}_p := \frac{1}{p\, F\left(K_p'\right)} \tag{4.84}$$

$\overline{\Lambda}_p$ has similar properties with Λ_p and in particular varies only slightly with p. For $p = 1$ it is readily seen that

$$\overline{\Lambda}_1 = 1/F(\mu) = \Lambda_1/(\Lambda_1 - 1) \tag{4.85}$$

For example, for a symmetrical distribution $\overline{\Lambda}_1 = \Lambda_1 = 2$. The limiting value $\overline{\Lambda}_\infty$ depends only on the lower tail index ζ of the distribution:

$$\overline{\Lambda}_\infty = \Gamma(1 + 1/\zeta)^{-\zeta} \tag{4.86}$$

For example, for $\zeta = 1$ (e.g. in exponential and Pareto distributions), $\overline{\Lambda}_\infty = 1$.

A simple approximation of $\overline{\Lambda}_p$ and hence of the non-exceedance probability is:

$$\overline{\Lambda}_p \approx \overline{\Lambda}_\infty + \frac{\overline{\Lambda}_1 - \overline{\Lambda}_\infty}{p}, \qquad F\left(K_p'\right) \approx \frac{1}{\overline{\Lambda}_\infty p + \left(\overline{\Lambda}_1 - \overline{\Lambda}_\infty\right)} \tag{4.87}$$

Conversely, for a given non-exceedance probability F, we can calculate the quantile x as the $\overline{K}_p'$ that corresponds to:

$$p \approx \frac{1}{\overline{\Lambda}_\infty F} + 1 - \frac{\overline{\Lambda}_1}{\overline{\Lambda}_\infty} \tag{4.88}$$

Combining the noncentral and tail moments we can produce estimates of quantiles for a wide range of probabilities of non-exceedance, namely for:

$$\frac{1}{\overline{\Lambda}_1 + \overline{\Lambda}_\infty(n-1)} \le F \le 1 - \frac{1}{\Lambda_1 + \Lambda_\infty(n-1)} \tag{4.89}$$

Additional information on several aspects of K-moments, including the adaptation of their estimators in the case that there is (long-range) dependence in the process of interest, can be found in Koutsoyiannis (2025).

Chapter 5. Stochastics as a tool to comprehend the microcosmos

5.1 Can common logic be reconciled with the quantum world?

The quantum world is considered to be incredibly weird, operating under rules that defy everyday logic and experience. This perception is reinforced by the (unnecessarily) weird mathematical notation adopted in quantum mechanics, which fosters the notion of mysteries that are unique to the microcosmos. Among the several behaviours regarded as weird, we examine in this chapter the two most prominent, the indistinguishability of particles and their duality as both waves and particles. The aim is to show that the weirdness disappears if we study them within a proper stochastic framework, similar to that used for macroscopic phenomena.

The idea that particles in the microscopic world are indistinguishable, interchangeable and without identity has been central in quantum physics. In quantum mechanics, two objects are regarded as identical whenever they have the same architecture and the same values of quantum numbers expressing their state. However, the same idea has been employed in statistical thermodynamics even in a classical framework of analysis (e.g. Wannier, 1987; Robertson, 1993; Stowe, 2007; Ben-Naim, 2008) to help make theoretical results agree with experience or perception, as well as with pre-existing thermodynamic results. Namely, the indistinguishability hypothesis has been central to determining the entropy in the kinetic theory of gases. In this case, the idea has been accepted despite the absence of direct experimental evidence supporting it (e.g., Papoulis, 1991, p. 11).

In the kinetic theory of gases, it is well known that the energy and momentum are taken to be continuous variables, as in classical physics, rather than discrete variables taking on a finite number of values as in quantum physics. Therefore, the probability that any two particles in motion have the same velocity, momentum and energy is zero. This can hardly justify their indistinguishability (even if the architecture of particles is identical) and suffices to dismiss the idea of the indistinguishability of molecules within the kinetic theory of gases. In Chapter 6 we will show that this idea resulted from superficial application of the entropy definition and that it can be fully abandoned without any problem. Rather, its dismissal resolves issues that are regarded as paradoxes—the well-known Gibbs paradox (see also Koutsoyiannis, 2013a).

In quantum mechanics, it looks difficult to dismiss the indistinguishability hypothesis. Also, the idea looks justifiable in typical quantum mechanical systems, as the number of states (dimension of Hilbert space) that describe what may happen in a finite volume is always finite (usually small) and therefore the probability of having particles in identical states is non-zero. However, as will be shown in section 5.2, it is rather easy to dismiss the indistinguishability idea and reestablish ordinary logic, as the related statistical behaviours can be recovered by using proper stochastics and assuming dependence among distinguishable particles.

The duality of particles as both waves and particles looks difficult to comprehend and reconcile with common logic developed from macroscopic phenomena. This duality is

exemplified by the famous double-slit experiment where particles act like waves (Figure 5.1).

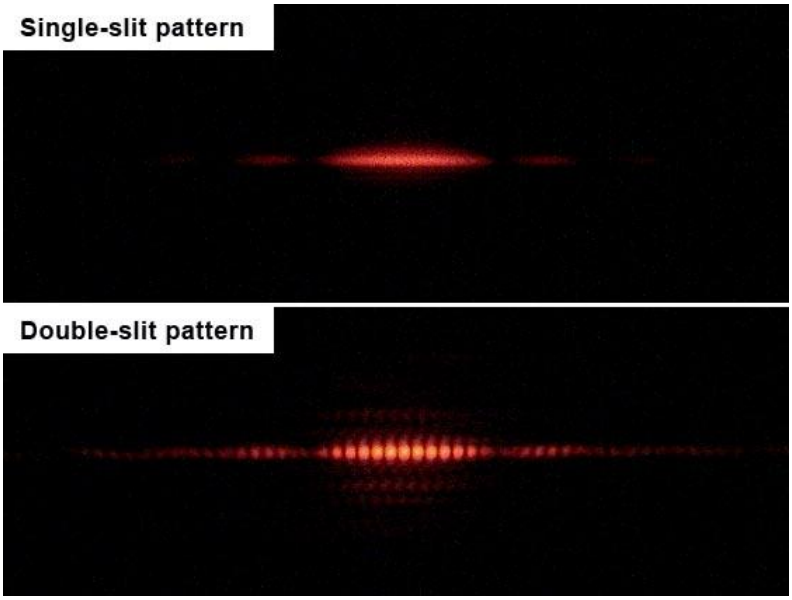

Figure 5.1 Patterns from the double-slit experiment, namely from a single slit (one of the two slits closed) vs. a double slit. Source: Wikimedia commons, https://commons.wikimedia.org/wiki/File:Single_slit_and_double_slit2.jpg, licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license.

This experiment is one of the most famous in modern physics and several versions thereof are described in experimental physics books (Beck, 2012; Prutchi, 2012), as well as in popular science books (Ananthaswamy, 2019). Its importance is highlighted in the following words by Feynman (1985):

> *I will take just this one experiment, which has been designed to contain all of the mystery of quantum mechanics, to put you up against the paradoxes and mysteries and peculiarities of nature one hundred per cent. Any other situation in quantum mechanics, it turns out, can always be explained by saying 'You remember the case of the experiment with the two holes? It's the same thing.*

But again we can tackle this paradox or mystery using probability. In the macroscopic world we are familiar with a random experiment (e.g. a die throw) and we are not surprised that initially there are many possibilities but eventually only one of them is realized. As already discussed in section 1.1, this was theorized by Aristotle in his dipole *potentiality* vs. *actuality*. This dismisses the deterministic dream that, given specified causes, only one outcome is possible. Aristotle's idea was adopted by Heisenberg (1962) and other modern physicists.

Hence it takes accepting probability as an abstract reality, to reinstate consistency and unity between the macroscopic random experiment and the behaviour of the quantum world. Initially, an emitted photon is a potentiality, described in probabilistic terms, and eventually it is realized as a particle. Not only does this glorify probability, but also confirms its existence, along with the sensible and other abstract objects. And its existential properties are so strong that we must also accept probability in the form of

waves — waves that travel exactly like waves of physical quantities. This is precisely the same logical structure we already used for macroscopic random experiments (section 1.1) — only now the potentialities propagate as waves. We examine this probabilistic explanation of the double-slit experiment in section 5.4, but before that we discuss the meaning of existence in Digression 5.A.

Digression 5.A: From the physical world to an abstract world

Most people agree that the physical world exists. For example, most will agree that the apple and the orange shown in Figure 5.2 exist, or did exist some time. Common-sense realism does not question the existence of the apple and the orange, but, notably, there are philosophical currents that do. For instance, solipsism would assert that only a person's (the photographer's or the eater's) experience was certain. Any "agreement" by another person (e.g. a reader of this text) could be part of the former person's dream or simulation or hallucination.

Figure 5.2 An apple and an orange on a table. This is not an image produced by artificial intelligence but a photograph (taken by the author) of two real fruits, which no longer exist (they were eaten shortly after the photograph was taken).

To continue, let's dismiss solipsism and other related '-isms', adopt the common-sense realism, and examine in this light the statement "the picture shows two fruits". Excepting those who do not mix apples with oranges, others would agree that this statement is true. But now either of the terms "fruit" and "two" describes not a physical but an abstract concept. Can we say that concept "two" or the number 2 exists? And if yes, can we also say that the number $\pi = 3.14159\ldots$ exists? What about the imaginary unit i and the complex numbers? Furthermore, does an ideal circle, whose ratio of its circumference to its diameter is π, exist? All these questions have diverse answers, depending on the ontology adopted. The replies that safeguard an easier life for one who studies physics using mathematics are affirmative to all these questions.

All these exist as abstract, non-spatiotemporal, mind-independent entities. This idea goes back to Plato's teaching, according to which the real world is a world of ideal or perfect forms (*αρχέτυπα*, archetypes). It is unchanging and unseen, and it can only be perceived by reason (*νοούμενα*, nooumena). The physical world is an imperfect image of the world of archetypes. Physical objects and events are "shadows" of their ideal forms, are subject to change and can be perceived by senses (*φαινόμενα*, phenomena). By turning Plato's theory upside-down we obtain a view that is more consistent with modern science: the physical world is the perpetually changing real world, but abstract concepts are also necessary to comprehend the real world (Koutsoyiannis and Montanari, 2015). Their indispensability to science pushes toward their existence—and this is supported by the philosophical current called *Mathematical Platonism*, according to which mathematical truths are discovered, not invented. Famous proponents of this current are Gottlob

Frege, Kurt Gödel, René Thom and Roger Penrose. But of course, there are many other '-isms', which do not accept such existence.

Stochastic variables, as we defined them and discussed in section 2.4, signify a higher level of abstraction. They are equivalent to a mathematical function, and as we prompted in section 2.4, an intuitive way to comprehend them is to think that, unlike common variables, they take on the entire set of possible values at the same time. The probabilities of taking on these values possibly differ among the different values, and are specified by their distribution functions. Probability per se is another concept of high-level abstraction, also a mathematical function, but one which maps sets onto numbers. And entropy is a concept relying on probability, as defined in section 2.3. Do stochastic variables, probability and entropy exist? Many would reply in the negative to this question, particularly those embracing the subjective, also known as Bayesian interpretation of probability. Others would agree about their existence—most prominently Popper (1982), who connected probability to *propensity*, a notion analogous to Aristotelian *potentiality*.

On the other hand, many would accept entropy as a physical entity, a property of a thermodynamic state that can be determined from observed quantities such as pressure, volume and temperature (recall Jaynes's view discussed in Digression 2.F). Without accepting entropy as existing, physics would collapse, the second law of thermodynamics would disappear, and engineering and technology would return to the pre-industrial era.

But is it possible for entropy to exist, if we refuse existence to its foundation, that is, probability? Even if we bypass this question trying to ground entropy in different ways, similar questions will emerge in quantum systems. In particular, the double-slit experiment (Figure 5.1) and Schrödinger's wavefunction would hardly be meaningful without invoking probability.

In this book we follow a pragmatic ontological approach based on the principle of parsimony: we shape the minimum set of assumptions that makes the world comprehensible and our lives easier (assuming that we live), both at an individual and a social level (assuming that others also live and interact with each other). Our minimal pragmatic approach includes the following premises:

- The physical world exists in an objective manner and its phenomena are mind-independent.
- Abstract concepts, which are not (but can be assigned to) physical entities, can exist, also being mind-independent.
- From the possible abstract concepts we adopt a minimal set that helps us comprehend the world.
- This minimal set includes natural, real and complex numbers, common variables and probabilistic concepts, including stochastic variables, probability per se, and entropy.

Clearly, including probabilistic concepts in the last premise is tantamount to extinguishing determinism. Most are reluctant to think outside of the deterministic framework (cf. Einstein's famous aphorisms in the beginning of this book) and it is reasonable to expect that theories trying to save determinism would emerge. Several attempts at this were made, most of which have been falsified by now. Currently, the most fashionable attempt is the multiverse hypothesis, in which a universe "splits" into two or more all the time. Whenever a quantum particle (e.g. an electron), which initially exists in multiple states at once (e.g., spinning both up and down), "decides" to realize in one of these states, a universe splits. Hence, instead of assuming that a stochastic variable realizes as a common variable, this hypothesis accepts that all possible realizations occur, inflating the number of universes to infinity.

However, there is no need to consider this hypothesis at all, as it does not affect our study. If it is false, we can disregard it from the outset. If it is true, we can continue our study while again disregarding it, leaving other copies of ourselves in other universes to work on it.

5.2 Particle indistinguishability vs. dependence

The indistinguishability hypothesis is not about whether or not we can distinguish or label particles. It goes far beyond the fact that the particles are identical, implying that this

property affects the probabilistic behaviour of particles and the entropy of the system. Thus, the hypothesis has resulted in different probabilistic behaviours or models, each one labelled by two names of some of the most respected physicists in history. They depart from standard probability and statistics, which we use in all sciences, including physics of the macroscopic world. The typical probability-theoretic problem of placing N particles into M boxes (representing particle states or locations) can serve as a basis for defining the models and distinguishing them from each other and from the standard probability model. Specifically, the number W of possible ways of placing the particles according to the different models are (Papoulis, 1991, pp. 11, 61; Ben-Naim, 2008, pp. 259-264, 319-320):

- Standard probabilistic model (for distinguishable particles):

$$W(M,N) = M^N \tag{5.1}$$

- Bose-Einstein statistics (each of N indistinguishable bosons occupies each of M states with no restriction on the occupation number in each state):

$$W_{\mathrm{BE}}(M,N) = \binom{N+M-1}{N} = \frac{(N+M-1)!}{N!\,(M-1)!} \tag{5.2}$$

- Fermi-Dirac statistics (each of N indistinguishable fermions occupies each of $M > N$ states with the restriction that no more than one particle can occupy a specific state):

$$W_{\mathrm{FD}}(M,N) = \binom{M}{N} = \frac{M!}{N!\,(M-N)!}, \qquad M \geq N \tag{5.3}$$

- Maxwell-Boltzmann statistics, lacking clear definition, but rather introduced as a mathematical trick to modify the entropy of systems of gas molecules, in which the standard model results in an entropic form that is non-extensive, while classical thermodynamics demands that the entropy be extensive (Koutsoyiannis, 2013a; see also section 6.6):

$$W_{\mathrm{MB}}(M,N) = \frac{W}{N!} = \frac{M^N}{N!} \tag{5.4}$$

An interesting property is that for $M >> N$, both the Bose-Einstein and the Fermi-Dirac statistics, become equivalent to the Maxwell-Boltzmann statistics, i.e.:

$$W_{\mathrm{BE}}(M,N) \approx W_{\mathrm{FD}}(M,N) \approx \frac{M^N}{N!} = W_{\mathrm{MB}}(M,N) \tag{5.5}$$

To show this we take the limit of $W_{\mathrm{BE}}(M,aM)/W_{\mathrm{MB}}(M,aM)$ as $a \to 0$, which turns out to equal 1. With a similar method, we can easily show that for $N >> M$:

$$W_{\mathrm{BE}}(M,N) \approx \frac{N^{M-1}}{(M-1)!} = W_{\mathrm{MB}}(N,M-1) \tag{5.6}$$

A relationship of this type for the Fermi-Dirac statistics is meaningless because by its definition it requires $N \leq M$.

The correctness of the above mathematical expressions is not questioned, except in the case of equation (5.4), where even simple inspection shows that it cannot be consistent. For example, taking $M = 3$ boxes and $N = 2$ particles, from (5.4) we obtain a number of possible arrangements $W_{\mathrm{MB}}(3,2) = 4.5$, which is absurd—this number cannot be non-integer (see additional problems in Koutsoyiannis, 2013a). By increasing the number of particles, i.e. $N = 9$ for $M = 3$, W_{MB} becomes lower than 1 ($W_{\mathrm{MB}}(3,9) = 0.054$), which is even more absurd, indicating a probability greater than 1 ($1/W_{\mathrm{MB}} = 18.44$).

Thus $W_{\mathrm{MB}}(M, N)$ is only useful as a numerical approximation for $M \gg N$, according to equation (5.5) (or for $N \gg M$ according to equation (5.6)). Otherwise it has no logical basis, it does not describe anything logically consistent, and it creates confusion. Therefore, we will not use it.

On the other hand, the Bose-Einstein and Fermi-Dirac equations are accurate and useful. However, the idea of indistinguishability, which serves as these equations' logical basis, is a weird one and leads to the paradoxical perception that common logic does not apply to the microcosmos. However, if microcosmos is logically different from macrocosmos, the entire body of logic becomes problematic. A fundamental question which reveals an inconsistency is this: What would be the threshold that distinguishes the two cosmoses and hence the two different logics?

Here to avoid such interpretation and the implied problematic situations, we reject the idea of indistinguishability and the different logics that it implies. Specifically, we derive all three equations (5.1)–(5.3) using a very simple assumption related to the dependence of particles, namely, zero, positive and negative dependence for equations (5.1) (classical model), (5.2) (Bose-Einstein) and (5.3) (Fermi-Dirac), respectively. From a physics point of view, zero, positive and negative dependence can be conceived as neutrality, attraction and repulsion between particles, respectively, in terms of their states. As dependence and independence are part of the standard logic and are quantified in stochastics through the concept of conditional probability, there is no need to introduce a new logic for microcosmos.

We consider the typical probability-theoretic problem of placing N particles into M boxes, representing particle states or locations. Let $n_m^N \le N$ denote the particular number of particles that are contained in box numbered m, $1 \le m \le M$, out of N particles in total. We consider the following event, which we call a *configuration* of $N - 1$ particles:

$$A^N := \{\underline{n}_1^N = n_1^N, \dots, \underline{n}_m^N = n_m^N, \dots, \underline{n}_M^N = n_M^N\} \tag{5.7}$$

Assuming, of course, that the particles are distinguishable, as common logic dictates, there are multiple *arrangements*, i.e. elementary events, corresponding to the configuration A^N. Their number is given by the multinomial coefficient, which, by definition, is the number of ways of partitioning N distinct objects into M sets of sizes $n_1, \dots, n_m, \dots, n_M$. This number is:

$$C(n_1, \dots, n_m, \dots, n_M) := \frac{N!}{n_1! \dots n_m! \ \dots n_M!} \tag{5.8}$$

We will study the transition of A^N into the configuration A^{N+1}, in which one box will contain one more particle. We let this to be the mth box, so that

$$A_m^{N+1} \coloneqq \{\underline{n}_1^{N+1} = n_1^N, , \dots, \underline{n}_m^N = n_m^N + 1, \dots, \underline{n}_M^N = n_M^N\} \tag{5.9}$$

We make the fundamental assumption that the probability of the new particle choosing a particular state (box) is linearly dependent on the number of particles that are already there:

$$P(A_m^{N+1}|A^N) \propto 1 + kn_m^N \tag{5.10}$$

where k is a dimensionless parameter (we call it a quantum number here) which physically encodes attraction (bosons, $k > 0$), neutrality ($k = 0$), or repulsion/exclusion (fermions, $k < 0$) in state space. Specifically, we set:

$$k = \begin{cases} -1, & \text{negative dependence} - \text{fermions} \\ 0, & \text{independence} - \text{classical particles} \\ +1, & \text{positive dependence} - \text{bosons} \end{cases} \tag{5.11}$$

There are M different configurations A_m^{N+1}, each corresponding to $m = 1, \dots, M$, with conditional probabilities $P\{A_m^{N+1}|A^N\}$ adding up to:

$$\sum_{m=1}^{M} P(A_m^{N+1}|A^N) = 1 \tag{5.12}$$

Summing probabilities for all m options we find:

$$\sum_{m=1}^{M} P(A_m^{N+1}|A^N) \propto M + kN \tag{5.13}$$

and hence:

$$P(A_m^{N+1}|A^N) = \frac{1 + kn_m^N}{M + kN} \tag{5.14}$$

This is our simple and fundamental equation, from which we can readily derive any of the cases corresponding to equations (5.1)–(5.3). More specifically, for the case of independence ($k = 0$), we have:

$$P(A_m^{N+1}|A^N) = \frac{1}{M} \tag{5.15}$$

For the case of positive dependence (bosons, $k = 1$), we have

$$P(A_m^{N+1}|A^N) = \frac{1 + n_m^N}{M + N} \tag{5.16}$$

Finally, for the case of negative dependence (fermions, $k = -1$), we have:

$$P(A_m^{N+1}|A^N) = \frac{1 - n_m^N}{M - N} = \begin{cases} 0, & n_m^N = 1 \\ \dfrac{1}{M - N}, & n_m^N = 0 \end{cases} \tag{5.17}$$

The zero probability when $n_m^N = 1$ is precisely the stochastic encoding of the Pauli exclusion principle.

Next we will see how this rule applies to the three different kinds of dependence. Specifically, we will show that equations (5.15)–(5.17) imply equations (5.1)–(5.3).

Classical particles

In the independence case, it is readily shown that the probability of any arrangement of the N particles in M boxes is independent of the arrangement and equal to:

$$\left(\frac{1}{M}\right)^N = \frac{1}{W(M,N)} \tag{5.18}$$

However, the probability of a configuration depends on $n_1, \dots, n_m, \dots, n_M$, i.e.:

$$P(A^N) = C(n_1, \dots, n_m, \dots, n_M)\left(\frac{1}{M}\right)^N = \frac{N!}{n_1! \dots n_m! \ \dots n_M!}\left(\frac{1}{M}\right)^N \tag{5.19}$$

This is known as the multinomial distribution.

Bosons

We will show that the probability of each configuration is independent of $n_1, \dots, n_m, \dots, n_M$, and only depends on their sum $N = n_1 + \cdots + n_m + \cdots + n_M$ and the number of boxes M, namely:

$$P(A^N) = P_{\mathrm{BE}}(M,N) = \frac{N!\,(M-1)!}{(N+M-1)!} = \frac{1}{W_{\mathrm{BE}}(M,N)} \tag{5.20}$$

To show that equation (5.20) holds true for any N we use the complete (mathematical) induction. Apparently, it holds for $N = 1$:

$$P(A^1) = P_{\mathrm{BE}}(M,1) = \frac{(M-1)!}{M!} = \frac{1}{M} \tag{5.21}$$

We assume that it holds for N and we will show that it also holds for $N+1$. We study the configuration A_m^{N+1} as in (5.9). This can result from (5.7) by placing the additional particle in box m. Its probability is

$$P(A_m^{N+1}|A^N)P(A^N) = \frac{1+n_m^N}{M+N}\frac{N!\,(M-1)!}{(N+M-1)!} = \frac{N!\,(M-1)!}{(N+M)!}(1+n_m^N) \tag{5.22}$$

However, it can also result from other configurations A^N. The configuration $A_{(1)}^N = \left\{\underline{n}_1^N = n_1^N - 1, \dots, \underline{n}_m^N = n_m^N + 1, \dots, \underline{n}_M^N = n_M^N\right\}$ will also give the same A_m^{N+1} if the new particle is placed in box 1. As the total number of particles in $A_{(1)}^N$ is again N and since the probability of a configuration for N particles does not depend on the particular arrangement, but only on N and M, the probability for this case is

$$P\left(A_m^{N+1}\middle|A_{(1)}^N\right)P\left(A_{(1)}^N\right) = \frac{1+n_1^N-1}{M+N}\frac{N!\,(M-1)!}{(N+M-1)!} = \frac{N!\,(M-1)!}{(N+M)!}n_1^N \tag{5.23}$$

Likewise, we can have the cases of the configuration $A_{(2)}^N, \dots, A_{(M)}^N = \left\{\underline{n}_1^N = n_1^N =, \dots, \underline{n}_m^N = n_m^N + 1, \dots, \underline{n}_M^N = n_M^N - 1\right\}$ and hence, adding the probabilities for all cases we get

$$P(A_m^{N+1}) = \frac{N!\,(M-1)!}{(N+M)!}(n_1^N + \cdots + 1 + n_m^N + \cdots + n_M^N) = \frac{N!\,(M-1)!}{(N+M)!}(N+1) \tag{5.24}$$

or

$$P(A_m^{N+1}) = \frac{(N+1)!\,(M-1)!}{(N+M)!} \tag{5.25}$$

Thus, we have proved that equation (5.20) holds true for $N+1$ and hence for any N.

It is tempting to conjecture that the equality of probabilities of all configurations is the reason why particles were long thought to be indistinguishable. But in our approach the probability becomes perfectly uniform per configuration, even though the particles themselves remain distinguishable.

Contrary to the independence case, in bosons the probability of a specific arrangement, one of the many elementary events forming the configuration A^N, does depend on the partitioning of N into $n_1, \dots, n_m, \dots, n_M$. It is readily found to be:

$$\frac{P(A^N)}{C(n_1, \dots, n_m, \dots, n_M)} = \frac{N!\,(M-1)!}{(N+M-1)!}\,\frac{n_1! \dots n_m!\ \dots n_M!}{N!} = \frac{n_1! \dots n_m!\ \dots n_M!\,(M-1)!}{(N+M-1)!} \tag{5.26}$$

Apparently, this is maximum when all N particles are placed in a single box (say $n_1 = N, n_2 = \cdots = n_M = 0$, but irrespective of what the number of the box is). In this case the configuration has only one arrangement (elementary event) is equal to $P(A^N) = 1/W_{\mathrm{BE}}(M,N)$.

Fermions

For the case of Fermions, we will again show that the probability of each configuration is independent of $n_1, \dots, n_m, \dots, n_M$. It only depends on their sum $N = n_1 + \cdots + n_m + \cdots + n_M$ and the number of boxes M, where now each n_m is either 0 or 1. It is given by:

$$P(A^N) = P_{\mathrm{FD}}(M,N) = \frac{N!\,(M-N)!}{M!} = \frac{1}{W_{\mathrm{FD}}(M,N)}, \qquad M \geq N \tag{5.27}$$

Again we use the complete (mathematical) induction. Apparently, equation (5.27) holds for $N = 1$. We assume that it holds for N and we will show that it also holds for $N+1$. We study the configuration A_m^{N+1}. If $n_m = 0$ then A_m^{N+1} results by placing the additional particle in box m. Utilizing equation (5.17), its probability is

$$P(A_m^{N+1}|A^N)P(A^N) = \frac{1}{M-N}\,\frac{N!\,(M-N)!}{M!} = \frac{N!\,(M-(N+1))!}{M!} \tag{5.28}$$

If $n_m = 1$, there are N other configurations A^N, which can produce A_m^{N+1}. Each of them has a 0 in one of the N elements in which A_m^{N+1} has 1. The probability of each one is equal to the above and hence:

$$P(A_m^{N+1}) = (N+1)\,\frac{N!\,(M-(N+1))!}{M!} \tag{5.29}$$

or

$$P(A_m^{N+1}) = \frac{(N+1)!\,(M-(N+1))!}{M!} \tag{5.30}$$

Thus, we have proved that equation (5.27) holds true for $N+1$ and hence for any N. In other words, from the hypothesis of negative dependence and assuming distinguishable particles, we recovered the typical result of Fermi-Dirac statistics that is given for indistinguishable particles. Again we have equality of probabilities of all configurations, hence again giving a hint why particles were long thought to be indistinguishable.

To calculate the probability of a specific arrangement, one of the many elementary events forming the configuration A^N, we observe that, since all n_m are 0 or 1 and since $0! = 1! = 1$, the probability in question is:

$$\frac{P(A^N)}{C(n_1, \dots, n_m, \dots, n_M)} = \frac{N!\,(M-N)!}{M!}\,\frac{1}{N!} = \frac{(M-N)!}{M!} \tag{5.31}$$

That is, it is precisely the same for all arrangements.

While the above proofs are general and hold precisely, for the sake of better understanding we provide an illustration in Digression 5.B.

In summary, dismissing the artificial hypothesis of indistinguishability* and replacing it with the natural concept of dependence (quantified by a single parameter $k = -1, 0, +1$) recovers Bose–Einstein, Fermi–Dirac, and classical statistics entirely within standard probability theory. As per the Maxwell-Boltzmann statistics, its only meaning is to view it as a limit of the Bose-Einstein and the Fermi-Dirac statistics. Common logic therefore applies to the microcosmos exactly as it does to the macrocosmos — exactly as we use it for planet orbits but also dice throws or gas molecules.

Digression 5.B: Illustration of probabilities by replacing indistinguishability with dependence

We illustrate the stochastic framework presented above for the problem of placing N particles into M boxes, with varying $N = 1,2,3$ and constant $M = 3$. For better illustration we depict the calculations in figures, without applying the final formulae of section 5.2. Specifically, Figure 5.3 shows the calculations for positive dependence (bosons) and Figure 5.4 those for negative dependence. Further explanations are provided in the figure captions.

For the bosons case (Figure 5.3) we have also summarized the resulting probabilities per arrangement and per configuration for $(M = 3, N = 3)$ in Table 5.1. It is seen that the most probable configuration is all N particles occupying a single state. It is also seen that while the probabilities per arrangement vary from $1/60$ to $1/10$ (contrasting with the case of independence where they are constant, $1/27$), the probabilities per configuration are constant. As already mentioned this can perhaps be the reason why particles were long thought to be indistinguishable.

The characteristic probabilities of each of the arrangements and configurations of all cases examined ($N = 1, 2, 3, M = 3$; with independent, positively dependent and negatively dependent particles) are compared in Table 5.2. We notice the constant probability per arrangement in the occasion of independence and the constant probability per configuration in the dependent cases. It may have become apparent by now that the constant probability per configuration does not imply indistinguishability of the particles.

---

* A probabilistic derivation avoiding the indistinguishability postulate also appears in Kolokoltsov (2021). The present approach is distinct in retaining particle distinguishability and encoding dependence via attraction/repulsion in state space.

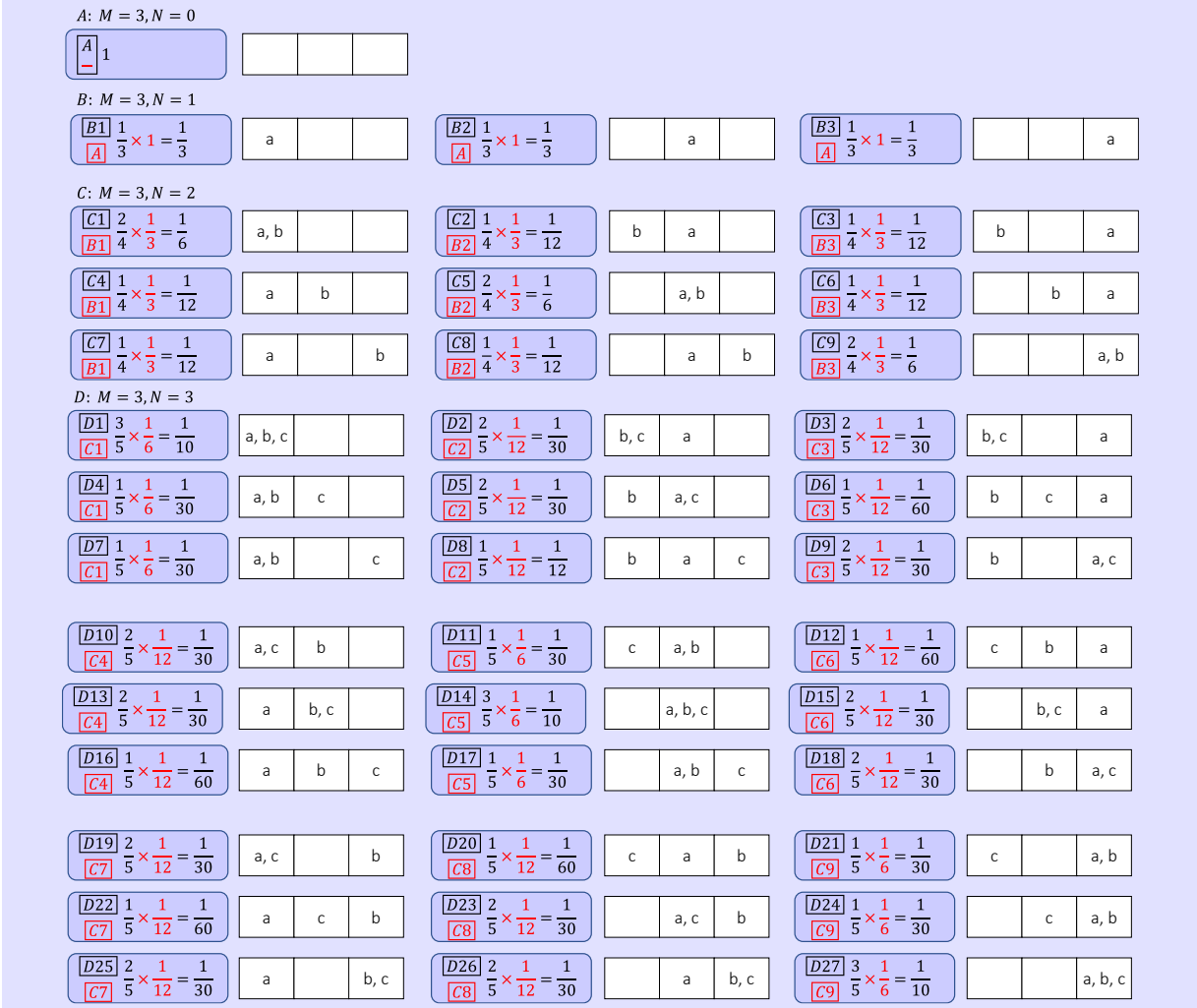

Figure 5.3 Illustration of calculating probabilities of all possible arrangements of placing N positively dependent particles (bosons), designated as a, b, c, into $M = 3$ boxes, where N is progressively increasing from 1 to 3. Each arrangement is assigned a particular label (B1, B2, etc.; upper left in each container of calculations) accompanied with the label of the predecessor arrangement (lower left in each container of calculations). In each calculation the first term is the conditional probability according to equation (5.16) and the second term is the probability that the condition holds (from the predecessor arrangement, i.e. the calculation is: conditional probability × predecessor probability equals probability of this arrangement).

Table 5.1 Characteristic probabilities of each of the arrangements and configurations in placing $N = 3$ positively dependent particles (bosons) into $M = 3$ boxes. The probability of each distinct arrangement (D1 to D27) has been calculated in Figure 5.3, which was also used to count the number of arrangements in each configuration (for example in configuration #2 there correspond the arrangements D2, D4, D10).

#	Characteristics of each configuration					Probability of each arrangement, $P_{\mathrm{BE}}(M, n_1, n_2, n_3)$	Probability of each configuration, $P_{\mathrm{BE}}(M, N) = P_{\mathrm{BE}}(M, n_1, n_2, n_3)C(n_1, n_2, n_3)$
	n_1	n_2	n_3	N	$C(n_1, n_2, n_3)$		
1	3	0	0	3	1	1/10	1/10
2	2	1	0	3	3	1/30	1/10
3	2	0	1	3	3	1/30	1/10
4	1	2	0	3	3	1/30	1/10
5	1	1	1	3	6	1/60	1/10
6	1	0	2	3	3	1/30	1/10
7	0	3	0	3	1	1/10	1/10
8	0	2	1	3	3	1/30	1/10
9	0	1	2	3	3	1/30	1/10
10	0	0	3	3	1	1/10	1/10

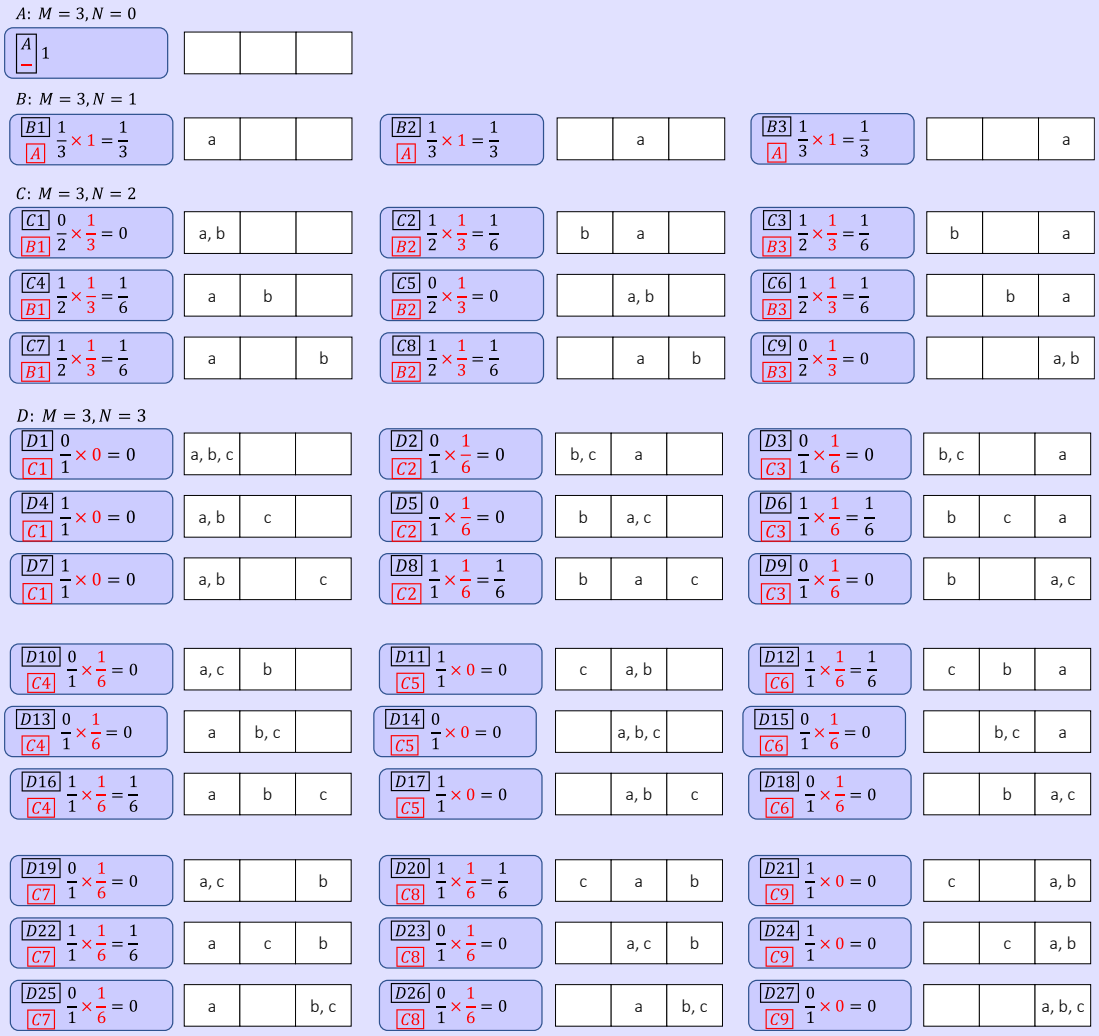

Figure 5.4 Illustration of calculating probabilities of all possible arrangements by putting up to $N = 3$ negatively dependent distinguishable particles (fermions), designated as a, b, c, into $M = 3$ boxes. Each arrangement is given a particular label (B1, B2, etc.; (upper in each container of calculations) accompanied with the label of the predecessor arrangement (lower left in each container of the calculations). In each calculation the first term is the conditional probability according to equation (5.17) and the second term is the probability that the condition holds, as calculated in the predecessor arrangement (i.e. the calculation is: conditional probability × predecessor probability equals probability of this arrangement).

Table 5.2 Characteristic probabilities of each of the arrangements and configurations in placing up $N = 1, 2$ or 3 distinct particles into $M = 3$ boxes. Note the constant probability per arrangement in the independence case of the constant probability per configuration in the dependence cases.

N	Total number of configu-rations*	Independence		Positive dependence (bosons)		Negative dependence (fermions)	
		Probability per arrangement	Probability per configuration	Probability per arrangement	Probability per configuration	Probability per arrangement	Probability per configuration
1	3	1/3	1/3	1/3	1/3	1/3	1/3
2	6 (3)	1/9	1/9 or 2/9	1/12 or 1/6	1/6	1/6	1/3
3	10 (1)	1/27	1/27 or 1/9 or 2/9	1/60 or 1/30 or 1/10	1/10	1/6	1

* In parentheses are given the numbers of legitimate configurations for Fermions.

5.3 From water waves to probability waves

We start discussing water waves because they are the most familiar ones as they enable a visual inspection (Figure 5.5). A disturbance at some point (e.g. due to a throw of a stone) will propagate radially, forming cylindrical waves. The energy will eventually be dissipated and the water surface will become calm again.

Figure 5.5 Realistic rendering of cylindrical (radially propagating) water waves (image generated by Grok—xAI).

Here we consider a continuous disturbance in a form of a harmonic oscillation at a single point ($r = 0$, where we use a polar coordinate system (r, θ) with r denoting the radial distance and θ the angle). We neglect energy dissipation and assume that the water surface motion reached a steady state. For easy illustration we study the simplest case of a wave with a single *frequency*, ν, and *wavelength*, λ (distance between two consecutive peaks). We assume a simplified wave model—even though as we will see below, this can only be used as a rough approximation for 2-dimensional waves. Specifically we assume that the displacement from the water level at rest, $h(r, t)$ at radial distance r and time t is given by

$$h(r, t) = a(r) \cos\, \varphi\,, \qquad \varphi \coloneqq \varphi_0 + 2\pi \frac{r - ct}{\lambda} \tag{5.32}$$

where $a(r)$ is the *amplitude* whose form will be specified below, φ the phase with φ_0 being an initial phase $r = 0$, $t = 0$, and c is the *wave velocity*, also known as *celerity*. The latter term is preferred when the water is in motion, where the water velocity V differs from the celerity c. The fact that in the assumed model the amplitude is a function of only the radial distance makes the waves cylindrical. Yet the waves shown in Figure 5.5 are more complex than those described by equation (5.32), which have an element of randomness.

For waves in shallow water the wave velocity is

$$c \approx \sqrt{gy} \tag{5.33}$$

where g is the gravitation acceleration and y is the mean water depth. The period T and the frequency $\nu \coloneqq 1/T$ of the wave are connected with the wave velocity through:

$$c = \frac{\lambda}{T} = \lambda\, \nu \tag{5.34}$$

When the wave at point r peaks, the kinetic energy is zero and the potential energy for an elementary area dA is calculated as follows. The mass of the volume displaced is $\rho h_\mathrm{m}\,\mathrm{d}A$, where ρ is the water density. The average displacement of this mass is $h_\mathrm{m}/2$ and therefore the potential energy is $(1/2)\rho\, g\, h_\mathrm{m}^2\,\mathrm{d}A$. Now, as we assumed a steady state, we have accepted that the total energy does not change in time. At lower displacement, kinetic energy is developed, being maximized when the displacement is zero, with maximum value $(1/2)\rho\, g\, h_\mathrm{m}^2\,\mathrm{d}A$.

As by virtue of equation (5.32) for t such that $\cos\varphi = 1$ we have $h_m = h(r,t) = a(r)$ and hence the total energy per unit area is

$$e(r) = (1/2)\,\rho\, g\, a(r)^2 \tag{5.35}$$

On the other hand, if we integrate the total energy per unit are at a circle of radius r, it should be constant:

$$2\pi\, r\, a(r)^2 = \text{constant} \tag{5.36}$$

and hence we can rewrite equation (5.32) as

$$h(r,t) = \frac{a}{\sqrt{r}}\cos\left(\varphi_0 + 2\pi\frac{r-ct}{\lambda}\right) \tag{5.37}$$

where a is now a constant with dimensions [L$^{3/2}$]. In this, the energy per unit area is $(1/2)\,\rho\, g\, a^2/r$ and that over the entire circle of radius r is $\pi\,\rho\, g\, a^2$. A characteristic depiction of the waves described by equation (5.37) is shown in Figure 5.6.

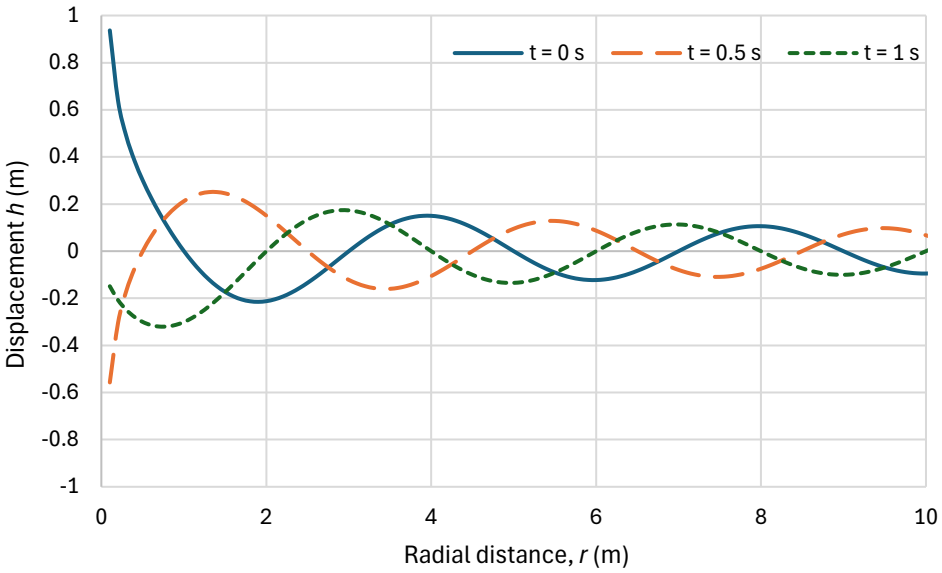

Figure 5.6 Displacement as a function of radial distance for water waves described by equation (5.37) for the indicated times. The assumed numerical constants are $\varphi_0 = 0, c = 3$ m/s (corresponding to water depth of about 1 m), $\lambda = 4$ m, $a = 0.3$ m$^{3/2}$, so that $T = c/\lambda = 0.75$ s, $\nu = 1/T = 1.33$ Hz. The total energy per unit radius is $2\pi\, r\, e(r) = 2\pi\, r\, (1/2)\,\rho\, g\, a(r)^2 = \pi\,\rho\, g\, a^2 = \pi$ 1000 kg m^{-3} 9.81 m s^{-2} 0.3^2 m^3 = 2773 kg m s^{-2} or 2.77 kJ/m.

From the standard theory of partial differential equations (e.g. Strauss, 2007), we recall that the wave equation is

$$\frac{\partial^2 h}{\partial t^2} - c^2 \nabla^2 h = 0 \tag{5.38}$$

where in two dimensions and for cylindrical symmetry (i.e. $\partial h/\partial\theta \equiv 0$, where θ is the angular coordinate), the Laplacian is

$$\nabla^2 h = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial h}{\partial r}\right) \tag{5.39}$$

After the algebraic manipulations we find

$$\frac{\partial^2 h}{\partial t^2} - \frac{1}{r}\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial h}{\partial r}\right) = -\frac{a\,c^2}{4\left(\sqrt{r}\right)^5}\cos\left(\varphi_0 + 2\pi\frac{r-ct}{\lambda}\right) \tag{5.40}$$

In other words, this is not zero (as it should), but tends to zero as r increases, at a rate of $r^{-5/2}$. The wave equation for large r is asymptotically satisfied, but not precisely. If we want to find an accurate representation for two dimensional waves, the exact solution is more complicated, involving Bessel or Hankel functions. However, for better illustration, here we preferred the simple approximate solution which better serves our focus.

When we have more intricate problems, such as many frequencies or interfering waves, it is too complicated to use trigonometric functions, and we replace them with exponential functions of complex numbers. That is, using Euler's formula, $\mathrm{e}^{\mathrm{i}x} = \cos x + \mathrm{i}\sin x$, where i is the imaginary unit, we rewrite equation (5.37) as

$$h(r,t) = \mathrm{Re}\,(\psi(r,t)), \qquad \psi(r,t) \coloneqq \frac{a}{\sqrt{r}}\exp\left(\mathrm{i}\left(\varphi_0 + 2\pi\frac{r-ct}{\lambda}\right)\right) \tag{5.41}$$

We may call $\psi(r,t)$ complex wave function and we note that:

- its real part Re (ψ) is the observable, in our case the displacement $h(r,t)$; and
- its absolute square, $|\psi|^2 = \psi\overline{\psi}$, which is the square of the coefficient of the exponential function (in our case $|\psi|^2 = \psi\overline{\psi} = \left(a/\sqrt{r}\right)^2 = a^2/r$) is related to the energy per unit space (in our case, unit area), or *energy density*.

If we go to three-dimensional waves, like the propagation of sound the same toolbox is applied with a slight modification. Namely, we should replace $\sqrt{r}$ with r because the waves are now spherical, rather than symmetrical, and the energy at distance r should be inversely proportional to r^2. The mathematical representation now becomes

$$h(r,t) = \mathrm{Re}\,(\psi(r,t)), \qquad \psi(r,t) \coloneqq \frac{a}{r}\mathrm{e}^{\mathrm{i}\varphi}, \qquad \varphi = \varphi_0 + 2\pi\frac{r-ct}{\lambda} \tag{5.42}$$

where the observable $h(r,t)$ now is the pressure amplitude in the carrying medium (air, water, etc.). The energy density is again proportional to $|\psi|^2 = \psi\overline{\psi}$, in our case $(a/r)^2$.

Unlike the case of water waves, in three dimensional waves the wave function is precisely satisfied by equation (5.42). Specifically, for spherical symmetry (i.e. $\partial h/\partial\theta = \partial h/\partial\eta \equiv 0$, where θ and η are the angular coordinates), the Laplacian is

$$\nabla^2 h = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial h}{\partial r}\right) \tag{5.43}$$

and after the algebraic manipulations we find

$$\frac{\partial^2 h}{\partial t^2} - \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial h}{\partial r}\right) = 0 \tag{5.44}$$

When moving from sound to light, the differences are conceptual, rather than in the mathematical treatment. We make the following points about them.

1. As we have discussed in Digression 1.A, light can simultaneously materialize multiple trajectories, once there are pertinent conditions, for example, mirrors in our experimental apparatus. This means that light fills the space, before materializing a specific trajectory (which eventually satisfies Fermat's principle). We can then conceptually perceive the light propagation in spherical waves.
2. Light is observable only when photons are emitted or absorbed. In between, nothing is observable. Therefore, if we express mathematically a quantity $\psi(r,t)$ as an amplitude, then the quantity $h(r,t) = \mathrm{Re}\,(\psi(r,t))$ is not observable and therefore it must be an abstract quantity.
3. This abstract quantity should imply energy propagation, proportional to the quantity $|\psi|^2 = \psi\overline{\psi} = (a/r)^2$ and becomes arbitrarily low for large r. On the other hand, a photon is a quantum that cannot be subdivided and has a specific energy content, $e = h\,\nu$, where h is the Planck constant. Given that there can be no lower energy amount than e, while $(a/r)^2$ can take arbitrarily low values, we understand that $f = \psi\overline{\psi}$ should be none other than a probability density that the photon should reach a certain point (r,θ,η). This is like when many people are vying for a car: Initially everyone has a probability of getting it, but eventually only one will have it.
4. Therefore, the situation becomes similar to a macroscopic random experiment. Let us consider for comparison a roulette wheel which is not divided into pockets, but its outcome is the angle θ measured on a circular scale, like in Digression 2.L (case c). When the wheel stops, the observed outcome is an actuality, but before that, when it spins, all possible outcomes are potentialities. The same happens with the photon. When it is absorbed at a point (r,θ,η), we have an actuality and a realization of the energy e at a single point via a single photon. Before that, we have infinite potentialities and the energy is probabilistically distributed to all accessible points (r,θ,η), so that its integral over (θ,η) at any r be exactly e. A deterministic distribution is impossible, for the explained reason that the photon cannot be subdivided.
5. The difference with the random experiment is that the photon has also frequency (reflected in its colour) and phase. Hence, we should also consider these in our study. To do this we apply the same rules as in other waves discussed above.
6. Light has a constant speed c, independent of the reference frame and, for any observer, there is a lapse time between its emission and absorption. However, for

the photon per se, according to special relativity, the lapse time is zero. Hence, according to its own frame, a single photon is emitted, propagated and absorbed instantaneously, and at its instantaneous propagation fills the entire space as a potentiality. Considering the open space admissible to it, a photon is an omnipresent and all-pervading entity. This is the only mystery that remains—but it is more related to special relativity than to quantum mechanics.

We will explain and implement these ideas in section 5.4 by studying the celebrated double-slit experiment.

5.4 The double-slit experiment

In our study of the double-slit experiment on a stochastic basis we try to represent the experimental results, as seen in Figure 5.1. Our aim is to make a simple conceptual model of the behaviours observed and not an accurate mathematical model. In brief, these behaviours, as seen in the experimental findings, are:

(a) when there are two options for a photon (or another particle, e.g. an electron) to pass, materialized by the double-slit device seen in Figure 5.7, then an intermittent pattern, known as interference fringe, appears on the screen collecting the photons, which reflects a multi-modal distribution function;
(b) when there is just one option then a unimodal distribution emerges on the screen.

The latter case emerges when one slit is closed, as well as when a measurement is made determining the slit from which the photon passed. As already noted, a photon cannot be subdivided and, thus, if it is observed it will necessarily be found in only one of the two slits. However, if it is not observed, it behaves like a stochastic variable with two options, each of which has a certain probability.

Figure 5.7 provides a sketch of an idealized double-slit experiment and explains the basic notation used. We denote $f_{\boldsymbol{v}'}(\boldsymbol{v}') \equiv f_{x'y'}(x', y')$ the probability density that a photon passes from point A with coordinates $\boldsymbol{v}' = (x', y')$ and $f_{\boldsymbol{v}}(\boldsymbol{v}) \equiv f_{xy}(x, y)$ the probability density that the photon hits the screen at a point with coordinates $\boldsymbol{v} = (x, y)$. For convenience, we will study the photons hitting the screen at the x axis (i.e., assuming $y = 0$ like in point B in the figure), and we denote the corresponding density $f_x(x)$. Actually, this is a conditional density, i.e., $f_x(x) = f(x|y = 0) = f_{xy}(x, 0)/f_y(0)$. Finally, we denote $f_{x|\boldsymbol{v}'}(x|\boldsymbol{v}')$ the conditional probability density that the photon reaches point B (with coordinates $\boldsymbol{v} = (x, 0)$), once it is known that it passed from point A (with coordinates $\boldsymbol{v}' = (x', y')$). We emphasize that all these densities are abstract quantities and the conditions like the latter one are not observable, because we cannot observe the exact position $\boldsymbol{v}' = (x', y')$ where the photon passed from the slit.

Because light propagates in the form of spherical waves, as already discussed, it is reasonable to assume that the conditional probability density is

$$f_{x|\boldsymbol{v}'}(x|\boldsymbol{v}') = \frac{a^2}{l^2} = \frac{a^2}{L^2 + (x - x')^2 + y'^2} \tag{5.45}$$

where and l is the Euclidean distance between A and B (see figure) and a is a proportionality constant determined so that

$$\int_{-\infty}^{\infty} f_{x|\boldsymbol{v}'}(x|\boldsymbol{v}')\mathrm{d}x = 1 \tag{5.46}$$

It is readily recognized that $f_{x|\boldsymbol{v}'}(x|\boldsymbol{v}')$ denotes a Cauchy distribution and it is easily found that

$$a^2 = \frac{\sqrt{L^2 + y'^2}}{\pi} \approx \frac{L}{\pi} \tag{5.47}$$

where the approximation is justified if $y' \ll L$.

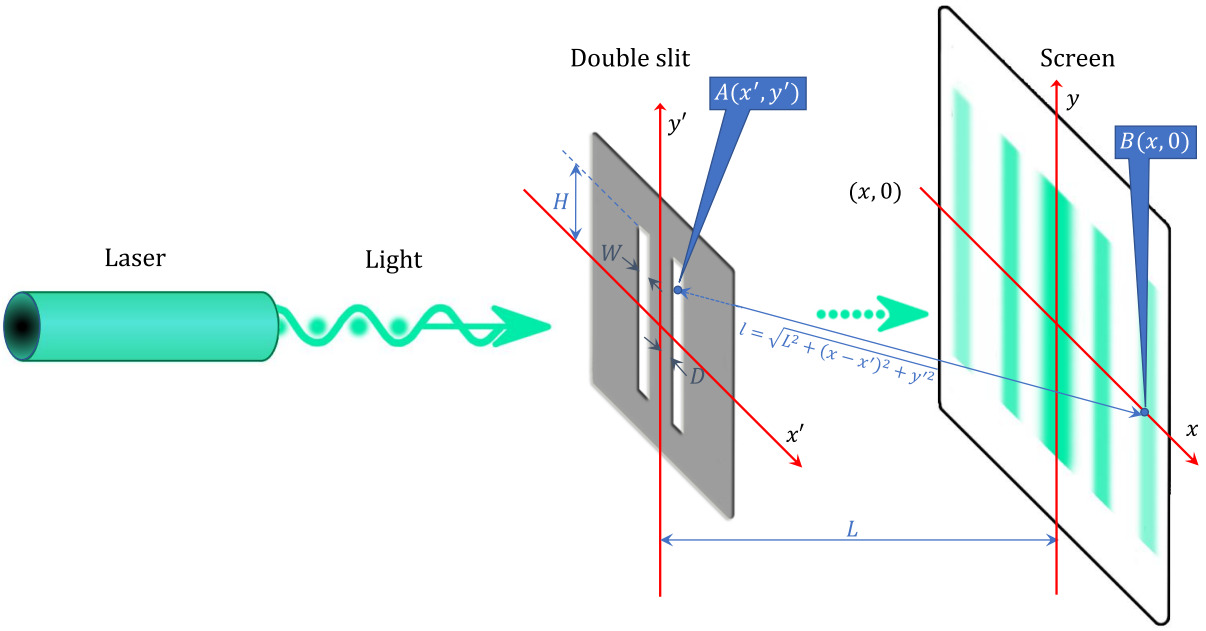

Figure 5.7 Sketch of the setup of an idealized double-slit experiment and notation used.

On the other hand, a photon is characterized by its wavelength λ, which we understand as colour. For example, a photon with the green colour shown in the figure has a wavelength of about 0.5 µm = 5×10^{-7} m. According to equation (5.42), the phase φ propagates linearly in time and space. When the phase at A is φ_0 (assuming this is the same for any point of the slit device), then the phase at B is

$$\varphi_{x|\boldsymbol{v}'} = \varphi_0 + \varphi, \qquad \varphi = \frac{2\pi l}{\lambda}, \qquad l = \sqrt{L^2 + (x - x')^2 + y'^2} \tag{5.48}$$

The phase φ is not reflected in the probability density f but it is in the complex number representation ψ. The three quantities are related by:

$$\psi = \sqrt{f}\mathrm{e}^{\mathrm{i}\varphi}, \qquad f = |\psi|^2 = \psi\overline{\psi} \geq 0, \qquad \varphi = \arg\psi \tag{5.49}$$

where $|\psi|$ is the modulus (absolute value) of ψ, $\overline{\psi}$ is the conjugate of ψ and arg is the function giving the argument (phase angle) of ψ.

Even if the point A with coordinates $\boldsymbol{v}' = (x', y')$ is the only option for the photon to pass, with probability 1, yet all points (x, y) on the screen are potential targets of the photon, and therefore we have again a continuous probability density $f_{\boldsymbol{v}}(\boldsymbol{v})$. If we are interested only on the points that hit the screen on the axis $y = 0$, then

$$\psi_x(x) = a\frac{1}{l}\exp\big(\mathrm{i}(\varphi_0 + \varphi)\big) \tag{5.50}$$

Because in this case there is just one option, the unconditional probability density equals the conditional one, i.e.,

$$f_x(x) = |\psi_x|^2 = \frac{a^2}{l^2} = \frac{a^2}{L^2 + (x - x')^2 + y'^2} \tag{5.51}$$

Now let us make the problem more interesting assuming two options for the photon to pass, namely points 1 and 2 with coordinates $\boldsymbol{v}_1' = (x_1', y_1')$ and $\boldsymbol{v}_2' = (x_2', y_2')$, and probabilities P_1 and $P_2 = 1 - P_1$, respectively. In this case we have

$$\psi_{x|1}(x) = \frac{a}{l_1}\exp\big(\mathrm{i}(\varphi_0 + \varphi_1)\big), \qquad \psi_{x|2}(x) = \frac{a}{l_2}\exp\big(\mathrm{i}(\varphi_0 + \varphi_2)\big) \tag{5.52}$$

The conditional probability density on the screen for a photon passing through point 1 or point 2 will be, respectively,

$$f_{x|1}(x) = \big|\psi_{x|1}(x)\big|^2 = \frac{a^2}{l_1^2}, \qquad f_{x|2}(x) = \big|\psi_{x|2}(x)\big|^2 = \frac{a^2}{l_2^2} \tag{5.53}$$

According to classical probability theory, the unconditional probability density should be:

$$f_x(x) = f_{x|1}(x)P_1 + f_{x|2}(x)P_2 = a^2\left(\frac{P_1}{l_1^2} + \frac{P_2}{l_2^2}\right) \tag{5.54}$$

However, this classical result does not agree with physical reality if the distance of the two points 1 and 2 is of the order of magnitude of the wavelength λ. So, there must be something wrong with this derivation. Tracing what is wrong, we may conclude that this derivation did not consider at all the phase, which certainly plays a role. To remedy this, we should take a different approach, considering additional variables (the phase angles) in the conditional distributions (e.g., Papoulis, 1991, p. 193). Indeed, if there are variables additional to x that affect the phenomenon, then that equation (5.54) does not hold.

We may observe that equation (5.54) calculates the marginal distribution as an expected value (cf. equation (2.134)), i.e.:

$$f_x(x) = \mathrm{E}\big[f_{x|\underline{w}}(x)\big] \tag{5.55}$$

where the stochastic variable $\underline{w}$ takes the values 1 and 2 with probabilities P_1 and P_2, respectively. To involve also the phases in the calculation, we may write in the analogous manner for the complex representation of the photon, which in addition considers the phase angles:

$$\psi_x(x) = \mathrm{E}\left[\psi_{x|\underline{w}}(x)\right] = \psi_{x|1}(x)P_1 + \psi_{x|2}(x)P_2 \tag{5.56}$$

This corresponds to the case where we do not have information about which of the points 1 or 2 the photon (as an indivisible quantum), actually passed through. If we have such information, say that it passed through point 1, then in equation (5.56) we simply set $P_1 = 1, P_2 = 0$. Otherwise, the form of equation (5.56) is exactly as if we assumed that the photon, still a potentiality, passed "at once" through both points. Specifically, a part P_1 of the energy of the photon, as a potentiality, passed through point 1 and the remaining part P_2 passed through point 2. According to this consideration, when we do not have information that a random event with two outcomes was *realized*, we treat both outcomes as if they existed *simultaneously* with *weights* equal to the respective probabilities.

Equation (5.56) yields

$$\psi_x(x) = a\left(\frac{P_1}{l_1}\exp\left(\mathrm{i}(\varphi_0 + \varphi_1)\right) + \frac{P_2}{l_2}\exp\left(\mathrm{i}(\varphi_0 + \varphi_2)\right)\right) \tag{5.57}$$

The unconditional probability density, under these considerations will be (omitting the coordinates in the notation for simplicity):

$$\begin{aligned} f_x(x) = |\psi_x|^2 = \psi_x\underline{\overline{\psi}}_x &= \left(\psi_{x|1}P_1 + \psi_{x|2}P_2\right)\left(\overline{\psi}_{x|1}P_1 + \overline{\psi}_{x|2}P_2\right) \\ &= P_1^2\psi_{x|1}\overline{\psi}_{x|1} + P_1^2\psi_{x|2}\overline{\psi}_{x|2} + P_1P_2\psi_{x|1}\overline{\psi}_{x|2} + P_1P_2\psi_{x|2}\overline{\psi}_{x|1} \end{aligned} \tag{5.58}$$

The first two terms denote $\left|P_1\psi_{x|1}\right|^2$ and $\left|P_2\psi_{x|2}\right|^2$, respectively. The last two terms yield

$$\begin{aligned} &P_1P_2\psi_{x|1}\overline{\psi}_{x|2} + P_1P_2\psi_{x|2}\overline{\psi}_{x|1} \\ &\quad= a^2\frac{P_1}{l_1}\exp\left(\mathrm{i}(\varphi_0 + \varphi_1)\right)\frac{P_2}{l_2}\exp\left(-\mathrm{i}(\varphi_0 + \varphi_2)\right) \\ &\quad+ a^2\frac{P_1}{l_1}\exp\left(-\mathrm{i}(\varphi_0 + \varphi_1)\right)\frac{P_2}{l_1}\exp\left(\mathrm{i}(\varphi_0 + \varphi_2)\right) = \\ &\quad= a^2\frac{P_1}{l_1}\frac{P_2}{l_2}\left(\exp\left(\mathrm{i}(\varphi_1 - \varphi_2)\right) + \exp\left(-\mathrm{i}(\varphi_1 - \varphi_2)\right)\right) \end{aligned} \tag{5.59}$$

This simplifies to

$$P_1P_2\psi_{x|1}\overline{\psi}_{x|2} + P_1P_2\psi_{x|2}\overline{\psi}_{x|1} = 2a^2\frac{P_1}{l_1}\frac{P_2}{l_2}\cos(\varphi_1 - \varphi_2) \tag{5.60}$$

Hence,

$$f_x(x) = |\psi_x|^2 = \left(a\frac{P_1}{l_1}\right)^2 + \left(a\frac{P_2}{l_2}\right)^2 + 2a^2\frac{P_1}{l_1}\frac{P_2}{l_2}\cos(\varphi_1 - \varphi_2) \tag{5.61}$$

and finally

$$f_x(x) = |\psi_x|^2 = a^2\left(\left(\frac{P_1}{l_1}\right)^2 + \left(\frac{P_2}{l_2}\right)^2 + 2\frac{P_1}{l_1}\frac{P_2}{l_2}\cos\left(\frac{2\pi(l_1 - l_2)}{\lambda}\right)\right) \tag{5.62}$$

We notice that this result does not depend on the phase at the slit device, φ_0, but only on the phase difference $\varphi_1 - \varphi_2$ and eventually on the travel time difference $l_1 - l_2$. Characteristic maximum and minimum values emerge when the cosine is 1 and –1, respectively. These are

$$f_x(x)_+ = \left(\frac{P_1}{l_1} + \frac{P_2}{l_2}\right)^2, \qquad f_x(x)_- = \left(\frac{P_1}{l_1} - \frac{P_2}{l_2}\right)^2 \tag{5.63}$$

It can easily be verified that these occur when $l_1 - l_2 = k\lambda$ and $l_1 - l_2 = (k + 1/2)\lambda$, respectively, where k is any integer. Setting $x_1' = \pm(D + W/2), y_1' = 0$ (the centres of the two slits), the distance difference is $l_1 - l_2 = \sqrt{L^2 + (x - D - W/2)^2} - \sqrt{L^2 + (x + D + W/2)^2}$. Peaks occur when the cosine in equation (5.62) becomes 1, which happens when $l_1 - l_2 = k\lambda$, where k is an integer. Solving for x and neglecting $k\lambda$ (assuming small k) over the much larger L, we find

$$x = \frac{1}{2}k\lambda\sqrt{1 + \left(\frac{L}{D + W/2}\right)^2} \approx \frac{k\lambda L}{2D + W} \tag{5.64}$$

The distance between two consecutive peaks is found by setting $k = 1$ in this equation.

In the case that $P_1 = 1, P_2 = 0$, it is readily seen that from equation (5.62) we recover equation (5.51) of the single option case. If $P_1 = P_2 = 1/2$, then

$$f_x(x) = |\psi_x|^2 = \frac{a^2}{4}\left(\left(\frac{1}{l_1}\right)^2 + \left(\frac{1}{l_2}\right)^2 + \frac{2}{l_1 l_2}\cos\left(\frac{2\pi(l_1 - l_2)}{\lambda}\right)\right) \tag{5.65}$$

We refer to equation (5.65) as the *simplified* representation of the double-slit experiment.

Now returning to the real case, in which the options for a photon to pass through the slit device are uncountably infinite, namely the entire area S which is open for the light to pass, similar to (5.52), we write

$$\psi_{x|\boldsymbol{v}'}(x) = \frac{a}{\sqrt{L^2 + (x - x')^2 + y'^2}}\exp\left(\mathrm{i}\left(\varphi_0 + \frac{2\pi}{\lambda}\sqrt{L^2 + (x - x')^2 + y'^2}\right)\right) \tag{5.66}$$

with

$$f_{x|\boldsymbol{v}'}(x) = \left|\psi_{x|\boldsymbol{v}'}(x)\right|^2 \tag{5.67}$$

Generalizing equation (5.62), we will have for the unconditional case

$$\psi_x(x) = \mathrm{E}\left[\psi_{x|\boldsymbol{v}'}(x)\right] = \int_S \psi_{x|\boldsymbol{v}'}(x) f_{\boldsymbol{v}'}(\boldsymbol{v}')\mathrm{d}\boldsymbol{v}' \tag{5.68}$$

or

$$\psi_x(x) = a \iint\limits_{(x',y')\in S} \frac{\exp\left(\mathrm{i}\frac{2\pi}{\lambda}\sqrt{L^2 + (x - x')^2 + y'^2}\right)}{\sqrt{L^2 + (x - x')^2 + y'^2}} f_{x'y'}(x', y')\mathrm{d}x'\mathrm{d}y' \tag{5.69}$$

where for simplicity we have set $\varphi_0 = 0$ because, as we have seen, the initial phase is eliminated in the calculation of $|\psi_x|$. Hence

$$f_x(x) = A^2 \left| \iint\limits_{(x',y')\in S} \frac{\exp\left(\mathrm{i}\frac{2\pi}{\lambda}\sqrt{L^2+(x-x')^2+y'^2}\right)}{\sqrt{L^2+(x-x')^2+y'^2}} f_{x'y'}(x',y')\,\mathrm{d}x'\mathrm{d}y' \right|^2 \quad (5.70)$$

If $f_{x'y'}(x',y')$ is uniform, we have

$$\psi_x(x) = A \iint\limits_{(x',y')\in S} \frac{\exp\left(\mathrm{i}\frac{2\pi}{\lambda}\sqrt{L^2+(x-x')^2+y'^2}\right)}{\sqrt{L^2+(x-x')^2+y'^2}}\mathrm{d}x'\mathrm{d}y' \quad (5.71)$$

where $A = a/A_S$ and A_S denotes the area of S. Hence, the final equation in our detailed framework is:

$$f_x(x) = A^2 \left| \iint\limits_{(x',y')\in S} \frac{\exp\left(\mathrm{i}\frac{2\pi}{\lambda}\sqrt{L^2+(x-x')^2+y'^2}\right)}{\sqrt{L^2+(x-x')^2+y'^2}}\mathrm{d}x'\mathrm{d}y' \right|^2 \quad (5.72)$$

Putting the above considerations in a pure stochastic framework we may summarize as follows, under the condition that the photon reached the screen at point $(x, y = 0)$.

1. The point at which the photon passed through the slit device is a stochastic variable $\underline{\boldsymbol{v}}' = (\underline{x}', \underline{y}')$ with probability density $f_{\boldsymbol{v}'}(\boldsymbol{v}') \equiv f_{x'y'}(x',y')$.
2. The position where the photon hits the screen is also a stochastic variable, $\underline{x}$ with probability density $f_x(x)$.
3. The two densities are generally related by equation (5.70).
4. To calculate $f_x(x)$ it suffices to specify $f_{x'y'}(x',y')$, which is based on the knowledge that we have about where the light passed through the slit device.
5. In the absence on specific knowledge, except for the area S available for a photon to pass, we may assume a uniform $f_{x'y'}(x',y')$ over that area. In this case we use equation (5.72) instead of (5.70).

With respect to point 5, we may distinguish the following cases.

- If both slits are open and we know nothing about the position in which the photon has passed through the slit device, then we may assume uniform distribution over the area defined as $-H \le y' \le H$ and $-D - W \le x' \le -D, D \le x' \le D + W$ with $A_S = 4WH$.
- If we know that the photon has passed through one of the slits, say the rightmost one, then we may assume uniform distribution over the area defined as $-H \le y' \le H$ and $D \le x' \le D + W$ with $A_S = 2WH$.
- For simplification we may assume an exact coordinate $y' = 0$ of the photon at the slit device and nothing about the coordinate x', i.e., $\underline{\boldsymbol{v}}' = (\underline{x}', 0)$, and uniform distribution over the unknown coordinate, i.e. over the intervals $(-D - W,\ -D)$

and $(D, D+W)$, replacing the double integral with a single one and setting $A_S = 2W$ (the length of the intervals).

These are realistic cases, while the ones examined before (equations (5.51) and (5.65)) are purely hypothetical because it is impossible to measure the exact position of a photon. Note that in the case where the slit through which the photon has passed is known, the way in which this knowledge was acquired does not matter. It could result from the fact that in our experiment we have closed one slit or from a measurement, if both slits are open. Also, the above framework is the same whether the emitted photons are one at a time or more than one simultaneously.

In equations (5.70) and (5.72), the double integral is difficult to determine analytically, but numerical integration is easy (using mathematical software). The constant A appearing in the above equation can be easily calculated a posteriori, so the $\int_{-\infty}^{\infty} f_x(x)\mathrm{d}x = 1$.

Replacing the double integral with a single one (assuming $y' = 0$ as discussed above) speeds up the calculations without discernible cost in accuracy. In this case, equation (5.72) becomes:

$$f_x(x) = A^2 \left| \int_{x' \in S} \frac{\exp\left(\mathrm{i}\frac{2\pi}{\lambda}\sqrt{L^2 + (x-x')^2}\right)}{\sqrt{L^2 + (x-x')^2}} \mathrm{d}x' \right|^2 \tag{5.73}$$

Another fast method is to replace the integration with stochastic simulation—a choice fully consistent with our stochastic framework. In this we generate a (large) number N of points $\boldsymbol{v}' = (x', y')$ from the uniform distribution over the admissible area S. Then from equation (5.66) we calculate $\psi_{x|\boldsymbol{v}'}(x)$ which now equals the unconditional $\psi_x(x)$, because $\boldsymbol{v}'$ is known for each generated point $\boldsymbol{v}'$. Finally we take the average $\psi_x(x)$ over the N points and the modulus square of it, which is the required $f_x(x)$. This can be expressed mathematically as

$$f_x(x) = A^2 \left| \mathrm{E}\left[\psi_{x|\underline{\boldsymbol{v}}'}(x)\right] \right|^2 \tag{5.74}$$

where the random variable over which the expectation is taken is $\underline{\boldsymbol{v}}'$.

We illustrate the framework using several examples depicted in the figures that follow. In all examples we assumed monochromatic green light with a wavelength $\lambda = 0.5$ μm $= 5 \times 10^{-7}$ m. In most of the examples, the slit width W and the half distance between the slits D are chosen close to their lower technically feasible limits, i.e. of the order of a few micrometers, because in this case the results are easier to visualize. We also include examples with much larger sizes for comparison. The particular sizes for each of the examples are noted in the figure captions.

In Figure 5.8 we compare the four computational methods described above. It is reasonable to assume that the two-dimensional numerical integration (equation (5.72)), which is least affected by assumptions, is the most accurate. At the other end, the simplified representation (equation (5.65)) is far too inaccurate. It correctly captures the interference fringes, correctly locating them in time, but not their number and intensity.

It is reminded that this assumes that the photons passed through two sharp positions, the centres of the slits. If this sharp locating of these positions would be possible, the extent of the interference fringe produced would be much wider, as seen in the figure. However, interference has another effect, the focus of the light in a narrower area, which the simplified representation is unable to capture.

The other two methods, the one-dimensional numerical integration (equation (5.73)) and the stochastic simulation (equation (5.74)) give accurate results, practically indistinguishable from those of the two-dimensional numerical integration, as seen in Figure 5.8.

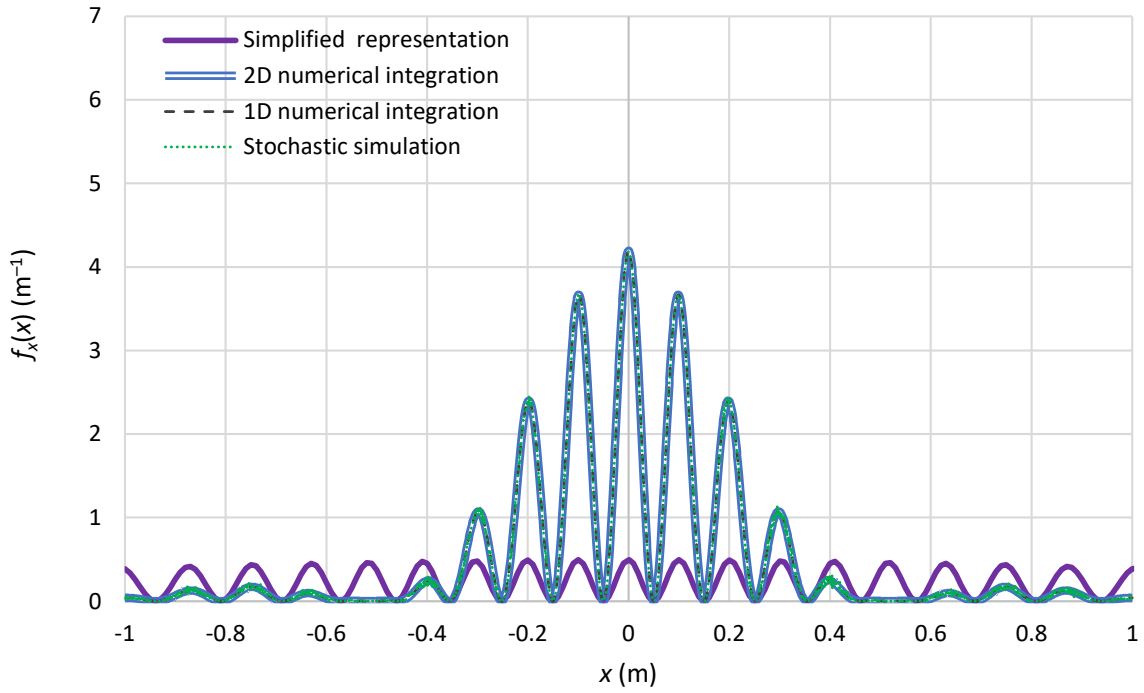

Figure 5.8 Comparison of different computational methods for the probability density $f_x(x)$ that a photon hits the screen at point $\boldsymbol{v} = (x, 0)$. The numerical constants are $\lambda = 0.5$ μm $= 5 \times 10^{-7}$ m (wavelength; green light), $W = 4\lambda = 2$ μm $= 2 \times 10^{-6}$ m (slit width), $D = 2W = 4$ μm $= 4 \times 10^{-6}$ m (half distance between slits), $H = 0.1$ mm $= 0.0001$ m (slit half height) and $L = 2$ m (distance between slit device and screen). The simplified representation is based on equation (5.65), the 2D numerical integration on equation (5.72) and the 1D numerical integration on equation (5.73). The stochastic simulation (equation (5.74)) used $N = 2000$ photons, equally distributed to each of the slits and uniformly distributed across the area of the slits. Aside from the simplified representation, the curves of the other methods are virtually identical.

Figure 5.9 makes a comparison of the case where both slits are open vs. the case where only one is open. The former case is the same as shown in Figure 5.8 with a multimodal distribution, while in the latter case, the distribution is clearly unimodal. Notice that, because of the chosen very small slit width and half-distance between slits, the peak of the distribution appears almost at the axis origin and there is no practical difference whichever of the two slits is open. The simplified cases of fixed coordinates of the photon at the slit are also shown in the figure which are similarly multimodal if the two slits are open (equation (5.65)) or one is closed (equation (5.51); Cauchy distribution). Both simplified cases magnify diffraction (high extent of the lit area) and fail to represent the interference's effect to focus the light in a narrow area.

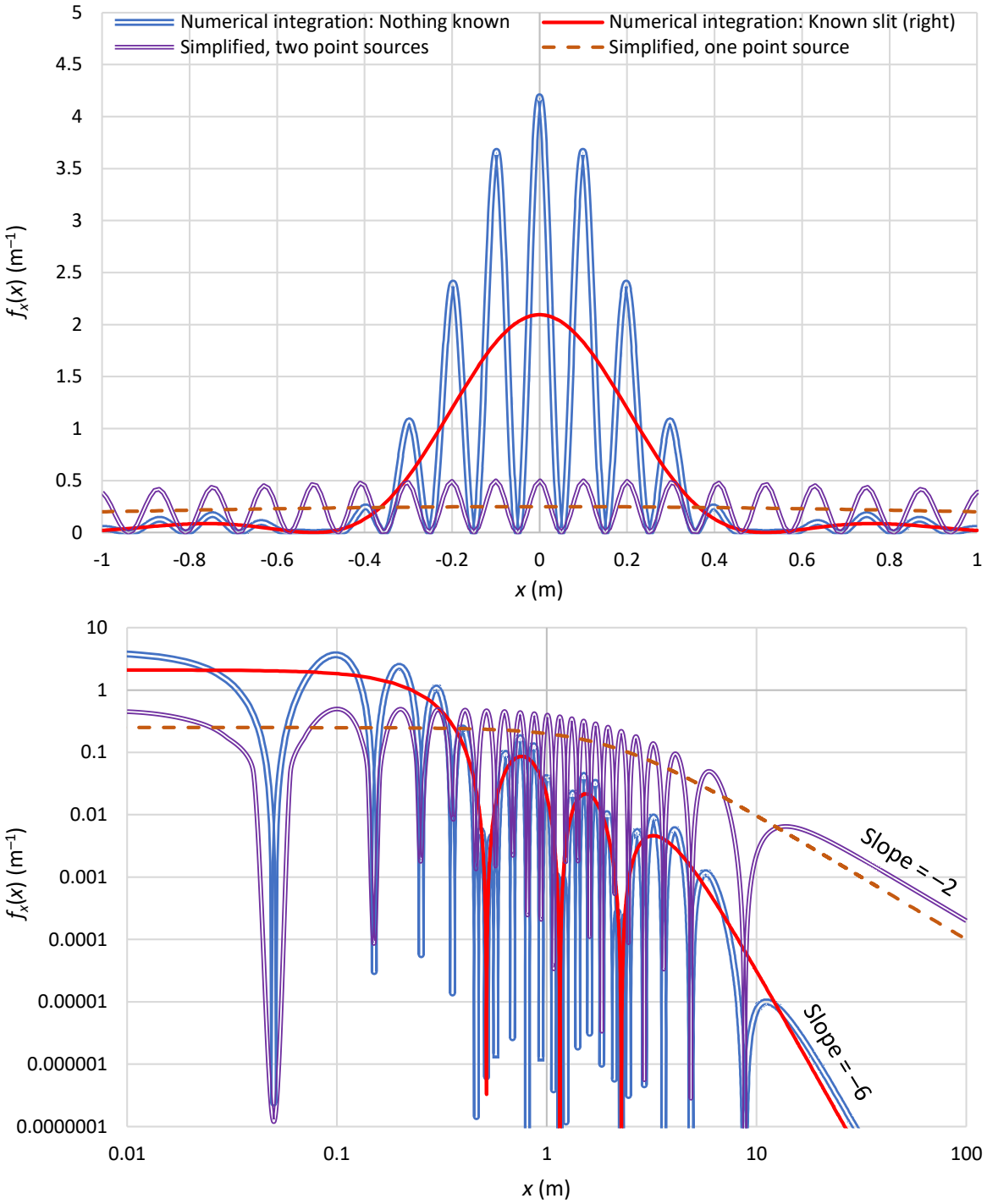

Figure 5.9 Numerical example for the probability density $f_x(x)$ (with x in m) that a photon hits the screen at point $v = (x, 0)$ for the indicated cases (see also text). The numerical constants are the same as in Figure 5.8. (**upper**) Cartesian plot; (**lower**) double logarithmic plot for positive x.

It is interesting to examine the distribution tails, which in all cases are of power type. These are shown in the lower panel of Figure 5.9. In the hypothetical case where the emission is from one or two single points, the slope is –2, as implied by equation (5.51) (Cauchy distribution). However, in the realistic cases where there is uncertainty (different options within an interval) along the x' direction, the slope is steeper than –2, namely –6. This again means larger concentration of the photons at the part of the screen that is

nearer to the slits and fewer photons at more distant points. Interestingly, this happens irrespectively of whether one or both slits are open.

It is important to note that these results emerge when both the slit width W and the half-distance between slits D are of the order of magnitude of the wavelength, i.e. extraordinary small. If we increase D to a typical macroscopical value, keeping W small, then the multimodality becomes so exceptionally high that it practically cancels itself. This is illustrated in the upper panel of Figure 5.10 for $D = 1$ cm. From a practical point of view, it is impossible to observe the small-scale fluctuations and what we observe if both slits are open is a locally averaged $f_x(x)$. This will be the same with the smooth curve when one slit is open.

If both W and D take typical macroscopical values (e.g., of the order of a cm, as specified in the example of the lower panel of Figure 5.10), then the pattern becomes quite different, and closer to our everyday macroscopic experience. As shown in the figure, the result would practically be two light beams behind the slits or a single beam if one slit is closed.

In other words, if the dimensions in the slit experiment are of macroscopic order of magnitude (compared to the wavelength) then we could build a macroscopic model that would not take into account the detailed representation of wave phases. Yet that would not be as simple as equation (5.54), because this would fail to represent the focused light beams seen in the lower panel of Figure 5.10. This illustrates that consistent macroscopization, in which all constituent uncertainties at a microscopic level are taken into account, results in a logical macroscopic picture, consistent with our experience from our senses. Therefore, there might be no meaning in isolating one microscopic element and treating it separately with equations pertinent to the microscopic behaviour, while macroscopizing all other constituents.

This is the case in the famous thought experiment of Schrödinger's cat, in which a random subatomic event (decay in a radioactive atom) is treated with full quantum detail, while the affected cat is treated as purely macroscopic. The paradox may result just from this inconsistency. It may be easy to make a Schrödinger's equation for a random subatomic event but impossible to make an "equation" of the same precision for the cat. Lacking such "equation" it is pointless to speculate about the cat's states or superposition thereof.

Another absurdity in the imagination of Schrödinger's cat, which is presumably alive and dead simultaneously (in a superposition) until observed, is the idea that it is the observation and perhaps the consciousness of the observer that makes the *realization* of a potentiality happen. This has triggered profound discussions on several dilemmas (e.g. "Is the moon there when nobody looks", Mermin, 1985). However, the slit experiment helps us understand that it is not the observer that causes the realization of the light as photon on the screen. It is the screen. Were it absent and the space behind the slit device open, the light would travel as potentiality until it perhaps reached another galaxy somewhere far away. In contrast, when the screen is there with two slits open, even if nobody observes the experiment, the fringes will be formed. If an observer visits it days

after the experiment, he will see them. If the experiment is done repeatedly several times, the result will be the same.

This treatment of the quantum world and the pragmatic ontological premises it rests on also serve as the foundation for the thermodynamic derivations that follow in Chapter 6.

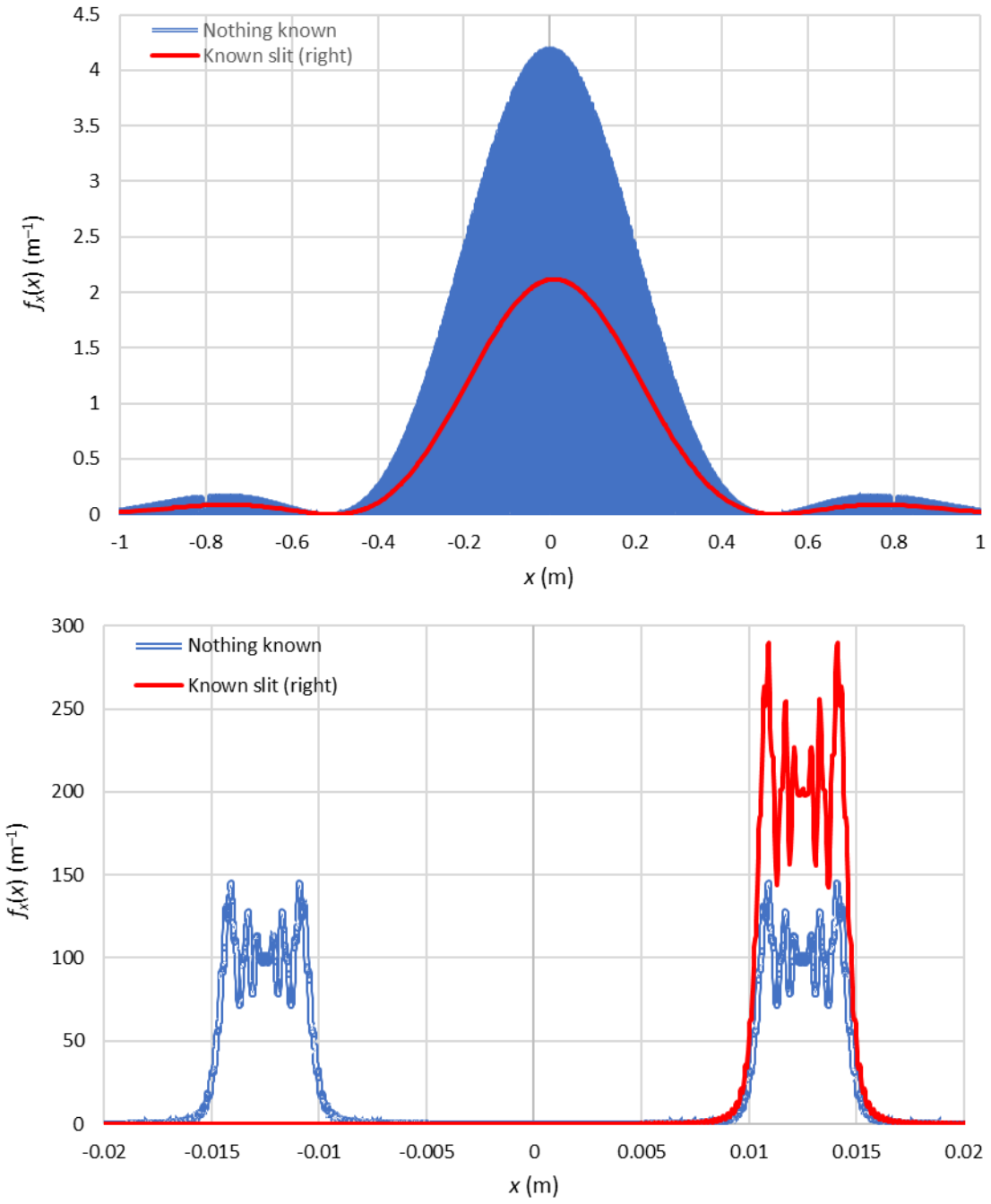

Figure 5.10 Numerical example for the probability density $f_x(x)$ (with x in m) that a photon hits the screen at point $v = (x, 0)$ for the indicated two cases (see also text). The numerical constants are $\lambda = 5 \times 10^{-7}$ m (wavelength; green light), $W = 2$ µm $= 2 \times 10^{-6}$ m for the upper panel and 5 mm $= 0.005$ m for the lower panel (slit width), $H = 5$ mm $= 0.005$ m (slit half height), $D = 10$ mm $= 0.01$ m (half distance between slits) and $L = 2$ m (distance between slit device and screen). The seemingly solid area in the upper graph is in fact a curve filling the entire surface via oscillations with a period of 50 µm.

Chapter 6. Atmospheric thermodynamics deduced by stochastics

6.1 Premises

In the modern period and up to the 19th century, heat was interpreted as the flow of the caloric fluid, assumed to be a weightless substance flowing from hotter to colder bodies, passing through pores of matter. In the mid-19th century this view was replaced by a mechanical theory of heat. This is marked by the introduction of the entropy concept in mechanical terms, and the development of thermodynamic principles (Zeroth, First, Second and Third Law; see section 6.26). As already mentioned in section 1.1, near the end of the 19th century, Boltzmann explained entropy in statistical terms. Despite the subsequent development of statistical thermophysics, still the dominant perception of thermodynamics remains mechanistic and deterministic, and in some respects still caloric.

In this chapter we highlight the fact that all thermodynamic laws in gases, with focus on the atmosphere, are stochastic laws. We deduce these laws using stochastics—and in particular the principle of maximum entropy—and without using the thermodynamic principles at all. We show that thermodynamic laws are stochastic laws—typically relationships of expectations of stochastic variables. We derive these laws by maximizing entropy, i.e. uncertainty, at a microscopic (molecular) level. Interestingly, though, they turn out to macroscopically express near certainties and are commonly misinterpreted as deterministic laws. The explanation of near certainty relies on these two facts:

- Typical thermodynamic systems are composed of hugely many identical elements: $N_\mathrm{a} = 6.022 \times 10^{23}$ per mole of material (N_a is the Avogadro's number).
- The random motion of each of the system elements makes their state practically independent of the others'.

As a consequence, a stochastic variable $\underline{x}$ expressing a macroscopic state of 1 mole of material, will have a coefficient of variation $\mathrm{std}[\underline{x}]/E[\underline{x}] \approx 1/\sqrt{N_\mathrm{a}} = 1.3 \times 10^{-12}$ (see also section 1.4). The fact that the macroscopic variability is practically zero should not mislead us to interpret the laws in deterministic terms.

We stress that here we do not assume the thermodynamic laws as first principles, nor do we dispute them. Rather, we derive them from conservation laws plus the principle of maximum entropy. These are our first principles.

6.2 The uncertain motion of a single monatomic molecule

We consider a motionless cube with edge a (volume $V = a^3$) containing spherical particles, namely monatomic molecules of mass m_0, moving rapidly*. Their exact position and velocity are unknown: actually, as illustrated in section 1.4, it is infeasible to know

* See 2D animations at https://commons.wikimedia.org/wiki/File:Translational_motion.gif and https://www.itia.ntua.gr/en/getfile/2537/50/documents/UnsettlingSupplementaryMaterialVideoLowRes.mp4. The latter is a simulation in a gravitational field.

them. Here we focus on one of these particles. Its state is described by 6 variables, 3 indicating its position $\underline{x}_i$ and 3 indicating its velocity $\underline{u}_i$ with $i = 1,2,3$, all represented as stochastic variables, forming the vector $\underline{\boldsymbol{z}} = (\underline{x}_1, \underline{x}_2, \underline{x}_3, \underline{u}_1, \underline{u}_2, \underline{u}_3)$. The constraints for the particle's position are:

$$0 \leq \underline{x}_i \leq a \tag{6.1}$$

We use a non-relativistic framework and therefore we do not constrain velocity. The feasible space, Ω, is thus $(0, a)$ for each x_i and $(-\infty, \infty)$ for each u_i ($\Omega := \{0 \leq x_i \leq a, -\infty < u_i < \infty; i = 1,2,3\}$)

The container (cube) as a whole is at rest, and therefore conservation of momentum demands that $\mathrm{E}[m_0 \underline{u}_i] = m_0 \int_\Omega u_i f(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} = 0$, or:

$$\mathrm{E}[\underline{u}_i] = 0, \qquad i = 1,2,3 \tag{6.2}$$

We note that the expectation $\mathrm{E}[\underline{u}_i]$ represents a macroscopic motion, while $u_i - \mathrm{E}[\underline{u}_i]$ represents fluctuation at a microscopic level.

The conservation of energy demands that : $\mathrm{E}\left[m_0 \|\underline{u}\|^2/2\right] = (m_0/2) \int_\Omega \|\underline{u}\|^2 f(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \varepsilon$ where ε is the energy per particle and $\|\underline{u}\|^2 = u_1^2 + u_2^2 + u_3^2$. Hence, the energy constraint is

$$\mathrm{E}\left[\|\underline{u}\|^2\right] = \frac{2\varepsilon}{m_0} \tag{6.3}$$

We note that if $\mathrm{E}[\underline{u}_i] \neq 0$, then the macroscopic and microscopic kinetic energies should be treated separately, the latter being $\varepsilon = \mathrm{E}\left[m_0 \|\underline{u} - \mathrm{E}[\underline{u}]\|^2/2\right]$—see also Digression 6.A. This is known as *internal* (or *thermal*) *energy* of the system.

Now we form the entropy of $\underline{\boldsymbol{z}}$ as in equation (2.73), recognizing that the background density $\beta(\boldsymbol{z})$ in $\ln(f(\boldsymbol{z})/\beta(\boldsymbol{z}))$ should have units $[\boldsymbol{z}^{-1}] = [x^{-3}]\,[u^{-3}] = [\mathrm{L}^{-6}\,\mathrm{T}^3]$. To form this, we utilize a universal constant, i.e., the Planck constant $h = 6.626 \times 10^{-34}$ J·s, whose dimensions are $[\mathrm{L}^2\,\mathrm{M}\,\mathrm{T}^{-1}]$. If we combine it with the particle mass m_0, we observe that the quantity $(m_0/h)^3$ has the required dimensions $[\mathrm{L}^{-6}\,\mathrm{T}^3]$, thereby giving the entropy as

$$\Phi[\underline{\boldsymbol{z}}] := \mathrm{E}\left[-\ln\left(\left(\frac{h}{m_0}\right)^3 f(\underline{\boldsymbol{z}})\right)\right] = -\int_\Omega \ln\left(\left(\frac{h}{m_0}\right)^3 f(\boldsymbol{z})\right) f(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \tag{6.4}$$

Application of the principle of maximum entropy with constraints (2.76), (6.1), (6.2) and (6.3) will give the distribution of $\underline{\boldsymbol{z}}$ as:

$$f(\boldsymbol{z}) = \left(\frac{1}{a}\right)^3 \left(\frac{3m_0}{4\pi\varepsilon}\right)^{3/2} \exp\left(-\frac{3m_0}{4\varepsilon} \|\underline{u}\|^2\right), \qquad 0 \leq x_i \leq a \tag{6.5}$$

To understand this, we recall that the entropy maximizing distribution is given by equation (2.83) and hence its density will be an exponential function of a second order polynomial of (u_1, u_2, u_3) involving no products of different u_i, i.e., $f(z) = A\,(m_0/h)^3 \exp\left(-\lambda_1 u_1 - \lambda_2 u_2 - \lambda_3 u_3 - \lambda_4 (u_1^2 + u_2^2 + u_3^2)\right)$. Indeed, $f(\boldsymbol{z})$ in (6.5) is of this

type, and thus it suffices to show that it satisfies the constraints. The inequality constraint (6.1) is not considered at this phase but only in the integration to evaluate the constraints. Denoting $\int_\Omega a(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}$ the integral of $a(\boldsymbol{z})$ over the domain Ω, it is trivial to show that:

$$\int_\Omega f(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = 1, \qquad \int_\Omega u_i f(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = 0, \qquad \int_\Omega (u_1^2 + u_2^2 + u_3^2)\, f(\boldsymbol{z})\mathrm{d}\boldsymbol{z} = \frac{2\varepsilon}{m_0} \tag{6.6}$$

Thus, all constraints are satisfied.

To find the marginal distribution of each of the variables we integrate over the domain of the remaining variables. Thus, the marginal distribution of each of the location coordinates $\underline{x_i}$ is easily found to be uniform in $[0, a]$, i.e.,

$$f(x_i) = \frac{1}{a}, \qquad 0 \le x_i \le a, \qquad i = 1,2,3 \tag{6.7}$$

The marginal distribution of each of the velocity coordinates $\underline{u_i}$ is derived from equation (6.5) as

$$f(u_i) = \left(\frac{3m_0}{4\pi\varepsilon}\right)^{3/2} \exp\left(-\frac{3m_0}{4\varepsilon} u_i^2\right) \tag{6.8}$$

This is Gaussian with mean 0 and variance $2\varepsilon/3m_0$. Given that in a monatomic molecule we have three degrees of freedom (one per direction of motion), the variance is twice the energy per unit mass per degree of freedom.

Since $\left\|\underline{u}\right\|^2 = u_1^2 + u_2^2 + u_3^2$, from (6.5) we readily deduce that the joint distribution $f(\boldsymbol{z})$ is a product of functions of $\boldsymbol{z}$'s coordinates $(x_1, x_2, x_3, u_1, u_2, u_3)$. This means that all six stochastic variables are jointly independent. The independence results from entropy maximization. From (6.5) and (6.8) we also observe a symmetry with respect to the three velocity coordinates, resulting in uniform distribution of the energy ε into $\varepsilon/3$ for each direction or degree of freedom. This is known as the *equipartition* principle and is again a result of entropy maximization.

Thus, neither independence nor equipartition are posed as assumptions here. Clearly, they are derived by the principle of maximum entropy. In other words, the kinetic energy is equally distributed among the different degrees of freedom because this maximizes entropy, that is, uncertainty.

To find the marginal distribution of the velocity magnitude $\left\|\underline{u}\right\|$, we recall that the sum of squares of n independent $N(0,1)$ stochastic variables has a $\chi^2(n)$ distribution (Papoulis, 1990, p. 219, 221) and then we use known results for the density of a transformation of a stochastic variable (Papoulis, 1990, p. 118) to obtain the distribution of the square root. The result is:

$$f\left(\left\|\underline{u}\right\|\right) = \left(\frac{2}{\pi}\right)^{1/2} \left(\frac{3m_0}{2\varepsilon}\right)^{3/2} \left\|\underline{u}\right\|^2 \exp\left(-\frac{3m_0}{4\varepsilon}\left\|\underline{u}\right\|^2\right) \tag{6.9}$$

This is known as the Maxwell–Boltzmann distribution.

Once $f(\boldsymbol{z})$ has been determined in equation (6.5), the final expression of entropy is then obtained as follows. We observe that

$$-\ln f(\boldsymbol{z}) = 3\ln a + \frac{3}{2}\ln\left(\frac{4\pi\varepsilon}{3m_0}\right) - \frac{3m_0}{4\varepsilon}(u_1^2 + u_2^2 + u_3^2), \qquad \ln\beta(\boldsymbol{z}) = 3\ln\frac{m_0}{h} \tag{6.10}$$

After performing the integration over Ω we find

$$\Phi[\underline{\boldsymbol{z}}] = \frac{3}{2}\ln\left(\frac{4\pi\mathrm{e}}{3}\,\frac{m_0}{h^2}\,\varepsilon V^{2/3}\right) = \frac{3}{2}\ln\frac{\varepsilon}{\varepsilon^*} + \ln\frac{V}{v^*} \tag{6.11}$$

where e is the base of natural logarithms and ε^*, v^* are constants with units of energy and volume, respectively, satisfying

$$\frac{4\pi\mathrm{e}}{3}\,\frac{m_0}{h^2}\,\varepsilon^* {v^*}^{2/3} = 1 \tag{6.12}$$

so that the middle term of the equation equals the rightmost one. Note that one of ε^*, v^* is a freely chosen parameter. Equation (6.11) reflects the fact that the entropy $\Phi[\underline{\boldsymbol{z}}]$ is a dimensionless quantity, as it should according to its definition in section 2.3, and its rightmost part is the sum of two dimensionless quantities representing energy and space available to motion.

An extended version of equation (6.11) (for many particles; see section 6.5), but with some differences (see section 6.6), is known as the Sackur-Tetrode equation (after H. M. Tetrode and O. Sackur, who developed it independently at about the same time in 1912). An interesting question is: When does the entropy, based on this formulation, become zero? This is discussed in Digression 6.B.

Digression 6.A: Is the entropy subjective or objective?

In physics most quantities are subjective in the sense that they depend on the *observer's* choices. There are also some objective quantities that are unaltered if the observer's choices change. Thus, the location coordinates (x_1, x_2, x_3) depend on the observer's choice of the coordinate frame and change if this frame is translated or rotated; however, the distance between two points remains constant if the frame changes. Also, the velocity depends on the relative motion of the frame of reference; the velocity of a car whose speedometer indicates 100 km/h is zero for an observer moving with the car, 100 km/h for an observer sitting at the road and about 107 000 km/h for a coordinate system attached to the sun.

The kinetic energy, as well as changes thereof, depend on the reference frame, too. If we have a system of N bodies with masses m_i and velocities u_i according to a specified reference frame, then the total kinetic energy of the system is $\sum_{i=1}^{N} m_i \|\boldsymbol{u}_i\|^2/2$. If we change the reference frame, so that the velocities in it are $\boldsymbol{v}_i = \boldsymbol{u}_i - \boldsymbol{c}$, where $\boldsymbol{c}$ is a constant, then the kinetic energy in this frame will be different, $\sum_{i=1}^{N} m_i \|\boldsymbol{u}_i - \boldsymbol{c}\|^2/2$. This quantity has a minimum among all frames, which can be easily found by taking the derivative with respect to $\boldsymbol{c}$ and equating it to zero. The result is $\boldsymbol{c} = \sum_{i=1}^{N} m_i \boldsymbol{u}_i / \sum_{i=1}^{N} m_i$. It can be easily seen that this is the velocity of the centre of gravity of the system with respect to the initial reference frame. This minimum kinetic energy gives the most characteristic representation of energy and does not depend on the choice of the initial reference frame.

Likewise, both the internal energy ε and the entropy $\Phi[\underline{\boldsymbol{z}}]$ of the gas molecule in a container of a fixed volume V, as given in (6.11), do not change when the reference frame is different, provided that the kinetic energy per gas molecule ε is defined based on the difference of velocity $\underline{u}$ from its mean $E[\underline{u}]$, i.e., $\varepsilon = \mathrm{E}\left[m_0\left\|\underline{u} - \mathrm{E}[\underline{u}]\right\|^2/2\right]$. In this case ε is also invariant, despite that $\underline{u}$ changes with the reference frame. The invariance extends to the entropy maximizing distribution.

To illustrate the invariance properties of entropy, we consider again a container filled with a gas with spherical particles (3 degrees of freedom). We assume that an observer is moving with velocity c parallel to the horizontal axis x_1 as shown in Figure 6.1. According to this observer each particle has location coordinates x_i' and velocity coordinates u_i' related to those of the fixed frame, x_i and u_i, by the relationships shown in figure.

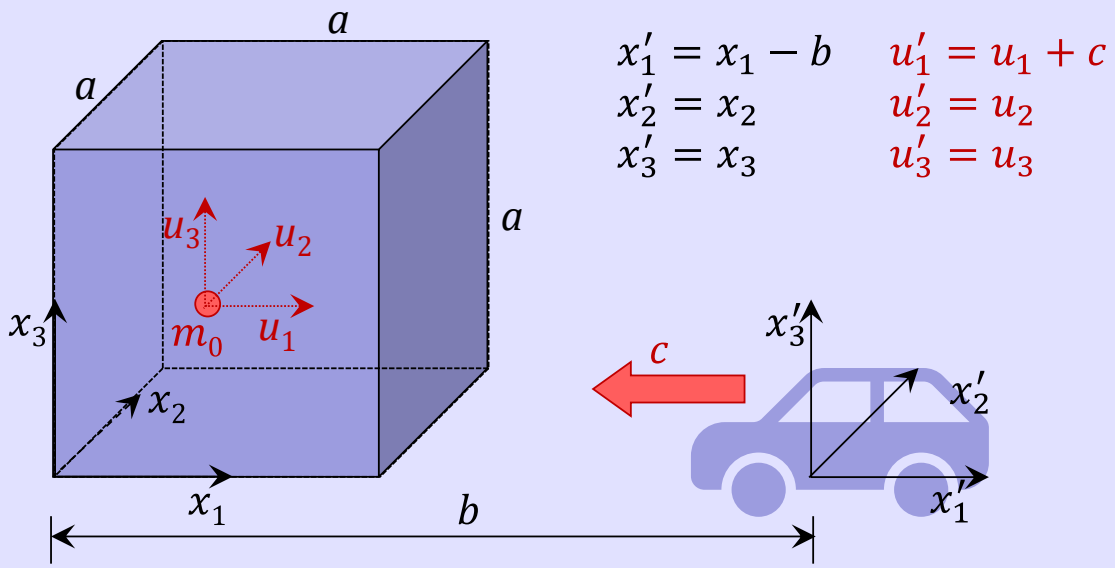

Figure 6.1 Sketch for the demonstration of invariance properties of entropy with respect to the coordinate frame.

For the frame attached to the car, $\mathrm{E}[u_1'] = c, \mathrm{E}[u_2'] = \mathrm{E}[u_3'] = 0$. Thus, $\mathrm{E}\left[m_0 \|\underline{u}' - \mathrm{E}[\underline{u}']\|^2/2\right] = \mathrm{E}\left[m_0\left(\left(\underline{u}_1' - c\right)^2 + \underline{u}_2'^2 + \underline{u}_3'^2\right)/2\right] = \varepsilon$, the same as for the fixed frame. For one molecule, the density function will be

$$f(\boldsymbol{z}') = \left(\frac{1}{a}\right)^3 \left(\frac{3m_0}{4\pi\varepsilon}\right)^{3/2} \exp\left(-\frac{3m_0}{4\varepsilon}\left(\left(\underline{u}_1' - c\right)^2 + \underline{u}_2'^2 + \underline{u}_3'^2\right)\right),$$

$$-b \le x_1' \le -b + a, 0 \le x_{2,3}' \le a$$

It can be verified, using the same method as in section 6.2, that it satisfies all constraints. To calculate the entropy, we observe that

$$-\ln\left(\left(\frac{h}{m_0}\right)^3 f(\boldsymbol{z}')\right) = -3\ln\frac{h}{m_0} - \frac{3}{2}\ln\left(\frac{3m_0}{4\pi\varepsilon}\right) + \ln a^3 + \frac{3m_0}{4\varepsilon}\left(\left(\underline{u}_1' - c\right)^2 + \underline{u}_2'^2 + \underline{u}_3'^2\right)$$

Thus, the entropy is calculated as

$$\Phi[\underline{\boldsymbol{z}}'] = \frac{3}{2}\ln\left(\frac{4\pi e}{3}\frac{m_0}{h^2}\varepsilon V^{2/3}\right)$$

which is the same as in (6.11).

Therefore, despite that entropy is based on probabilities, it is an objective quantity that can be determined from measurements, and its magnitude does not depend on the reference frame. We note though that the above expression of entropy is for the complete microscopic description of the phenomenon, consisting of all elementary events. One could define partitions of the elementary events into composite events—and in this case the entropy depends on the specific partition. Evidently, it also depends on the chosen background measure.

Digression 6.B: When is the entropy zero?

According to the Third Law of thermodynamics, the entropy is zero when the temperature is zero, and hence the mean energy ε is also zero (see section 6.8 for the relationship of mean energy and temperature). However, from equation (6.11), in our framework this happens when the mean energy and volume are related by

$$\varepsilon V^{2/3} = \frac{3}{4\pi e}\frac{h^2}{m_0} \Leftrightarrow \varepsilon = \frac{3}{4\pi e}\frac{h^2}{m_0 L^2}$$

where L is the edge of a cube with volume V. The Planck's constant h is 6.626×10^{-34} m^2 kg s^{-1}. To make the above quantity as large as possible, we consider the monatomic molecule with the lowest atomic number, i.e. helium, with $m_0 = 6.65 \times 10^{-27}$ kg, and we find that $\varepsilon V^{2/3} = 5.8 \times 10^{-42}$ J m^2. This super-tiny quantity is practically no different from zero, yet one could say that even such a miniscule difference violates the Third Law.

However, it is better to acknowledge that for such super-tiny scales, the formulations based on the classical laws of thermodynamics cease to be applicable. In this case, one would better think in terms of quantum mechanics. Even though such energy levels are out of the scope of this chapter, we may recall from quantum mechanics that the lowest possible energy of a particle is not zero. Rather, due to Heisenberg's uncertainty principle, there is a lower limit, called the *zero-point energy* and given as

$$\varepsilon_1 = \frac{1}{8}\frac{h^2}{m_0 L^2}$$

Comparing the above two equations, we infer that the entropy-zeroing energy is $\varepsilon = (6/\pi e)\varepsilon_1 \approx 0.7\varepsilon_1 < \varepsilon_1$ and this means it cannot be attainable. Hence our version of the Third Law is in better agreement with quantum physics than the classical version which implies that zero energy is in principle feasible.

It is reminded that for continuous stochastic variables, as the variance (in our case represented by the average energy) tends to zero, the entropy tends to minus infinity. It does not tend to zero, which is the case for discrete variables. And our framework is consistent with this fact, because both energy and volume are regarded as continuous variables. If we changed our approach so as to quantize the variables, then the entropy would be zero when the energy were zero. However, treating the variables as continuous is more advantageous and simpler—and the result is in better agreement with quantum mechanics.

6.3 Extension to the motion of a diatomic molecule

We proceed to examine what happens if the particle is a diatomic molecule. Analysis of diatomic gases is important for atmospheric physics because the dominant atmospheric gases are diatomic (N_2, O_2; see Digression 6.E). In a diatomic gas, in addition to the kinetic energy, we have rotational energy at two axes x_4 and x_5 perpendicular to the axis defined by the two atoms. These energies are $L_4^2/2I$ and $L_5^2/2I$, where L denotes angular momentum at the two axes x_4 and x_5 (dimensions [M L^2 T^{-1}] and I denotes rotational inertia with dimensions [M L^2]. Due to symmetry, $I_4 = I_5 = I$.

We consider again the same cube with edge a, containing identical diatomic molecules, each one with mass m_0, rotational inertia I, and total internal energy (sum of kinetic and rotational energy) ε. Each molecule is described by eight variables, three indicating its position, $\underline{x_i}$, three indicating its velocity $\underline{u_i}$ $(i = 1,2,3)$ and two indicating its rotation, $\underline{u}_4 = \underline{L}_4/\sqrt{Im_0}$ and $\underline{u}_5 = \underline{L}_5/\sqrt{Im_0}$. Note that the coordinates $\underline{u}_4$ and $\underline{u}_5$ were chosen so as have the same dimensions as all other $\underline{u_i}$ and $m_0 u_i^2/2$, $i = 4,5$ represent the

rotational energy ($m_0 u_4^2/2 = m_0 \underline{L}_4^2/2Im_0 = \underline{L}_4^2/2I$ and likewise for $i = 5$). The coordinates $\underline{x}_1, \underline{x}_2, \underline{x}_3$ and the five $\underline{u}_i$ are represented as stochastic variables, forming the vector $\underline{\boldsymbol{z}} = \left(\underline{x}_1, \underline{x}_2, \underline{x}_3, \underline{u}_1, \underline{u}_2, \underline{u}_3, \underline{u}_4, \underline{u}_5\right)$.

The background density $\beta(x)$ in $\ln(f(x)/\beta(x))$ should have units $[\boldsymbol{z}^{-1}] = [x^{-3}]\,[u^{-5}] = [\mathrm{L}^{-8}\,\mathrm{T}^5]$. Combining the Planck constant h with the particle mass m_0 and rotational inertia I, we observe that the required dimensions are attained by the quantity $m_0^4 I/h^5$, thereby giving the entropy as

$$\Phi\left[\underline{\boldsymbol{z}}\right] \coloneqq \mathrm{E}\left[-\ln\left(\frac{h^5}{m_0^4 I} f\left(\underline{\boldsymbol{z}}\right)\right)\right] = -\int_{\Omega} \ln\left(\frac{h^5}{m_0^4 I} f(\boldsymbol{z})\right) f(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} \tag{6.13}$$

Application of the principle of maximum entropy with constraints (2.76), (6.1), (6.2) and (6.3) will give the density of $\underline{\boldsymbol{z}}$ as:

$$f(\boldsymbol{z}) = \left(\frac{1}{a}\right)^3 \left(\frac{5m_0}{4\pi\varepsilon}\right)^{5/2} \exp\left(-\frac{5m_0}{4\varepsilon}\left\|\underline{u}\right\|^2\right), \qquad 0 \le x_i \le a \tag{6.14}$$

which is again uniform for the location coordinates and Gaussian for the translational and rotational coordinates. This density indicates independence of all eight components and equipartitioning of the internal energy ε, which now is the sum of kinetic and rotational energy. Specifically, the energy per degree of freedom i is $\mathrm{E}[m_0 u_i^2/2] = \varepsilon/5$. For the rotational components it equals $\mathrm{E}[\underline{L}_4^2/2I]$ and $\mathrm{E}[\underline{L}_5^2/2I]$. The entropy is then calculated as

$$\Phi\left[\underline{\boldsymbol{z}}\right] = \frac{5}{2}\ln\left(\frac{4\pi\mathrm{e}}{5}\,\frac{m_0^{3/5} I^{2/5}}{h^2}\,\varepsilon V^{2/5}\right) = \frac{5}{2}\ln\frac{\varepsilon}{\varepsilon^*} + \ln\frac{V}{v^*} \tag{6.15}$$

for appropriately chosen constants ε^*, v^*, so that the middle term of the equation equal the rightmost one.

6.4 Generalization of the entropy of a single particle

In the previous sections we have maximized the entropy of a monatomic molecule, which has three translational degrees of freedom, and a diatomic molecule, which has two additional rotational degrees of freedom. A molecule with more atoms and a nonlinear structure has three rotational degrees of freedom. It may also have vibrational degrees of freedom associated with vibrations of atoms in the molecule's structure. Vibrations appear also in solids, but in this case these rely on connections among molecules which are also associated with vibrational degrees of freedom. Apparently, though, molecules of solids do not have translational degrees of freedom. In liquids there may also be connections among molecules—a "social behaviour", which is prominent in water molecules—and again correspond to vibrational degrees of freedom. As liquids are fluids, their molecules have also translational degrees of freedom.

In all these types of motion, the energy is proportional to the square of a characteristic quantity, which is velocity for the translational motion, angular velocity for rotational motion and relative extension or compression in a vibrational motion. We may thus denote the energy of the *i*th degree of freedom of the molecule as:

$$\underline{\varepsilon}_i = \frac{1}{2} m_i \underline{u}_i^2 \tag{6.16}$$

where $\underline{u}_i$ is the characteristic quantity and m_i is a proportionality constant. As we have seen, for the translational degrees of freedom m_i is the mass molecule ($m_i = m_0$) and $\underline{u}_i$ is the velocity, but in the other types of motion these quantities are different. The internal energy of the molecule is thus the sum of energy for all degrees of freedom, denoted as β^*:

$$\underline{\varepsilon}_{\mathrm{tot}} = \sum_{i=1}^{\beta} \frac{m_i \underline{u}_i^2}{2} \tag{6.17}$$

Conservation of energy entails

$$\mathrm{E}\left[\underline{\varepsilon}_{\mathrm{tot}}\right] = \mathrm{E}\left[\sum_{i=1}^{\beta} \frac{m_i \underline{u}_i^2}{2}\right] = \varepsilon \tag{6.18}$$

The vector of stochastic variables representing the molecule's state is $\underline{\boldsymbol{z}} = \left(\underline{x}_1, \underline{x}_2, \underline{x}_3, \underline{u}_1, \underline{u}_2, \dots, \underline{u}_\beta\right)$. To find its probability density $f(\boldsymbol{z})$ we first form the entropic representation:

$$\Phi\left[\underline{\boldsymbol{z}}\right] \coloneqq \mathrm{E}\left[-\ln\left(\frac{f\left(\underline{\boldsymbol{z}}\right)}{C^\beta}\right)\right] = -\int_{\Omega} \ln\left(\frac{f\left(\underline{\boldsymbol{z}}\right)}{C^\beta}\right) f(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \tag{6.19}$$

where C is a physical constant. Obviously, the quantity $f\left(\underline{\boldsymbol{z}}\right)\mathrm{d}\boldsymbol{z} = f\left(\underline{\boldsymbol{z}}\right)\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3\mathrm{d}u_1 \dots \mathrm{d}u_\beta$ is dimensionless, while $[u_i] = [\varepsilon/m_i]$, $\left[\mathrm{d}u_1 \dots \mathrm{d}u_\beta\right] = [\varepsilon/m_{\mathrm{G}}]^\beta$, where m_{G} is the geometric mean of m_i, i.e.,

$$m_{\mathrm{G}} \coloneqq \left(m_1 \dots m_\beta\right)^{1/\beta} \tag{6.20}$$

Hence $[\mathrm{d}\boldsymbol{z}] = [x^3]\,[\varepsilon/m_{\mathrm{G}}]^\beta = [C^{-\beta}]$ and $[C] = \left[x^{-3/\beta}\right][\varepsilon^{-1}][m_G]$.

Application of the principle of maximum entropy with constraints (2.76), (6.1), (6.2) and (6.3) will give the density of $\underline{\boldsymbol{z}}$ as:

$$f(\boldsymbol{z}) = \left(\frac{\beta m_{\mathrm{G}}}{4\pi\varepsilon}\right)^{\beta/2} \frac{1}{V} \exp\left(-\frac{\beta}{2\varepsilon}\sum_{i=1}^{\beta} \frac{m_i u_i^2}{2}\right), \qquad 0 \le x_i \le a \tag{6.21}$$

which is again uniform for the location coordinates and Gaussian for the motional coordinates. Again, it indicates independence of all components and equipartitioning of the internal energy ε, which now is the sum of different types of energy. Specifically, the energy per degree of freedom i is $\mathrm{E}[m_i u_i^2/2] = \varepsilon/\beta$. For appropriately chosen constants ε^*, v^*, the entropy is then calculated as

---

* The symbol β is used here for the number of quadratic degrees of freedom of a molecule (as well as for the background-measure density $\beta(x)$ introduced in Chapter 2). The multitude of concepts arising in the union of stochastics and thermodynamics has exhausted both the Greek and the Latin alphabets.

$$\Phi\left[\underline{\boldsymbol{z}}\right] = \frac{\beta}{2}\ln\left(\frac{4\pi e C^2}{\beta m_{\mathrm{G}}}\varepsilon\, V^{2/\beta}\right) = \frac{\beta}{2}\ln\frac{\varepsilon}{\varepsilon^*} + \ln\frac{V}{v^*} \tag{6.22}$$

This is a generalization of equations (6.11) and (6.15), for which we stress again that the entropy $\Phi[\underline{\boldsymbol{z}}]$ and each of its components are dimensionless quantities.

6.5 The principle of maximum entropy applied to *N* molecules

The generalization from the stochastics of one particle, studied in sections 6.2–6.4, to many, say N, molecules of a gas in a container of volume V is not difficult. Once there is no constraint imposing dependence, the principle of maximum entropy will result in independence as already seen for a single particle), which makes things easy. Here the N molecules are assumed to be of the same kind and thus each one has the same degrees of freedom β and the same characteristic constant m_i for each v_i.

We denote by $\underline{x}_{ij}$ the ith location coordinate of the jth molecule ($i = 1,2,3; j = 1, \dots, N$) and $\underline{u}_{ij}$ the ith motion coordinate of the jth molecule ($i = 1, \dots, \beta; j = 1, \dots, N$), E the total internal energy of the N molecules, $\varepsilon = E/N$ is the average energy per particle, and $\underline{\boldsymbol{Z}} = (\underline{\boldsymbol{z}}_1, \dots, \boldsymbol{z}_N)$ the vector with $3N$ location coordinates and βN motion coordinates. Conservation of energy yields:

$$\mathrm{E}\left[\sum_{j=1}^{N}\sum_{i=1}^{\beta}\frac{m_i \underline{u}_{ij}^2}{2}\right] = E \tag{6.23}$$

Application of the principle of maximum entropy with constraints (2.76), (6.1), (6.2) and (6.23) gives:

$$f(\boldsymbol{Z}) = \left(\frac{\beta m_{\mathrm{G}}}{4\pi\varepsilon}\right)^{\beta N/2}\frac{1}{V^N}\exp\left(-\frac{\beta}{2\varepsilon}\sum_{j=1}^{N}\sum_{i=1}^{\beta}\frac{m_i \underline{u}_{ij}^2}{2}\right), \qquad 0 \le x_{ij} \le a \tag{6.24}$$

which is again uniform for the location coordinates and Gaussian for all other coordinates, and indicates independence of all components and equipartitioning of energy into each of the βN degrees of freedom

The entropy for N particles is then calculated as

$$\Phi\left[\underline{\boldsymbol{Z}}\right] = \frac{\beta N}{2}\ln\frac{\varepsilon}{\varepsilon^*} + N\ln\frac{V}{v^*} \tag{6.25}$$

While equation (6.25) has been deduced for a gas, a similar equation should hold for liquids, while for solids the rightmost term ($N \ln V$) should be dropped because the molecules are not in free translational motion. Note though that the number of degrees of freedom in liquids and solids are difficult to infer by deduction and it is more convenient to infer by induction through experimental measurements.

An interesting property that can be observed in equation (6.25) and has important consequences is that, for constant mean energy per particle ε and constant volume V, the entropy is proportional to the number of particles N, i.e., $\Phi\left[\underline{\boldsymbol{Z}}\right] = AN$, where A depends on

ε, V, and the constants appearing in (6.25), but not on N. Now, within the fixed volume V let us consider a part $V' < V$, whose boundaries are not walls, thus allowing particles entering and leaving V'. In this case, the number $\underline{n}$ of particles in V' is not fixed and it can be regarded as a stochastic variable. Due to uniformity, its average will be $N' \coloneqq \mathrm{E}[\underline{n}] = NV'/V$. For the volume V', the conditional entropy for known number of particles n is, obviously, $\Phi[\underline{\boldsymbol{Z}}'|n] = An$, where in $\underline{\boldsymbol{Z}}'$ only those coordinates that fall within V' are counted. Since entropy is an expected value, the unconditional entropy will be

$$\Phi[\underline{\boldsymbol{Z}}'] = \sum_{i=0}^{N} \Phi[\underline{\boldsymbol{Z}}'|n]P(n) = A\sum_{i=0}^{N} nP(n) = AN' \tag{6.26}$$

where $P(n)$ is the probability mass function of n and the sum of $nP(n)$ is by definition the average of n. This indicates that the entropy for the partial volume V' is given by the same formula that provides the entropy of the fixed volume V. In other words, the same entropic expression is valid irrespective of its shape and of whether the boundaries of a certain volume have fixed or free of walls, where in the latter case the average number of particles is used. This result would not hold if there was no proportionality between Φ and N.

6.6 The standardized entropies

In very large systems, such as the atmosphere as a whole, all physical quantities change with location and the homogeneity (independence of expected values from location) assumed in our gas container example does not hold. Still the equations we have derived are valid but at a local scale, i.e. at a small volume V containing N particles (on average), for which homogeneity can be assumed. In this case, however, the choice of V becomes subjective. For a local view, it is convenient to define an objective entropic quantity.

We note first that equation (6.25) expresses the entropy of the finest partition (which in Chapter 2 was denoted as $\mathbb{V}$ and on which the entropy was maximized). It is easily seen that it satisfies both properties of additivity (cf. equation (2.15)) and extensivity. Namely, with respect to Corollary 4.3, we have $\Phi[\underline{\boldsymbol{Z}}]/N = (\beta/2)\ln(\varepsilon/\varepsilon^*) + \ln(V/v^*)$, which is constant, independent of N, when V is specified. Setting $\Phi(E, V, N) \coloneqq \Phi[\underline{\boldsymbol{Z}}]$, for any $a > 0$ this results in

$$\Phi(aE, V, aN) - a\Phi(E, V, N) = 0 \tag{6.27}$$

On the other hand,

$$\Phi(aE, aV, aN) - a\Phi(E, V, N) = aN\ln a \neq 0 \tag{6.28}$$

which means that the entropy is not extensive with respect to the volume V. However, we can define an entropic quantity that is extensive also with respect to V. This can be achieved by defining the entropy of a partition different from the finest.

Specifically, we consider a partition $\mathbb{B}$, in which only the particle location, discretized into N bins, matters. The number N of bins was chosen in order for the discretization of the volume V not to be a subjective choice. Clearly, in gases (and fluids in general) there

are N and N^N ways of placing one and N particles, respectively, in the N bins, so that this partition's entropies are $\Phi_1(\mathbb{B}) = \ln N$ and $\Phi_N(\mathbb{B}) = N \ln N$, respectively*.

Based on equation (2.26), noting that $\Phi\left[\underline{\boldsymbol{Z}}\right] = \Phi[\mathbb{V}]$ is the entropy of the finest partition, which is necessarily a refinement of any other partition, including $\mathbb{B}$, we find that the conditional entropy for 1 and N particles are:

$$\begin{aligned} \Phi\left[\underline{\boldsymbol{z}}|\mathbb{B}\right] &= \Phi_1[\mathbb{V}] - \Phi_1(\mathbb{B}) = \Phi\left[\underline{\boldsymbol{z}}\right] - \Phi_1(\mathbb{B}) = \Phi\left[\underline{\boldsymbol{z}}\right] - \ln N \\ \Phi\left[\underline{\boldsymbol{Z}}|\mathbb{B}\right] &= \Phi_N[\mathbb{V}] - \Phi_N(\mathbb{B}) = \Phi\left[\underline{\boldsymbol{Z}}\right] - \Phi_N(\mathbb{B}) = \Phi\left[\underline{\boldsymbol{Z}}\right] - N \ln N \end{aligned} \tag{6.29}$$

Denoting

$$v \coloneqq \frac{V}{N} \tag{6.30}$$

the volume per particle, which is an intensive quantity that can be assumed fairly uniform at a local level, we identify the above conditional entropies with what we call *standardized entropies* (intensive and extensive). More specifically, we define the *standardized entropy per particle* as:

$$\varphi^*(\varepsilon, v) \coloneqq \Phi\left[\underline{\boldsymbol{z}}|\mathbb{B}\right] = \frac{\beta}{2} \ln \frac{\varepsilon}{\varepsilon^*} + \ln \frac{v}{v^*} \tag{6.31}$$

and the *standardized total entropy*:

$$\Phi^*(E, V, N) \coloneqq \Phi\left[\underline{\boldsymbol{z}}|\mathbb{B}\right] = \ln\left(\frac{E}{N\varepsilon^*}\right) + N \ln\left(\frac{V}{Nv^*}\right) \tag{6.32}$$

where obviously

$$\Phi^*(E, V, N) = N\varphi^*(\varepsilon, v) \tag{6.33}$$

Both quantities $\varphi^*(\varepsilon, v)$ and $\Phi^*(E, V, N)$ are invariant under change of V, provided that the volume per particle v is fairly uniform. The standardized values φ^* and Φ^* quantify how much larger the finest-partition entropy $\Phi\left[\underline{\boldsymbol{z}}\right]$ or $\Phi\left[\underline{\boldsymbol{Z}}\right]$, respectively, is from the discretized location partition entropies $\Phi_1(\mathbb{B})$ and $\Phi_N(\mathbb{B})$, respectively. Put it otherwise, $\Phi^*(E, V, N)$ is identical to the probabilistic entropy of a system of N particles, conditional on each being in a volume v. Likewise, $\varphi^*(\varepsilon, v)$ is identical to the probabilistic entropy of a molecule with a fixed volume equal to v.

Like the energy per particle, ε, and the volume per particle, v, the standardized entropy per particle $\varphi^*(\varepsilon, v)$, is an *intensive* property in the sense that it does not depend on the size of system that an observer, justifiably or arbitrarily, considers.

In contrast, the total energy, E, the volume, V, and the number of particles, N, are *extensive* properties in the sense that they depend on the observer's selection of the system and are proportional one another; that is, a system of volume αV, where α is any positive number contains αN particles with a total energy αE. Likewise, the standardized

* Notably in solids the locations of particles are fixed (only one possible way) and thus $\Phi_1(\mathbb{B}) = \Phi_N(\mathbb{B}) = 0$ (because $\ln 1 = 0$). Thus, the standardized entropies φ^* and Φ^* in solids become identical to φ and Φ, respectively, which agrees with the classical result for solids.

total entropy $\Phi^*(E,V,N)$ is indeed an extensive property in terms of all (E,V,N), as it is easily seen that

$$\Phi^*(\alpha E,\alpha V,\alpha N)=\alpha\Phi^*(E,V,N) \tag{6.34}$$

We note that in classical statistical thermodynamics the name 'entropy' is used for the quantity Φ^* calculated from equation (6.32) and multiplied by the Boltzmann constant k, i.e. $S=k\Phi^*$. However, it is derived in a different manner, assuming indistinguishability of particles. As discussed in section 5.1, here we have rejected this interpretation because it has several logical problems (see also Koutsoyiannis, 2013a, for additional details for the case of gases).

Descriptions using either actual entropies Φ or standardized entropies Φ^* are generally equivalent and describe the same thing. As will be seen in section 6.7, using Φ^* is mathematically simpler and more convenient. It should be stressed, though, that the actual entropy of the finest partition is given by Φ, not Φ^*. Therefore, in most of derivations here we use Φ. In some cases descriptions based on Φ^* may yield paradoxical results. In such cases, we should resort to the actual entropy Φ. Contrary to Φ^*, Φ works always, offers additional insights, and avoids paradoxical results. We see this in the example of the well-known Gibbs paradox in Digression 6.D (again, see Koutsoyiannis, 2013a, for additional details).

6.7 Equivalence of descriptions by actual and standardized entropies

To illustrate the particularities and eventually the equivalence of the descriptions by probabilistic entropy (Φ) and standardized entropy (Φ^*) we assume again that N molecules are in motion in a container of volume V and entropy per particle φ. We assume a partition of the container into two parts A and B with volumes V_A and V_B, respectively, with $V_\mathrm{A}+V_\mathrm{B}=V$.

First off, we will clarify the concepts by examining the case of just two particles 1 and 2 and using the more insightful actual entropies, rather than the standardized entropies. We examine two cases: (a) the separation into parts A and B is mental, so that in fact both particles move into the entire volume V, and (b) the separation is physical, through a diaphragm, and each of the particles 1 and 2 moves into parts A and B respectively. The (probabilistic) ground sets for particles 1 and 2 are denoted Ω_1 and Ω_2, respectively and the ground set for both is the Cartesian product $\Omega_1\times\Omega_2$. We examine the finest partitions for each of the particles, denoted as $\mathbb{V}_1$ and $\mathbb{V}_2$.

In case (a), the entropy of each of the two particles is

$$\varphi_1=\Phi(\mathbb{V}_1)=\frac{\beta}{2}\ln\frac{\varepsilon}{\varepsilon^*}+\ln\frac{V}{v^*}=\varphi_2=\Phi(\mathbb{V}_2) \tag{6.35}$$

If each of the particles moves independently of the other, then according to equation (2.15) the total entropy is

$$\Phi=\Phi(\mathbb{V}_1\otimes\mathbb{V}_2)=\varphi_1+\varphi_2=2\left(\frac{\beta}{2}\ln\frac{\varepsilon}{\varepsilon^*}+\ln\frac{V}{v^*}\right) \tag{6.36}$$

If there is dependence, then, according to inequality (2.16), the entropy will be smaller. As the most extreme example, we assume that the two particles are tied together by an attractive force, so that the total entropy will be

$$\Phi_{\mathrm{a}} = \Phi_{\mathrm{a}}(\mathbb{V}_1 \otimes \mathbb{V}_2) = \varphi_1 = \frac{\beta}{2}\ln\frac{\varepsilon}{\varepsilon^*} + \ln\frac{V}{v^*} < \varphi_1 + \varphi_2 \tag{6.37}$$

We will also calculate the total entropy from conditional entropies assuming that, at any instance, any particle can be either in part A with probability P_{A}, or in part B with probability $P_{\mathrm{B}} = 1 - P_{\mathrm{A}}$. We denote the conditional entropy per particle for each of the two parts as φ_{A} and φ_{B}, respectively. The unconditional entropy (for the entire volume) can be calculated from the conditional entropies according to equation (2.27), where the bipartition $\mathbb{B}$ is in this case $\mathbb{B} = \{\boldsymbol{z} \text{ in A}, \boldsymbol{z} \text{ in B}\}$ and has partition entropy

$$\varphi_P \coloneqq \Phi(\mathbb{B}) = -P_{\mathrm{A}}\ln P_{\mathrm{A}} - P_{\mathrm{B}}\ln P_{\mathrm{B}} \tag{6.38}$$

Hence the total entropy is

$$\varphi_1 = P_{\mathrm{A}}\varphi_{\mathrm{A}} + P_{\mathrm{B}}\varphi_{\mathrm{B}} + \varphi_P > P_{\mathrm{A}}\varphi_{\mathrm{A}} + P_{\mathrm{B}}\varphi_{\mathrm{B}} \tag{6.39}$$

where

$$\varphi_{\mathrm{A}} = \frac{\beta}{2}\ln\frac{\varepsilon}{\varepsilon^*} + \ln\frac{V_{\mathrm{A}}}{v^*}, \qquad \varphi_{\mathrm{B}} = \frac{\beta}{2}\ln\frac{\varepsilon}{\varepsilon^*} + \ln\frac{V_{\mathrm{B}}}{v^*} \tag{6.40}$$

Hence:

$$\varphi_1 = \varphi_{\mathrm{A}} = \frac{\beta}{2}\ln\frac{\varepsilon}{\varepsilon^*} + P_{\mathrm{A}}\ln\frac{V_{\mathrm{A}}}{v^*} + P_{\mathrm{B}}\ \ln\frac{V_{\mathrm{B}}}{v^*} + P_{\mathrm{A}}\ln V_{\mathrm{A}} + P_{\mathrm{B}}\ln V_{\mathrm{B}} - P_{\mathrm{A}}\ln P_{\mathrm{A}} - P_{\mathrm{B}}\ln P_{\mathrm{B}} \tag{6.41}$$

or

$$\varphi_1 = \frac{\beta}{2}\ln\frac{\varepsilon}{\varepsilon^*} + P_{\mathrm{A}}\ln\frac{V_{\mathrm{A}}}{v^* P_{\mathrm{A}}} + P_{\mathrm{B}}\ \ln\frac{V_{\mathrm{B}}}{v^* P_{\mathrm{A}}} \tag{6.42}$$

and since the probabilities should be in proportion with volume we have $V_{\mathrm{A}}/P_{\mathrm{A}} = V_{\mathrm{B}}/P_{\mathrm{B}} = V$ and thus we arrive again in equation (6.35). The total entropy of the two particles is, again, twice φ_1.

In case (b), assuming that the energies in the two particles are the same as in case (a), the entropies per particle are:

$$\varphi_1 = \Phi(\mathbb{V}_1) = \varphi_{\mathrm{A}}, \qquad \varphi_2 = \Phi(\mathbb{V}_2) = \varphi_B \tag{6.43}$$

Since each of the particles moves in a separate part, independence can be assumed and, according to equation (2.15), the total entropy is

$$\Phi_{\mathrm{b}} = \Phi_{\mathrm{b}}(\mathbb{V}_1 \otimes \mathbb{V}_2) = \varphi_1 + \varphi_2 = 2\left(\frac{\beta}{2}\ln\frac{\varepsilon}{\varepsilon^*}\right) + \ln\frac{V_{\mathrm{A}}}{v^*} + \ln\frac{V_{\mathbf{B}}}{v^*} < \Phi_{\mathrm{a}} \tag{6.44}$$

The inequality in the rightmost part is inferred from the fact that $\ln V_A + \ln V_B < 2\ln(V_A + V_B)$, which is always true, as can be readily seen (e.g. by rewriting the inequality as $\ln(V_A V_B) < \ln((V_A + V_B)^2)$ or $V_A V_B < (V_A + V_B)^2$). The fact that $\Phi_{\mathrm{b}} < \Phi_{\mathrm{a}}$ is justified from the larger phase space of case (a) compared to case (b).

Now we examine the case (a) (no material separation) for very many particles, so that we can substitute N_A/N for P_A and N_B/N for P_B, where N_A and N_B are the number of particles in parts A and B, respectively. According to equation (6.39), the entropy for each of the particles is

$$\varphi = \frac{N_\mathrm{A}}{N}\varphi_\mathrm{A} + \frac{N_\mathrm{B}}{N}\varphi_\mathrm{B} - \frac{N_\mathrm{A}}{N}\ln\frac{N_\mathrm{A}}{N} - \frac{N_\mathrm{B}}{N}\ln\frac{N_\mathrm{B}}{N} \tag{6.45}$$

and hence

$$N\varphi = N_\mathrm{A}\varphi_\mathrm{A} + N_\mathrm{B}\varphi_\mathrm{B} - N_\mathrm{A}\ln N_\mathrm{A} - N_\mathrm{B}\ln N_\mathrm{B} + N\ln N \tag{6.46}$$

from which we obtain

$$\Phi = \Phi_\mathrm{A} + \Phi_\mathrm{B} + N\ln N - N_\mathrm{A}\ln N_\mathrm{A} - N_\mathrm{B}\ln N_\mathrm{B} \tag{6.47}$$

This can be written in terms of standardized entropies as

$$\Phi^* = \Phi_\mathrm{A}^* + \Phi_\mathrm{B}^* \tag{6.48}$$

Equations (6.47) and (6.48) are in this case precisely equivalent and describe the same thing, using either actual entropies Φ or standardized entropies Φ^*.

Now let us examine case (b) (separation by a diaphragm that does not allow molecules to move from one part to the other) for many particles. The two parts should be treated as separate and independent. The entropy of the whole should correspond to the ground set $\Omega_\mathrm{A} \times \Omega_\mathrm{B}$, where Ω_A and Ω_B are the ground sets of the two parts. The Cartesian product is justified because we are simultaneously interested in what happens in each of the two parts, while the two parts are separate. Hence the entropy we are looking for is that of the product partition, i.e., $\Phi(\mathbb{A} \otimes \mathbb{B})$, and, according to equation (2.15) it equals the sum of the two entropies:

$$\Phi = \Phi_\mathrm{A} + \Phi_\mathrm{B} \tag{6.49}$$

which is quite different from equation (6.47). We can assume (without delving into further into conceptual details) that equation (6.48) for the standardized entropies, but now we cannot assume that the standardized entropy Φ^* is given by the difference $\Phi - N\ln N$ for the total system. Rather it is given by the sum $\Phi_\mathrm{A}^* + \Phi_\mathrm{B}^*$.

The characteristics of the two cases (a) and (b) for many particles, including the results of the analyses are summarized in Table 6.1.

Table 6.1 Total and per particle characteristic quantities, including the results of these analyses, for a gas container comprising two parts for two cases: (a) without a diaphragm between the two parts (mass exchange is allowed) and (b) with a diaphragm between the two parts.

	No diaphragm (mass exchange)		Diaphragm (no mass exchange)	
	Total	Per particle	Total	Per particle*
Number of particles	$N = N_A + N_B$	1	$N = N_A + N_B$	1
Volume	$V = V_A + V_B$	$v = V/N$	$V = V_A + V_B$	$v_\mathrm{A} = V_\mathrm{A}/N, v_\mathrm{B} = V_\mathrm{B}/N$
Energy	$E = E_A + E_B$	$\varepsilon = E/N$	$E = E_A + E_B$	$\varepsilon_\mathrm{A} = E_\mathrm{A}/N, \varepsilon_\mathrm{B} = E_\mathrm{B}/N$
Actual entropy	$\Phi = \Phi_\mathrm{A} + \Phi_\mathrm{B} + N\varphi_P$	$\varphi = \Phi/N$	$\Phi_\mathrm{A} + \Phi_\mathrm{B}$	$\varphi_\mathrm{A} = \Phi_\mathrm{A}/N, \varphi_\mathrm{B} = \Phi_\mathrm{B}/N$
Standardized entropy	$\Phi^* = \Phi_\mathrm{A}^* + \Phi_\mathrm{B}^*$	$\varphi^* = \Phi^*/N$	$\Phi^* = \Phi_\mathrm{A}^* + \Phi_\mathrm{B}^*$	$\varphi_\mathrm{A}^* = \Phi_\mathrm{A}^*/N, \varphi_\mathrm{B}^* = \Phi_\mathrm{B}^*/N$

* The quantities per particle are different for the two parts.

Digression 6.C: Does the standardized entropy per particle depend on N?

Considering the volume partition in two parts as in section 6.8, the total entropy in part A is

$$\Phi_{\mathrm{A}}^{*}=\Phi_{\mathrm{A}}-N_{\mathrm{A}}\ln N_{\mathrm{A}}=\frac{\beta N}{2}\ln\left(\frac{E_{\mathrm{A}}}{N_{\mathrm{A}}\varepsilon^{*}}\right)+N\ln\left(\frac{V_{\mathrm{A}}}{N_{\mathrm{A}}v^{*}}\right)$$

According to equation (6.33), the standardized entropy per particle is

$$\varphi_{\mathrm{A}}^{*}=\frac{\Phi_{\mathrm{A}}^{*}}{N}=\frac{\beta}{2}\ln\frac{E_{\mathrm{A}}}{N_{\mathrm{A}}\varepsilon^{*}}+\ln\frac{V_{\mathrm{A}}}{N_{\mathrm{A}}v^{*}}=\frac{\beta}{2}\ln\frac{\varepsilon_{\mathrm{A}}}{\varepsilon^{*}}+\ln\frac{V_{\mathrm{A}}}{v^{*}}-\ln P_{A}-\ln N$$

or

$$\varphi_{\mathrm{A}}^{*}=\varphi_{\mathrm{A}}-\ln P_{A}-\ln N$$

Since neither the actual entropy φ_{A} nor the probability P_{A} depend on N, the standardized entropy φ_{A}^{*} does depend on N. However, we may omit the term $-\ln N$ and write

$$\varphi_{\mathrm{A}}^{*}=\varphi_{\mathrm{A}}-\ln P_{A}$$

in most cases, i.e.:

- when $N=1$;
- when we maximize entropy for a fixed N, as in this case it is a constant; and
- when we take differences such as: $\varphi_{\mathrm{A}}^{*}-\varphi_{\mathrm{B}}^{*}=(\varphi_{\mathrm{A}}-\ln P_{A})-(\varphi_{\mathrm{B}}-\ln P_{B})$.

Digression 6.D: The Gibbs paradox

We consider two identical parts of a box separated by a diaphragm, each containing N identical particles with energy E in a volume V. The entropy in each part is $\Phi(E,V,N)$. Since independence between the two parts can be assumed, the total entropy is $2\Phi(E,V,N)$. If we remove the diaphragm, the entropy becomes $\Phi(2E,2V,2N)$ and according to (6.28) with $\alpha=2$, there is an increase of entropy equal to $2N\ln 2$. The same result is obtained from (6.47) by replacing N with $2N$ and N_{A} and N_{B} with N. If we reinsert the diaphragm, entropy will become again $2\Phi(E,V,N)$ and there will be a decrease of entropy, $-2N\ln 2$.

On the other hand, if we describe the process in terms of standardized entropies, before we remove the diaphragm, the standardized entropy in each part will be $\Phi^{*}(E,V,N)$ and totally $2\Phi^{*}(E,V,N)$. After the removal of the diaphragm, the standardized entropy becomes $\Phi^{*}(2E,2V,2N)$ and according to (6.34) with $\alpha=2$, this is again $2\Phi^{*}(E,V,N)$. So no change of standardized entropy is seen. When no change in entropy occurs, the process is reversible. This, however, is incorrect, as can be understood if in the two compartments there were, e.g., different monatomic gases before the diaphragm was removed. Apparently, after its removal, there will be a mixture and the mixing cannot be reversed. These have been regarded as contradictions to thermodynamics, known as the Gibbs paradox (Jaynes, 1992; Swendsen, 2008). And since mixing happens, there must be an explanation why it does.

This explanation is given by the actual entropy Φ and its interpretation as quantification of uncertainty. The increase of Φ by $2N\ln 2$ shows that the process is irreversible and there is no contradiction or paradox. The increase of entropy after the removal of the diaphragm, quantifies the fact that we have greater uncertainty about the location of each particle. Initially, we knew each particle's location with "precision" V and after the removal we lost information, with the "precision" becoming $2V$. The increase of entropy objectively expresses the lost information, and reflects the physical fact that the motion of each particle has a larger domain. Likewise, the decrease of entropy after reinsertion of the diaphragm (which can hardly be thought of as a spontaneous natural phenomenon) reflects the gain of information and the decreased uncertainty about the position of a particle.

One may contend that the introduction of the diaphragm gives us no information at all about which part any particular particle might be found in. Such an argument is interesting and in fact can be fully addressed by the proposed probabilistic approach. Specifically, if we do not know

whether a particle is in one or the other part, then the entropy of the particle is φ as given by equation (6.39) and is not affected by the diaphragm. However, once the diaphragm is there and once we know that at some time the particle is in a specific part, then the entropy of the particle has been reduced to $\varphi_A < \varphi$ (the leftmost term in (6.39)). The information that the specific particle is in the specific part at a specific time is important because it is transferred to subsequent times. That is, if we know in which part a particle is found and there is a diaphragm, then the entropy is φ_A now and will be the same in the future. In contrast, if the diaphragm is absent, then knowing in which part a particle is found makes the entropy φ_A now, but soon after in the future this information is lost and the entropy becomes φ. Thus, the introduction of the diaphragm is associated with reduction of uncertainty.

According to the proposed approach, this situation does not change if on either side of the diaphragm the gases are different, thus removing the asymmetry (or discontinuity) in the two cases where in the two boxes the gases are different or the same. In both cases, the entropy Φ after the mixing (by removal of the diaphragm) will be by $2N \ln 2$ greater than the sum of the entropies in the initial state. And again, if we consider the mixture of gases as an equivalent single gas (as is the standard practice, for instance, in the atmospheric mixture; see section 6.12) the entropy Φ^* will not change, in comparison to the initial state.

If we assumed that the particles are indistinguishable, then there would be no distinction between Φ and Φ^*, which would equal each other. In this case the paradox would hold in both cases. This is another reason why we should reject the indistinguishability hypothesis, additional to those already stated in Chapter 5.

6.8 Definition of temperature in the stochastic and the classical framework

We define temperature as the inverse of the partial derivative of entropy with respect to internal energy, i.e.,

$$\frac{1}{\theta} \coloneqq \frac{\partial \Phi}{\partial E} \tag{6.50}$$

From equations (6.22), (6.25), (6.31), (6.32) we obtain

$$\frac{1}{\theta} = \frac{\partial \Phi}{\partial E} = \frac{\partial \Phi^*}{\partial E} = \frac{\partial \varphi}{\partial \varepsilon} = \frac{\partial \varphi^*}{\partial \varepsilon} = \frac{\beta}{2\varepsilon} \tag{6.51}$$

We highlight the rightmost term, from which we find:

$$\theta = \frac{2\varepsilon}{\beta} \tag{6.52}$$

That is, the temperature equals twice the particle's internal energy per degree of freedom. Therefore, if we increase the internal energy of matter, the temperature will increase in proportion. If we offer the same amount of internal energy per particle to two objects with different number of degrees of freedom, the temperature increase of the object with the fewer degrees of freedom will be higher.

It is useful to note that, by virtue of equation (6.52), the entropic expression (6.31) can be written in terms of θ as

$$\varphi^*(\theta, v) = \frac{\beta}{2} \ln \frac{\theta}{\theta^*} + \ln \frac{v}{v^*}, \qquad \theta^* \coloneqq \frac{2\varepsilon^*}{\beta} \tag{6.53}$$

We highlight the fact that in our approach energy and entropy are the fundamental quantities, and temperature is a derivative quantity, defined in terms of these

fundamental quantities. Since Φ is dimensionless and E has dimensions of energy, temperature has also dimensions of energy (e.g. joules, J = kg m^2 s^{-2}).

However, in classical thermodynamics, temperature, denoted as T, is the principal quantity, for which a base unit of the *Système International* (SI), is defined, the kelvin (K). Then the entropy, denoted as S, is defined by $\mathrm{d}S \coloneqq \delta Q/T$, where Q denotes heat. The definition is perhaps affected by circularity, as it is valid for reversible processes, which are those in which $\mathrm{d}S = \delta Q/T$, while irreversible are those in which $\mathrm{d}S > \delta Q/T$ (see details in section 6.11, where this classical definition is derived by our stochastic framework, rather than premised).

The classical thermodynamic quantities are related to the natural quantities of our approach by

$$S = k\,\Phi^*, \qquad T = \frac{\theta}{k} \tag{6.54}$$

where k is the Boltzmann constant. Both this constant and the unit of kelvin are historical accidents related to the inadequate understanding of the probabilistic nature of thermodynamic phenomena and the arbitrary introduction of temperature scales. This is clearly manifested in the current (since 2019) definition of the kelvin. Specifically, according to the SI Brochure (Bureau International des Poids et Mesures, 2019):

> "[The kelvin] is defined by taking the fixed numerical value of the Boltzmann constant, k, to be 1.380 649 × 10^{-23} when expressed in the unit J K^{-1}, which is equal to kg m^2 s^{-2} K^{-1} […]
>
> $$1\ \mathrm{K} = \left(\frac{1.380\,649\ \times\ 10^{-23}}{k}\right)\ \mathrm{kg\ m^2 s^{-2}}$$
>
> […] The effect of this definition is that one kelvin is equal to the change of thermodynamic temperature that results in a change of thermal energy kT by 1.380 649 × 10^{-23} J" [=13.80649 yJ (yoctojoules) =0.01380649 zJ (zeptojoules)].

This definition clearly shows that the entire classical thermodynamic setting is artificial and arbitrary, and this extends to the unit of kelvin and the Boltzmann constant k. A more natural setting would be to take $k = 1$ (dimensionless). Then, T and S would be identical to θ and Φ, respectively. Here we obviously prefer the stochastic framework, which is superior to the classical framework for many reasons:

- It is more parsimonious and logically consistent.
- It provides explanations for the processes described by classical thermodynamics.
- It advances awareness about the maximum uncertainty in thermodynamic processes.
- It dispels the delusion of deterministic laws in classical thermodynamics.
- It can be generalized to other processes, involving higher levels of macroscopization.

On the other hand, given the popularity of the classical approach, while we insist on deriving all equations by the stochastic framework, in the subsequent sections we will

also convert the natural quantities to the classical ones. This is not difficult: the only conversion needed is that 1 K = 13.80649 yJ. Occasionally, we will also use the degree Celsius unit (°C), which is by definition equal in magnitude to the unit kelvin. However, the numerical value of temperature expressed in degrees Celsius, T_C is related to the numerical value of the absolute temperature T expressed in kelvins by the relation $T_C/°C = T/K - 273.15$. To distinguish our stochastic quantities from those of common practice, we call θ the *natural temperature* and T the *thermodynamic* or *absolute temperature*. Likewise, we call Φ *probabilistic* (or *actual*) *entropy*, Φ^* *standardized entropy* and S *thermodynamic entropy*.

It is easily seen that equation (6.50) is retained in the classical framework in the form:

$$\frac{1}{T} = \frac{\partial S}{\partial E} \tag{6.55}$$

6.9 The law of ideal gases

We consider again (as in section 6.2) the cube of edge a containing N identical molecules of a gas, each with mass m_0 and β degrees of freedom. Considering a small time interval δt, with the help of Figure 6.2 we deduce that any particle at distance from the bottom edge of the cube $x_3 \leq -u_3 \delta t$ will collide with that edge ($x_3 = 0$).

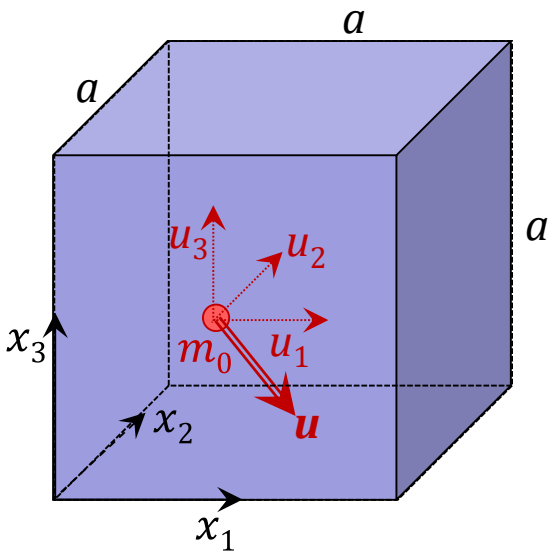

Figure 6.2 Explanation sketch for the collision of a molecule to the bottom edge of a cube. The particle shown, located at coordinate x_3 (distance from the bottom edge), will collide with the bottom edge ($x_3 = 0$) within time δt if $x_3 \leq -u_3 \delta t$. The other location and velocity coordinates do not play any role: for example, if x_1 and/or u_1 are large, a collision with the rightmost edge of the cube may occur within δt, which however will result in horizontal reflection that will not alter the velocity coordinate u_3.

From equation (6.21), we calculate the joint distribution function of the coordinates (x_3, u_3) of a single particle as:

$$f(x_3, u_3) = \frac{1}{a}\left(\frac{\beta m_0}{4\pi\varepsilon}\right)^{1/2} \exp\left(-\frac{\beta m_0}{4\varepsilon} u_3^2\right) \tag{6.56}$$

Thus, the expected value of the momentum $\underline{q}(\delta t)$ of molecules colliding at the cube edge ($x_3 = 0$) within the time interval δt is

$$\mathrm{E}\left[\underline{q}(\delta t)\right] = N \int_0^a \int_{-\infty}^{-x_3/\delta t} m_0 u_3 \; f(x_3, u_3) \mathrm{d}u_3 \mathrm{d}x_3 \tag{6.57}$$

After the calculations we find

$$\mathrm{E}\left[\underline{q}(\delta t)\right] = \frac{N\varepsilon\,\delta t}{a\beta} \tag{6.58}$$

Here we note that the integral originally includes a multiplicative term $\mathrm{Erf}\left((a/2\delta t)\sqrt{\beta m_0/\varepsilon}\right)$, which by virtue of (6.23) (written for one molecule as $\mathrm{E}[u_3^2] = 2\varepsilon/\beta m_0$) becomes $\mathrm{Erf}\left(a/\left(\delta t\sqrt{2\mathrm{E}[u_3^2]}\right)\right)$. As $\delta t \to 0$, this term tends to 1 and thus it is omitted.

According to Newton's Second Law, the force exerted on the edge is $F = 2\mathrm{E}\left[\underline{q}(\delta t)\right]/\delta t$, and the pressure is $p = F/a^2 = 2N\varepsilon/(\beta V)$, or finally (by using (6.52)),

$$p = N\theta/V = \theta/v \Leftrightarrow pV = N\theta \Leftrightarrow pv = \theta \tag{6.59}$$

This is the well-known *law of ideal gases* written in terms of natural temperature. Its above derivation shows that it is not an empirical relationship, but it is deduced by maximizing entropy.

To write the same law in the classical thermodynamic framework we first introduce the well-known *gas constant*:

$$R_* \coloneqq kN_\mathrm{a} = 8\,314.463\ \mathrm{J\,K^{-1}\,kmol^{-1}} \tag{6.60}$$

where $N_\mathrm{a} = 6.022{\times}10^{23}$ mol^{-1} is the Avogadro constant, representing the number of particles per mole of substance. We also define the *molar mass* (kg/kmol),

$$M_0 \coloneqq N_\mathrm{a} m_0 \tag{6.61}$$

the *density* (kg/m^3),

$$\rho \coloneqq \frac{m}{V} = \frac{nM_0}{V} = \frac{m_0}{v} \tag{6.62}$$

where m and n denote the mass and the number of moles in the volume V, and the *specific gas constant* (J K^{-1} kg^{-1})

$$R \coloneqq R_*/M_0 \tag{6.63}$$

Hence, the law of ideal gases takes the following classical thermodynamic forms:

$$pV = nR_*T \;\Leftrightarrow\; \frac{p}{\rho} = RT \tag{6.64}$$

In all this analysis we have not considered gravity, which is justified if the cube edge a is small and gravity does not play any role for the energy. If we consider gravity, then the kinetic energy of the gas will decrease as the molecule goes up and a potential energy

$-m_0 g x_3$ develops. An easy way to find the relationship between pressures at the two horizontal edges is to see that the total force on the bottom edge (B) equals that of the top edge (T) plus the weight of the gas, i.e.,

$$p_{\mathrm{B}} a^2 = p_{\mathrm{T}} a^2 + N m_0 g \tag{6.65}$$

which results in

$$\Delta p \coloneqq p_{\mathrm{T}} - p_{\mathrm{B}} = -\frac{N m_0 g}{a^2} = -\frac{a m_0 g}{v} \tag{6.66}$$

We can write this in differential form, substituting $\mathrm{d}p$ for Δp and $\mathrm{d}x_3$ for a:

$$\frac{\mathrm{d}p}{\mathrm{d}x_3} = -\frac{m_0 g}{v} = \rho g \tag{6.67}$$

as the quantity m_0/v is the density ρ of the gas. This is the well-known hydrostatic law of pressure. We note that equation (6.67) is the simplest case of a pressure profile, which is valid when the vertical component of air motion is negligible (otherwise we should make more sophisticated hydrodynamic considerations). To derive equation (6.67) we have ignored the vertical component of the Coriolis force, as indeed this is a negligent fraction of the gravity force ρg.

6.10 Alternative expressions of entropy

The definition of temperature θ and its relationship with kinetic energy per particle ε (equation (6.52)) along with the law of ideal gases (equation (6.59)) allows expressing the standardized entropy per particle φ^* (equation (6.31)) in terms of temperature and pressure as follows:

$$\varphi^*(\theta, p) = \left(1 + \frac{\beta}{2}\right) \ln \frac{\theta}{\theta^*} - \ln \frac{p}{p^*}, \qquad p^* \coloneqq \frac{\theta^*}{v^*} = \frac{2\varepsilon^*}{\beta v^*} \tag{6.68}$$

We may also derive several useful differential forms of entropy. From (6.25), (6.22), (6.32), (6.31), the partial derivatives of entropy with respect to V and v are

$$\frac{\partial \Phi}{\partial V} = \frac{\partial \Phi^*}{\partial V} = \frac{\partial \varphi}{\partial v} = \frac{\partial \varphi^*}{\partial v} = \frac{p}{\theta}, \qquad \frac{\partial \varphi}{\partial V} = \frac{1}{V} \tag{6.69}$$

where the term p/θ was obtained from the ideal gas law. From the same equations, the partial derivatives with respect to N are

$$\frac{\partial \Phi}{\partial N} = N \frac{\partial \varphi}{\partial N}, \qquad \frac{\partial \Phi^*}{\partial N} = \varphi^* + N \frac{\partial \varphi^*}{\partial N} \tag{6.70}$$

The quantity $\partial \Phi^*/\partial N$ defines the so-called *chemical potential*, μ (cf., Wannier, 1987, p. 139), through the relationship

$$-\frac{\mu}{\theta} \coloneqq \frac{\partial \Phi^*}{\partial N} = \varphi^* + N \frac{\partial \varphi^*}{\partial N} \tag{6.71}$$

Based on this, the probabilistic and the standardized entropy can be written, respectively, in differential form as

$$\mathrm{d}\Phi = \frac{1}{\theta}\mathrm{d}E + \frac{p}{\theta}\mathrm{d}V + \left(-\frac{\mu}{\theta} + 1 + \ln N\right)\mathrm{d}N, \qquad \mathrm{d}\Phi^* = \frac{1}{\theta}\mathrm{d}E + \frac{p}{\theta}\mathrm{d}V - \frac{\mu}{\theta}\mathrm{d}N \tag{6.72}$$

The latter can be written in the equivalent form:

$$\theta\mathrm{d}\Phi^* = \mathrm{d}E + p\mathrm{d}V - \mu\mathrm{d}N \tag{6.73}$$

Furthermore, by using an anchoring point (θ_0, p_0), calculating $\varphi^*(\theta_0, p_0)$ and subtracting the latter from $\varphi^*(\theta, p)$ we find:

$$\varphi^*(\theta, p) = \varphi^*(\theta_0, p_0) + \left(1 + \frac{\beta}{2}\right)\ln\frac{\theta}{\theta_0} - \ln\frac{p}{p_0} \tag{6.74}$$

In classical thermodynamics, considering a reference point $(T^* \coloneqq \theta^*/k\,, p^*)$, to find the classical entropy per unit mass, s, we multiply φ^* by $R \coloneqq k/m_0$ and obtain

$$s = \left(\left(1 + \frac{\beta}{2}\right)\ln\frac{T}{T^*} - \ln\frac{p}{p_0}\right)R \tag{6.75}$$

where usually in atmospheric thermodynamics the anchoring values are arbitrarily chosen $T^* = 200$ K and $p^* = 1000$ hPa. By setting

$$\left(1 + \frac{\beta}{2}\right)R =: c_p, \qquad \frac{\beta}{2}R =: c_v, \qquad c_p - c_v = R \tag{6.76}$$

where c_p and c_v are termed the *isobaric specific heat* (for constant pressure) and *isochoric specific heat* (for constant volume), respectively, we obtain the final classical entropic formula:

$$s = c_p \ln\frac{T}{T_0} - R\ln\frac{p}{p_0} \tag{6.77}$$

We have seen that the statistical framework provided here can produce the classical thermodynamic framework successfully (see Digression 6.E), yet it is more general because the principle of maximum entropy can be applied in other systems.

For easy reference, Table 6.2 provides a summary of classical thermodynamic variables and their connection with the quantities used in our stochastic framework.

Table 6.2 Summary of classical thermodynamic variables and their connection with the quantities used in our stochastic framework.

Quantity (and standard SI unit)	Total	Eqn. no.
Boltzmann constant	$k = 1.380\,649 \times 10^{-23}\ \mathrm{J\,K^{-1}}$,	
Avogadro constant	$N_a = 6.022{\times}10^{23}\ \mathrm{mol^{-1}} = 6.022{\times}10^{26}\ \mathrm{kmol^{-1}}$	(6.60)
Gas constant	$R_* \coloneqq kN_a = 8\,314.463\ \mathrm{J\,K^{-1}\,kmol^{-1}}$	(6.60)
Molecular (or molar) mass (kg/kmol)	$M_0 \coloneqq N_a m_0$	(6.61)
Specific gas constant ($\mathrm{J\,K^{-1}\,kg^{-1}}$)	$R \coloneqq \frac{R_*}{M_0} = \frac{k}{m_0}$	(6.63)
Entropy (J/K)	$S = k\Phi^*$	(6.54)
Entropy per unit mass ($\mathrm{J\,K^{-1}\,kg^{-1}}$)	$s = \frac{S}{m} = R\varphi^*$	
Temperature (K)	$T = \frac{\theta}{k}$	(6.54)
Density ($\mathrm{kg/m^3}$)	$\rho = \frac{m_0}{v}$	(6.62)
Isobaric specific heat ($\mathrm{J\,K^{-1}\,kg^{-1}}$)	$c_p = \left(1 + \frac{\beta}{2}\right)R = \left(1 + \frac{\beta}{2}\right)\frac{k}{m_0}$	(6.76)
Isochoric specific heat ($\mathrm{J\,K^{-1}\,kg^{-1}}$)	$c_v = \frac{\beta}{2}R = \frac{\beta}{2}\frac{k}{m_0}$	(6.76)

Digression 6.E: Specific heat of atmospheric gases

Using our stochastic framework, which is based on the principle of maximum entropy, and knowing the molecule structure of a specific gas, (hence its degrees of freedom β), and the molecular mass m_0, we can readily infer the macroscopic thermodynamic properties defined in section 6.10. Table 6.3 provides the results of such calculations for the gases most abundant in Earth's atmosphere. It also compares these theoretical values with experimental values, whose deviation from the theoretical ones is small.

Table 6.3 Gases appearing in Earth's atmosphere in proportions (mole fraction) > 0.1%, and their characteristics as derived by our stochastic framework. Conventional experimental values for the isobaric specific heat are also given in the table and compared with theoretical values. (Trace gases not included in the table constitute <0.03% altogether.)

Gas	Structure	Mole fraction (%)*	β	M_0, kg/kmol	R, $\mathrm{J\,K^{-1}kg^{-1}}$	c_v, $\mathrm{J\,K^{-1}kg^{-1}}$	c_p, $\mathrm{J\,K^{-1}kg^{-1}}$	Experimental c_p,† $\mathrm{J\,K^{-1}kg^{-1}}$	% deviation
Nitrogen, N_2	Diatomic	78.1 (77.8)	5	28.014	296.8	742.0	1038.8	1040	0.1
Oxygen, O_2	Diatomic	21.0 (20.9)	5	31.998	259.8	649.6	909.5	918	0.9
Argon, Ar	Monatomic	0.9 (0.9)	3	39.950	208.1	312.2	520.3	522	0.3
Water vapour, H_2O	Triatomic, nonlinear	0 (0.4)‡	6	18.015	461.5	1384.6	1846.1	1884	2.0

* The percentages without parentheses are for dry air and those in parentheses for wet air for the entire atmosphere.
† Different experimental values are given from different experiments in different texts. The value given here for water vapour was taken from Wagner and Pruss (2002) for the triple point of water.
‡ It typically varies to 0–3%, averaging at 0.4%

Digression 6.F: On the absence of hydrogen in the Earth's atmosphere

In the composition of Earth's atmosphere given in Table 6.3, nitrogen, oxygen and argon are present but hydrogen is absent. However, in other planets farther away from the Sun, hydrogen is present. It is easy to understand why this happens. We have seen that entropy maximization yields that the kinetic energy is equally distributed among different molecules. Hence, hydrogen (H_2), which has molecular mass 2 kg/kmol, much lower than oxygen and nitrogen, moves on average faster (~4 times) and escapes to space, while nitrogen, oxygen and argon do not reach the escape velocity. Planets farther from the Sun have a lower temperature, which is proportional to the kinetic energy, and thus hydrogen cannot reach the escape velocity.

6.11 Some classical concepts: heat, work and enthalpy

In classical thermodynamics, *heat* (Q) is the energy transfer across a boundary associated with microscopic motion (internal energy) due to a temperature difference. *Work* (W) is the energy transfer resulting from a force causing displacement at a macroscopic level.

While the internal energy of a gas in a motionless container is kinetic energy, the expected value of the velocity of a molecule is zero and thus macroscopically it cannot produce work. In contrast, the forces exerted on the walls of the container by the colliding molecules are all of the same direction and give a resultant force (pressure times area) whose expected value is not zero. Macroscopically, this can produce work. This is illustrated in Figure 6.3.

By virtue of energy conservation, when a differential heat δQ is supplied to a system it will equal the increase in internal energy of the system $\mathrm{d}E$ plus the work done by the system, δW. The delta notation in Q and W expresses the fact that these are inexact differentials, meaning that these quantities are not state functions, but path functions. Internal energy is an exact differential, so that when the system moves from state 1 to state 2 the change is $\int \mathrm{d}E = E_2 - E_1$. On the contrary, a similar equation does not hold for Q and W. Moreover, there is no meaning in writing Q_1 etc. We can only write $\int \delta Q = \Delta Q_{1\to 2}$.

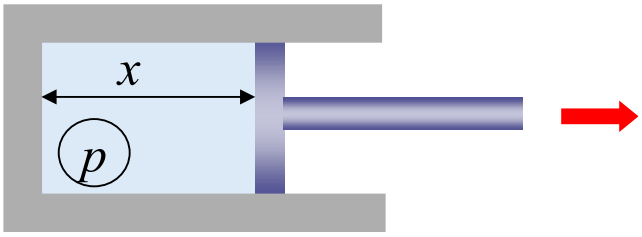

Figure 6.3 Explanation sketch showing a piston moving under the influence of the inside gas pressure p (the external pressure is taken zero).

Mathematically, we can write the energy conservation for the system of Figure 6.3 as

$$\delta Q = \mathrm{d}E + \delta W \tag{6.78}$$

If the piston whose area is A is moved by $\mathrm{d}x$, then the work produced is $\delta W = pA\mathrm{d}x = p\mathrm{d}V$, or per particle $\delta w = p\ A\ \mathrm{d}x/N = p\ \mathrm{d}v$. Hence we can write

$$\delta Q = \mathrm{d}E + p\mathrm{d}V \tag{6.79}$$

or per particle

$$\delta q = \mathrm{d}\varepsilon + p\,\mathrm{d}v \tag{6.80}$$

These equations hold true for any shape of the container (as already seen in section 6.9) and any macroscopic motion of parts, because the force due to pressure is always perpendicular to the surface. Now, because in the change we examine no leak or addition of particles occurs, equation (6.73) can be simplified to

$$\theta \mathrm{d}\Phi^* = \mathrm{d}E + p\mathrm{d}V \tag{6.81}$$

By comparing equations (6.79) and (6.81) we conclude that $\delta Q = \theta \mathrm{d}\Phi^*$, or

$$\mathrm{d}\Phi^* = \frac{\delta Q}{\theta} \tag{6.82}$$

In classical thermodynamics, this serves as the definition of entropy.

It should be stressed that equation (6.82) is valid only for reversible processes, in which the entropy is already maximized. As a counterexample, we consider a container, in which the molecules have temperatures θ_1 and θ_2 in the left and right half respectively. Without adding heat ($\Delta Q = 0$) the entropy will increase as the temperatures will become equal at the equilibrium state, in which the entropy will be maximized (see section 6.14), so that $\Delta\Phi^* > 0$. Hence, we rewrite equation (6.82) as

$$\mathrm{d}\Phi^* \geq \frac{\delta Q}{\theta} \tag{6.83}$$

and clarify that equality holds for reversible processes. A process with zero heat transfer ($\Delta Q = 0$) is called an *adiabatic* process. If it is reversible, then equation (6.82) holds and we have $\Delta\Phi^* = 0$. Such a process is called *isentropic* or *reversible adiabatic*.

Another useful concept expressing the total heat content of a system is the enthalpy defined as (in natural units, per molecule)

$$\eta \coloneqq \varepsilon + pv \tag{6.84}$$

which by virtue of the law of ideal gases can be written as

$$\eta \coloneqq \varepsilon + \theta = \left(\frac{\beta}{2} + 1\right)\theta \tag{6.85}$$

Converting to classical units, we find that the enthalpy per unit mass is

$$H = c_p T \tag{6.86}$$

Enthalpy, like internal energy, pressure, temperature and entropy, is an exact differential. To find heat content changes we just multiply enthalpy changes by the mass. We will use enthalpy to this aim in section 6.24 and beyond.

6.12 Gas mixtures

While in a laboratory we may make an experiment with a single gas, in natural systems (e.g., the atmosphere, see Digression 6.E) we usually have mixtures of gases. Up to now, we have seen that it is easy to deal with mixtures of degrees of freedom (translational, rotational, etc.), in which the constants that relate them to energy (mass, rotational

inertia, etc.) can be different in different degrees of freedom. As we have seen, the principle of maximum entropy implies that the internal energy is equally distributed among the molecules, as well as among the degrees of freedom of a molecule. We expect a similar behaviour with mixtures of gases, which then can be dealt with as single gases with appropriately specified constants, depending on the proportions of numbers of molecules in the mixture.

We consider a mixture of gas A with comprising N_A molecules with β_A degrees of freedom with a gas B comprising N_B molecules with β_B degrees of freedom. The total number of molecules N is

$$N = N_\mathrm{A} + N_\mathrm{B} \tag{6.87}$$

The two gases share the common volume V and have total internal energy

$$N_\mathrm{A}\varepsilon_\mathrm{A} + N_\mathrm{B}\varepsilon_\mathrm{B} = E \tag{6.88}$$

From equation (6.25) we find that the total entropy (the sum of the two partial ones) is

$$\Phi = \frac{\beta_\mathrm{A}N_\mathrm{A}}{2}\ln\frac{\varepsilon_\mathrm{A}}{\varepsilon_\mathrm{A}^*} + \frac{\beta_\mathrm{B}N_\mathrm{B}}{2}\ln\frac{\varepsilon_\mathrm{B}}{\varepsilon_\mathrm{B}^*} + N_\mathrm{A}\ln\frac{V}{v_\mathrm{A}^*} + N_\mathrm{B}\ln\frac{V}{v_\mathrm{B}^*} \tag{6.89}$$

Maximization of entropy for varying $\varepsilon_\mathrm{A}, \varepsilon_\mathrm{B}$ subject to the constraint (6.88)

$$\frac{\varepsilon_\mathrm{A}}{2\beta_\mathrm{A}} = \frac{\varepsilon_\mathrm{B}}{2\beta_\mathrm{B}} = \theta \tag{6.90}$$

Thus, the total entropy can be written as

$$\begin{aligned}\Phi = \left(\frac{\beta_\mathrm{A}N_\mathrm{A}}{2} + \frac{\beta_\mathrm{B}N_\mathrm{B}}{2}\right)\ln\theta - \left(\frac{\beta_\mathrm{A}N_\mathrm{A}}{2}\ln\theta_\mathrm{A}^* + \frac{\beta_\mathrm{B}N_\mathrm{B}}{2}\ln\theta_\mathrm{B}^*\right)\\ + N_\mathrm{A}\ln V + N_\mathrm{B}\ln V - (N_\mathrm{A}\ln v_\mathrm{A}^* + N_\mathrm{B}\ln v_\mathrm{B}^*)\end{aligned} \tag{6.91}$$

and after algebraic manipulations it becomes

$$\Phi = \frac{\beta N}{2}\ln\frac{\theta}{\theta^*} + N\ln\frac{V}{v^*} \tag{6.92}$$

where

$$\beta := \frac{N_\mathrm{A}\beta_\mathrm{A} + N_\mathrm{B}\beta_\mathrm{B}}{N} \tag{6.93}$$

is the average number of degrees of freedom per molecule, and

$$\ln\theta^* := \frac{\beta_\mathrm{A}N_\mathrm{A}}{\beta N}\ln\theta_\mathrm{A}^* + \frac{\beta_\mathrm{B}N_\mathrm{B}}{\beta N}\ln\theta_\mathrm{B}^*, \qquad \ln v^* := \frac{N_\mathrm{A}}{N}\ln v_\mathrm{A}^* + \frac{N_\mathrm{B}}{N}\ln v_\mathrm{B}^* \tag{6.94}$$

The entropy Φ is equivalent to that of a single gas with β degrees of freedom per molecule.

In summary, equation (6.92) expressing the (maximized) entropy of the mixture does not differ from that of a single gas (equation (6.25)), if we appropriately calculate the characteristic constant of the mixture, c, as above. All quantities of the mixture in terms of totals and means are given in Table 6.4.

Table 6.4 Characteristic quantities of a mixture of two gases A and B in terms of totals and averages.

	Total	Mean
Number of particles	$N = N_\mathrm{A} + N_\mathrm{B}$	
Mass	$M = N_\mathrm{A} m_{0_\mathrm{A}} + N_\mathrm{B} m_{0_\mathrm{b}}$	$m_0 = \left(N_\mathrm{A} m_{0_\mathrm{A}} + N_\mathrm{B} m_{0_\mathrm{b}}\right)/N$
Energy	$E = E_\mathrm{A} + E_\mathrm{B} = N_\mathrm{A}\varepsilon_\mathrm{A} + N_\mathrm{B}\varepsilon_\mathrm{B}$	$\varepsilon = E/N$
Degrees of freedom	$N\beta = N_\mathrm{A}\beta_\mathrm{A} + N_\mathrm{B}\beta_\mathrm{B}$	$\beta = (N_\mathrm{A}\beta_\mathrm{A} + N_\mathrm{B}\beta_\mathrm{B})/N$
Reference temperature		$\ln\theta^* = \frac{\beta_\mathrm{A} N_\mathrm{A}}{\beta N}\ln\theta_\mathrm{A}^* + \frac{\beta_\mathrm{B} N_\mathrm{B}}{\beta N}\ln\theta_\mathrm{B}^*$
Reference volume		$\ln v^* = \frac{N_\mathrm{A}}{N}\ln v_\mathrm{A}^* + \frac{N_\mathrm{B}}{N}\ln v_\mathrm{B}^*$

The interpretation of the above results is that a mixture of gases, statistically behaves like a hypothetical single gas with molecular mass, energy per particle and degrees of freedom equal to the corresponding averages in the mixture of gases (for an illustration see Digression 6.G). Noticeably, the number of degrees of freedom in a mixture is no longer an integer, but a real number. Each of the gases exert a partial pressure determined by the ideal gas law, and the sum of all constitutes the total (e.g. atmospheric) pressure. Nonetheless, a mixture is not identical to a single gas as illustrated in Digression 6.F.

Digression 6.G: Specific heat of Earth's atmosphere

Based on the characteristics of the gases constituting the Earth's atmosphere given in Table 6.3 and using the relations of Table 6.4, we find that the degrees of freedom per molecule of the mixture are $\beta = 4.98$. From equation (6.61), it becomes evident that the microscopic relation involving m_0 in Table 6.4 extends to the macroscopic molecular mass M_0. Hence, for this mixture we find $M_0 = 28.96$ kg/kmol, and $R = 8314.463/28.96 = 287.1$ J K^{-1} kg^{-1}. Thus, $c_p = (1 + 4.98/2) \times 287.1 = 1002.2$ J K^{-1} kg^{-1}, while the experimental value is $c_p = 1004$ J K^{-1} kg^{-1}. The deviation is only 0.18%.

6.13 Closed interaction of two bodies in contact

We consider a container isolated from the environment and separated into two parts by a diaphragm, as seen in Figure 6.4. Part A with volume V_A contains N_A molecules with average internal energy ε_A and β_A degrees of freedom; part B with volume V_B contains N_B molecules with average energy ε_B and β_B degrees of freedom.

The device does not allow exchange of mass between the two parts, or change of volumes and thus the quantities V_A, N_A, β_A, V_B, N_B, β_B are constant. The diaphragm allows exchange of energy but not mass. Therefore, whatever the initial energies per particle, $\varepsilon_\mathrm{A}^{(0)}$ and $\varepsilon_\mathrm{B}^{(0)}$, be, the system will reach an *equilibrium* state in which these energies will change so that the entropy is maximized. Yet the total energy should be conserved, i.e.:

$$N_\mathrm{A}\varepsilon_\mathrm{A} + N_\mathrm{B}\varepsilon_\mathrm{B} = E = N_\mathrm{A}\varepsilon_\mathrm{A}^{(0)} + N_\mathrm{B}\varepsilon_\mathrm{B}^{(0)} \tag{6.95}$$

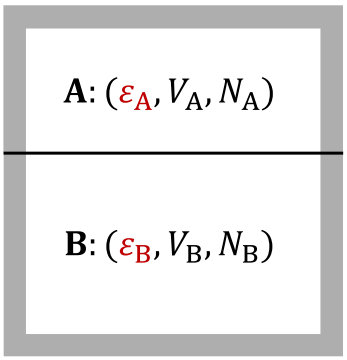

Figure 6.4 Explanation sketch for the closed interaction of two bodies in contact, showing the characteristics in each part of the system, where those typeset in black are known and those in red are determined by entropy maximization.

The entropies of the two systems are:

$$\Phi_A = \frac{\beta_A N_A}{2} \ln \frac{\varepsilon_A}{\varepsilon_A^*} + N_A \ln \frac{V}{v_A^*}, \qquad \Phi_B = \frac{\beta_B N_B}{2} \ln \frac{\varepsilon_B}{\varepsilon_B^*} + N_B \ln \frac{V}{v_B^*} \tag{6.96}$$

Since the parts are isolated, the total entropy will be the sum of the two partial ones:

$$\Phi = \frac{\beta_A N_A}{2} \ln \frac{\varepsilon_A}{\varepsilon_A^*} + \frac{\beta_B N_B}{2} \ln \frac{\varepsilon_B}{\varepsilon_B^*} + N_A \ln \frac{V}{v_A^*} + N_B \ln \frac{V}{v_B^*} \tag{6.97}$$

In the final state Φ will be at maximum. Its maximization with respect to ε_A and ε_B subject to constraint (6.95), results in

$$\frac{\varepsilon_A}{\beta_A} = \frac{\varepsilon_B}{\beta_B} \tag{6.98}$$

and by virtue of (6.52),

$$\theta_A = \theta_B \tag{6.99}$$

In other words, the temperatures of two systems brought in contact, like the two parts of the above compound system, will become equal. This reflects the Zeroth Law of thermodynamics (see section 6.26). It is readily understood that, since Φ^* has been maximized,

$$\Phi - \Phi^{(0)} = \Phi^* - \Phi^{*(0)} \geq 0 \tag{6.100}$$

where the equality sign applies to the case that the initial entropy was already at maximum. The spontaneous increase of entropy, due to entropy maximization, reflects the Second Law of thermodynamics (see section 6.26 again).

6.14 Open interaction of two systems in contact

In an *open* interaction, the two systems are allowed to exchange both energy and mass. This is assured by a hole seen in Figure 6.5, which allows molecules to move from one system to the other. Here, mass conservation should also be considered as a constraint, in addition to the energy constraint. The possibility of mass exchange allows us to simplify

the derivation by considering the entropy of just a single molecule, because, as we saw in section 6.7, in this case (and contrary to the case of closed interaction examined in section 6.13) we have unique intensive quantities, representative for both systems. Since we study intensive quantities, we formulate the problem in terms of the probabilities that the particle is in system A or B, namely P_A and P_B, respectively, instead on numbers of particles.

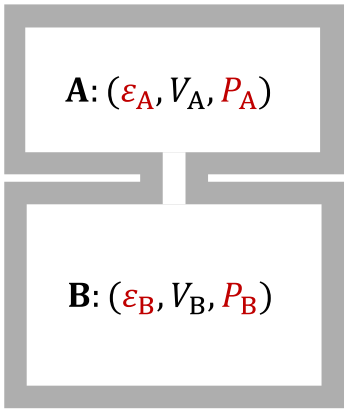

Figure 6.5 Explanation sketch for the open interaction of two systems in contact, showing the characteristics in each system, where those typeset in black are known and those in red are determined by entropy maximization.

For one molecule, the mass conservation is expressed as

$$P_A + P_B = 1 \tag{6.101}$$

and the energy conservation as

$$P_A \varepsilon_A + P_B \varepsilon_B = \varepsilon = E^{(0)}/N \tag{6.102}$$

where $E^{(0)}$ is the initial internal energy and N the number of particles. Because of mass exchange, in both systems the molecule characteristics will be the same in the two systems, i.e. $\beta_A = \beta_B = \beta$, $\varepsilon_A^* = \varepsilon_B^* = \varepsilon^*$ and $v_A = v_B = v$. Based on equation (6.22), the entropies per particle will be

$$\varphi_A = \frac{\beta}{2} \ln \frac{\varepsilon_A}{\varepsilon^*} + \ln \frac{V_A}{v^*}, \qquad \varphi_B = \frac{\beta}{2} \ln \frac{\varepsilon_B}{\varepsilon^*} + \ln \frac{V_B}{v^*} \tag{6.103}$$

and from equation (6.39) the entropy per particle of the compound system is

$$\begin{aligned} \varphi &= P_A \varphi_A + P_B \varphi_B + \varphi_P \\ &= \frac{\beta P_A}{2} \ln \frac{\varepsilon_A}{\varepsilon^*} + \frac{\beta P_B}{2} \ln \frac{\varepsilon_B}{\varepsilon^*} + P_A \ln \frac{V_A}{v^*} + P_B \ln \frac{V_B}{v^*} - P_A \ln P_A - P_B \ln P_B \end{aligned} \tag{6.104}$$

The Lagrangian to maximize is

$$\begin{aligned} L &= \varphi - \lambda_1 (P_A + P_B - 1) - \lambda_2 (P_A \varepsilon_A + P_B \varepsilon_B - \varepsilon) = \\ &= \frac{\beta P_A}{2} \ln \frac{\varepsilon_A}{\varepsilon_A^*} + \frac{\beta P_B}{2} \ln \frac{\varepsilon_B}{\varepsilon_B^*} + P_A \ln \frac{V_A}{v^*} + P_B \ln \frac{V_B}{v^*} - P_A \ln P_A - P_B \ln P_B \\ &\qquad - \lambda_1 (P_A + P_B - 1) - \lambda_2 (P_A \varepsilon_A + P_B \varepsilon_B - \varepsilon) \end{aligned} \tag{6.105}$$

Taking derivatives with respect to ε_A and ε_B we find

$$\frac{\partial L}{\partial \varepsilon_{\mathrm{A}}} = \frac{\beta P_{\mathrm{A}}}{2\varepsilon_{\mathrm{A}}} - \lambda_2 P_{\mathrm{A}}, \qquad \frac{\partial L}{\partial \varepsilon_{\mathrm{B}}} = \frac{\beta P_{\mathrm{B}}}{2\varepsilon_{\mathrm{B}}} - \lambda_2 P_{\mathrm{B}} \tag{6.106}$$

and equating them to zero we readily obtain

$$\varepsilon_{\mathrm{A}} = \varepsilon_{\mathrm{B}} = \varepsilon \tag{6.107}$$

Which, by virtue of equation (6.52), results in equality of temperatures:

$$\theta_{\mathrm{A}} = \theta_{\mathrm{B}} = \theta = \frac{2\varepsilon}{\beta} \tag{6.108}$$

Taking derivatives with respect to P_{A} and P_{B} we find

$$\begin{aligned} \frac{\partial L}{\partial P_{\mathrm{A}}} &= -1 - \lambda_1 - \lambda_2 \varepsilon_{\mathrm{A}} + \frac{\beta}{2} \ln \frac{\varepsilon_{\mathrm{A}}}{\varepsilon^*} + \ln \frac{V_{\mathrm{A}}}{v^*} - \ln P_{\mathrm{A}} \ , \\ \frac{\partial L}{\partial P_{\mathrm{B}}} &= -1 - \lambda_1 - \lambda_2 \varepsilon_{\mathrm{B}} + \frac{\beta}{2} \ln \frac{\varepsilon_{\mathrm{B}}}{\varepsilon^*} + \ln \frac{V_{\mathrm{A}}}{v^*} - \ln P_{\mathrm{B}} \end{aligned} \tag{6.109}$$

Subtracting these two, equating to zero, and utilizing $\varepsilon_{\mathrm{A}} = \varepsilon_{\mathrm{B}}$ we find

$$-\ln P_{\mathrm{A}} + \ln P_{\mathrm{B}} + \ln V_{\mathrm{A}} - \ln V_{\mathrm{B}} = 0 \tag{6.110}$$

which yields

$$P_{\mathrm{A}} = \frac{V_{\mathrm{A}}}{V}, \qquad P_{\mathrm{B}} = \frac{V_{\mathrm{B}}}{V} \tag{6.111}$$

where $V \coloneqq V_{\mathrm{A}} + V_{\mathrm{B}}$.

The maximized entropies per particle are then obtained as

$$\varphi_{\mathrm{A}} = \frac{\beta}{2} \ln \frac{\varepsilon}{\varepsilon^*} + \ln \frac{V_{\mathrm{A}}}{v^*}, \qquad \varphi_{\mathrm{B}} = \frac{\beta}{2} \ln \frac{\varepsilon}{\varepsilon^*} + \ln \frac{V_{\mathrm{B}}}{v^*}, \qquad \varphi = \frac{\beta}{2} \ln \frac{\varepsilon}{\varepsilon^*} + \ln \frac{V}{v^*} \tag{6.112}$$

and the standardized entropies are

$$\varphi_{\mathrm{A}}^* = \varphi_{\mathrm{B}}^* = \varphi^* = \frac{\beta}{2} \ln \frac{\varepsilon}{\varepsilon^*} + \ln \frac{v}{v^*} \tag{6.113}$$

This is justified by the fact that, as a result of the proportionality of probability, and hence the number of particles, the volume per particle is the same, $v = V/N$, everywhere. As a result, the density $\rho = m_0/v$ will also be the same and, by virtue of the law of the ideal gases, the pressure will also be equal everywhere.

In conclusion, while a closed interaction results in equal temperatures of the two parts, an open interaction results in equality of all intensive quantities: temperature, volume per particle, density, pressure and standardized entropy per particle. This expresses the macroscopic ultimate simplicity of a microscopically complex system. The macroscopic equality is accompanied by the ultimate microscopic diversity, so that the microscopic uncertainty be maximal. This is implied by entropy maximization. If the initial conditions differ from the ultimate macroscopic equality, thus corresponding to smaller entropy, these conditions will change. The resulting spontaneous increase of entropy, due to entropy maximization, again reflects the Second Law of thermodynamics.

6.15 Open interaction of two systems under gravitation

We will now reexamine the problem of section 6.14, but also taking gravitation into account. We now assume that the upper container A is far away from B, at an elevation z, while A is at zero elevation, as seen in Figure 6.6.

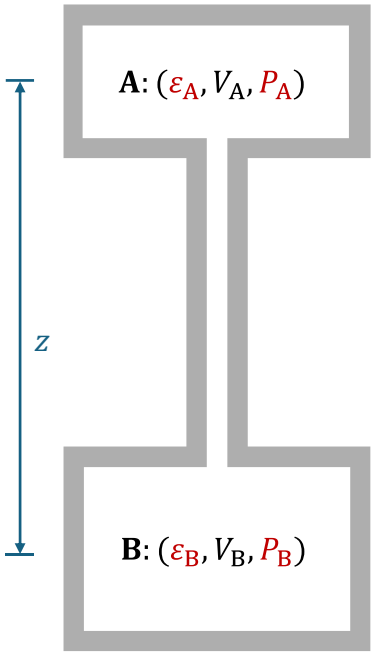

Figure 6.6 Explanation sketch for the open interaction of two systems affected by gravity, because of their vertical distance of z, showing the characteristics in each system, where those typeset in black are known and those in red are determined by entropy maximization.

In this case the energy conservation equation (6.102) should be modified to include the potential energy due to gravitation of the molecules that are found in A, i.e.

$$P_A(\varepsilon_A + m_0 g z) + P_B \varepsilon_B = \varepsilon_{tot} \tag{6.114}$$

where ε_{tot} is the total energy per particle. The expressions of the entropies of the two systems do not change and therefore the entropy per particle of the compound system is again given by equation (6.104). Thus, the quantity to maximize is

$$\begin{aligned} L = \varphi - \lambda_1(P_A + P_B - 1) - \lambda_2(P_A(\varepsilon_A + m_0 g z) + P_B \varepsilon_B - \varepsilon_{tot}) = \\ = \frac{\beta P_A}{2} \ln \frac{\varepsilon_A}{\varepsilon^*} + \frac{\beta P_B}{2} \ln \frac{\varepsilon_B}{\varepsilon^*} + P_A \ln \frac{V_A}{v^*} + P_B \ln \frac{V_B}{v^*} - P_A \ln P_A - P_B \ln P_B \\ - \lambda_1(P_A + P_B - 1) - \lambda_2(P_A(\varepsilon_A + m_0 g z) + P_B \varepsilon_B - \varepsilon_{tot}) \end{aligned} \tag{6.115}$$

Taking derivatives with respect to ε_A and ε_B we find

$$\frac{\partial L}{\partial \varepsilon_A} = \frac{\beta P_A}{2\varepsilon_A} - \lambda_2 P_A, \qquad \frac{\partial L}{\partial \varepsilon_B} = \frac{\beta P_B}{2\varepsilon_B} - \lambda_2 P_B \tag{6.116}$$

and equating them to zero we readily find again

$$\varepsilon_A = \varepsilon_B = \varepsilon \tag{6.117}$$

This means that the internal energy and hence the temperature will be the same in the two systems. Gravitation does not affect the isothermal state of the compound system.

Taking derivatives with respect to P_A and P_B we find

$$\frac{\partial L}{\partial P_\mathrm{A}} = -1 - \lambda_1 - \lambda_2(\varepsilon + m_0 g z) + \frac{\beta}{2}\ln\frac{\varepsilon}{\varepsilon^*} + \ln\frac{V_\mathrm{A}}{v^*} - \ln P_\mathrm{A}\ ,$$
$$\frac{\partial L}{\partial P_\mathrm{B}} = -1 - \lambda_1 - \lambda_2\varepsilon_\mathrm{B} + \frac{\beta}{2}\ln\frac{\varepsilon}{\varepsilon^*} + \ln\frac{V_\mathrm{B}}{v^*} - \ln P_\mathrm{B} \tag{6.118}$$

Subtracting these two, equating to zero, and utilizing $\varepsilon_\mathrm{A} = \varepsilon_\mathrm{B}$ we find

$$-\frac{\beta}{2}\frac{m_0 g z}{\varepsilon} - \ln P_\mathrm{A} + \ln P_\mathrm{B} + \ln V_\mathrm{A} - \ln V_\mathrm{B} = 0 \tag{6.119}$$

where we have made a substitution for λ_2 from equation (6.116). Exponentiating the latter equation, we obtain

$$\exp\left(-\frac{\beta g\, m_0\, z}{2\,\varepsilon}\right)\frac{P_\mathrm{B}V_\mathrm{A}}{P_\mathrm{A}V_\mathrm{B}} = 1 \tag{6.120}$$

By expressing potential energy as a fraction of the total, i.e., $m_0 g z = a\,\varepsilon_\mathrm{tot}$ (where a is dimensionless), and substituting $1 - P_\mathrm{A}$ for P_B we find the following implicit expression whose solution gives P_A:

$$\exp\left(-\frac{\beta}{2}\frac{a}{1 - aP_A}\right)(1 - P_\mathrm{A})V_\mathrm{A} = P_\mathrm{A}V_\mathrm{B} \tag{6.121}$$

For large z (or a) the probability P_A is small. Then the equation becomes

$$\exp\left(-\frac{\beta a}{2}\right)(1 - P_\mathrm{A})V_\mathrm{A} \approx P_\mathrm{A}V_\mathrm{B} \tag{6.122}$$

and admits an explicit solution

$$P_\mathrm{A} \approx \frac{V_\mathrm{A}}{V_\mathrm{A} + \exp\left(\frac{\beta a}{2}\right)V_\mathrm{B}} \tag{6.123}$$

which suggests a near-exponential decay with increasing a.

Accurate numerical solutions for a range of a are shown in Figure 6.7, from which we can make the following observations:

- For $z = 0$ the probability for either system is 0.5 and the entire energy is internal. This switches to the case we examined in section 6.14.
- For large z (or a) the probability P_A drops exponentially.
- The more degrees of freedom the gas has, the higher the rate of exponential decrease is.
- For varying z, the fraction of the internal to total energy, which equals $1 - aP_A$, is initially decreasing with z (or a) up to a certain level of a. This is explained by the fact that the potential energy increases with z. However, beyond that level this fraction increases again as a result of the decreased probability that a molecule stays at the upper system.

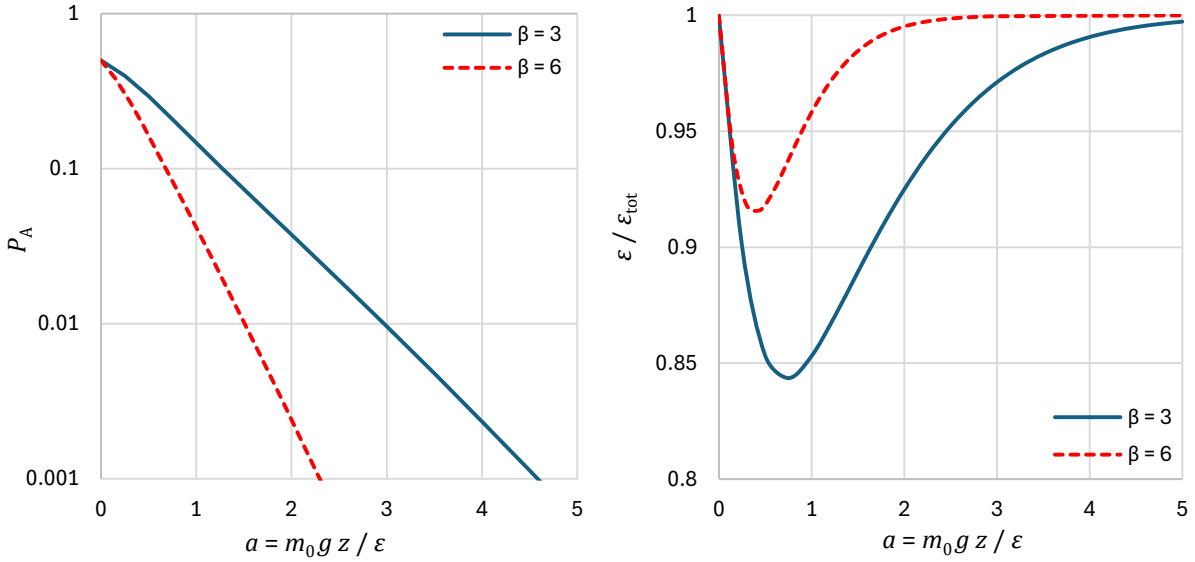

Figure 6.7 Variation of characteristic quantities in open interaction of two systems affected by gravitation, as functions of the fraction of potential to total energy, $a = m_0 g z/\varepsilon_{\text{tot}}$ for a gas with 3 or 6 degrees of freedom: (**left**) probability that a particle is found in the upper system A, located at vertical distance z above B; (**right**) fraction of internal to total energy per particle.

6.16 Equilibrium state of the air column

The system we studied in section 6.15 is not a real-world system but an idealized one, yet it has helped us to develop intuition and know what to expect in the case of the entire air column being at the equilibrium state. For example, we may expect that it is isothermal. We will now reexamine the problem of section 6.14, but for the entire atmospheric column and taking gravitation into account. While in section 6.15 we examined two containers with two probabilities, now we have an infinite number thereof, expressed through a probability densify function $f(z)$, where z is the altitude varying for zero to infinity. The partition entropy, which in section 6.14 was the sum of two terms, now vecomes the integral

$$\varphi_P = -\int_0^\infty \ln f(z)\, f(z)\, \mathrm{d}z \tag{6.124}$$

Assuming a constant reference volume V at each z, the conditional entropy for altitude z is

$$\varphi(z) = \frac{\beta}{2} \ln \frac{\varepsilon(z)}{\varepsilon^*} + \ln \frac{V}{v^*} \tag{6.125}$$

Hence, the total entropy across the column is

$$\varphi = \int_0^\infty \varphi(z) f(z)\, \mathrm{d}z + \varphi_P \tag{6.126}$$

The mass conservation equation is

$$\int_0^\infty f(z)\,\mathrm{d}z = 1 \tag{6.127}$$

and the energy conservation equation is

$$\int_0^\infty (\varepsilon(z) + m_0 g\, z) f(z)\,\mathrm{d}z = \varepsilon_{\mathrm{tot}} \tag{6.128}$$

where $\varepsilon(z)$ is the internal energy at altitude z and $\varepsilon_{\mathrm{tot}}$ is the sum of internal and potential energy.

We form the Lagrangian by neglecting the constant terms in entropy as

$$\begin{aligned} L = \frac{\beta}{2}\int_0^\infty \ln \varepsilon(z)\, f(z)\,\mathrm{d}z - \int_0^\infty \ln f(z)\, f(z)\,\mathrm{d}z - \lambda_1\left(\int_0^\infty f(z)\,\mathrm{d}z - 1\right) \\ - \lambda_2\left(\int_0^\infty (\varepsilon(z) + m_0 g\, z) f(z)\,\mathrm{d}z - \varepsilon_{\mathrm{tot}}\right) \end{aligned} \tag{6.129}$$

Taking derivatives with respect to $\varepsilon(z)$ we find

$$\frac{\partial L}{\partial \varepsilon(z)} = \frac{\beta f(z)}{2\varepsilon(z)} - \lambda_2 f(z) \tag{6.130}$$

and equating it to zero we readily find

$$\varepsilon(z) = \frac{\beta}{2\lambda_2} =: \varepsilon \tag{6.131}$$

where ε is constant, independent of z. This means that the temperature will also be constant, and the atmospheric column will be *isothermal*.

Taking derivatives with respect to $f(z)$ we find

$$\frac{\partial L}{\partial f(z)} = -1 - \lambda_1 - \lambda_2(\varepsilon + m_0 g z) + \frac{\beta}{2}\ln \varepsilon - \ln f(z) \tag{6.132}$$

Solving this for $f(z)$, making a substitution for λ_2 from equation (6.131) and calculating λ_1 from (6.127) we obtain an exponential distribution

$$f(z) = \frac{\beta g m_0}{2\varepsilon} \exp\left(-\frac{\beta g m_0 z}{2\varepsilon}\right) \tag{6.133}$$

On the other hand, from (6.128) finally we find

$$\frac{\beta + 2}{\beta}\varepsilon = \varepsilon_{\mathrm{tot}} \tag{6.134}$$

which suggests that the distribution of the total energy over the entire column into internal, ε, and potential due to gravitation, ε_{G}, is:

$$\varepsilon = \frac{\beta}{\beta+2}\varepsilon_{\text{tot}}, \qquad \varepsilon_{\text{G}} = \frac{2}{\beta+2}\varepsilon_{\text{tot}} \tag{6.135}$$

In other words, in terms of energy distribution, gravitation is equivalent to two additional degrees freedom.

We can thus write equation (6.133) in the following two forms, by means of either the total energy ε_{tot} or the natural temperature, $\theta = 2\varepsilon/\beta$ as:

$$f(z) = \frac{(\beta+2)gm_0}{2\varepsilon_{\text{tot}}}\exp\left(-\frac{(\beta+2)gm_0 z}{2\varepsilon_{\text{tot}}}\right), \qquad f(z) = \frac{gm_0}{\theta}\exp\left(-\frac{gm_0 z}{\theta}\right) \tag{6.136}$$

Apparently, the air density $\rho(z)$ will be proportional to $f(z)$. Considering the ideal gas law, $pv = \theta$ (equation (6.59)), and the relationship between density ρ and volume per particle, v, i. e. $\rho = m_0/v$ (equation (6.62)), v and p will also be exponential functions of z. Specifically,

$$\rho(z) = \rho_0\exp\left(-\frac{gm_0 z}{\theta}\right), \quad v(z) = v_0\exp\left(\frac{gm_0 z}{\theta}\right), \quad p(z) = p_0\exp\left(-\frac{gm_0 z}{\theta}\right) \tag{6.137}$$

where p_0 is the pressure at $z = 0$, $v_0 \coloneqq \theta/p_0$ and $\rho_0 \coloneqq m_0 p_0/\theta$. Taking the derivative of pressure with respect to altitude, we find:

$$\frac{\mathrm{d}p}{\mathrm{d}z} = -\frac{gm_0}{\theta}p(z) = -g\rho(z) \tag{6.138}$$

In other words, the hydrostatic law of pressure (equation (6.67)) is confirmed.

The entropy per particle for specified volume V will be constant, independent of z:

$$\varphi(z) = \frac{\beta}{2}\ln\frac{\theta}{\theta^*} + \ln\frac{V}{v^*} \tag{6.139}$$

The standardized entropy per particle is

$$\varphi_z^*(z) = \frac{\beta}{2}\ln\frac{\theta}{\theta^*} + \ln\frac{v(z)}{v^*} = \varphi_0^* + \ln\frac{v(z)}{v_0} = \varphi_0^* + \frac{gm_0 z}{\theta}, \quad \varphi_0^* \coloneqq \frac{\beta}{2}\ln\frac{\theta}{\theta^*} + \ln\frac{v_0}{v^*} \tag{6.140}$$

This is a linearly increasing function of z. To be consistent with the literature, we do not call the resulting profile *isentropic*, despite the fact that $\varphi(z)$ is constant. Rather, we call a state *isentropic* when the standardized entropy, and not the actual, has constant entropy. As we have already explained, the standardized entropy, not the actual one, is an intensive property and therefore it is meaningful to make comparisons (e.g. if they are equal or which one is greater) in terms of intensive properties.

Summarizing, an air column at equilibrium is isothermal but gravitation does not allow it to be isobaric or isentropic. On the other hand, several quantities have exponential vertical profiles, which are the simplest possible extending up to infinity. The exceptions are the temperature and the actual entropy, which are constant, and the standardized entropy, which has a linear profile.

Different derivations of precisely the same equations have been given by Koutsoyiannis and Tsakalias (2025), who notably also confirmed the equation using simulations of the molecules' motion and collisions.

6.17 Non-equilibrium open interaction of two systems in contact

We assume two systems of same composition. According to what we have seen in previous sections, if they are brought in contact and are allowed to exchange energy (closed interaction), then they reach a state of equal energy per particle, or equal temperature. If they are also allowed to exchange mass (open interaction), then they reach full uniformity, with equal temperature and volume per particle, thus practically forming a single system. These states correspond to equilibrium, in which entropy is maximized

However, if there are external effects, the two systems may not reach full uniformity and equilibrium state. Such cases are necessarily accompanied by externally imposed energy flow to counteract the heat flow from the hotter to the cooler body. Let us examine the case shown in Figure 6.8, where we pump heat from system A and transfer it to system B.

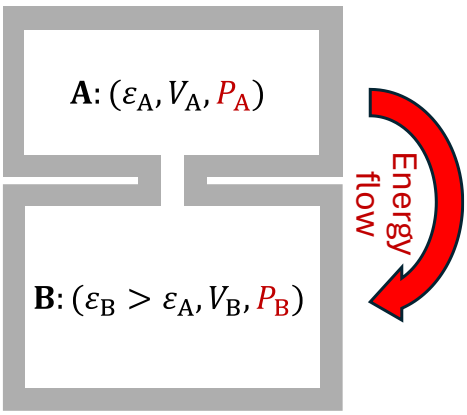

Figure 6.8 Explanation sketch for the non-equilibrium open interaction of two systems in contact, showing the characteristics in each system, where those typeset in black are known and those in red are determined by entropy maximization.

We assume that there is uniformity within each container A and B, but not for the compound system. System B is hotter than A (like in a heated room, beside one without heat), and we assume the energies per particle ε_A and $\varepsilon_B > \varepsilon_A$ are known. Unknown are the probabilities P_A and P_B that a particle is in A and B, respectively.

Since the systems are in close contact, we may invoke the momentum conservation, at the border between system A and B, i.e. the momentum directed from A to B in time δt, $\mathrm{E}\left[\underline{q}_{A\to B}(\delta t)\right]$, equals that from B to A, $\mathrm{E}\left[\underline{q}_{B\to A}(\delta t)\right]$. As explained in section 6.9, momentum is proportional to pressure and hence the pressure in the two systems will be the same:

$$p = \frac{\theta_A}{v_A} = \frac{\theta_B}{v_B} \tag{6.141}$$

Since $\theta_A = 2\varepsilon_A/\beta$ and $v_A = V_A/N_A = V_A/(P_A N)$, and likewise for B, we will have

$$\frac{\varepsilon_A P_A}{V_A} = \frac{\varepsilon_B P_B}{V_B} \tag{6.142}$$

and hence

$$P_{\mathrm{A}} = \frac{V_{\mathrm{A}}/\varepsilon_{\mathrm{A}}}{V_{\mathrm{A}}/\varepsilon_{\mathrm{A}} + V_{\mathrm{B}}/\varepsilon_{\mathrm{B}}}, \qquad P_{\mathrm{B}} = \frac{V_{\mathrm{B}}/\varepsilon_{\mathrm{B}}}{V_{\mathrm{A}}/\varepsilon_{\mathrm{A}} + V_{\mathrm{B}}/\varepsilon_{\mathrm{B}}} \tag{6.143}$$

The entropies per particle for each of A and B are given by equation (6.103) and are constant. The entropy per particle of the compound system is given by equation (6.104) after substitution of $P_{\mathrm{A}}, P_{\mathrm{B}}$ from (6.143). The standardized entropies per particle, which are intensive properties, are given by equation (6.68):

$$\varphi_{\mathrm{A}}^{*} = \left(1 + \frac{\beta}{2}\right) \ln \frac{\theta_{\mathrm{A}}}{\theta^{*}} - \ln \frac{p}{p^{*}}, \qquad \varphi_{\mathrm{B}}^{*} = \left(1 + \frac{\beta}{2}\right) \ln \frac{\theta_{\mathrm{A}}}{\theta^{*}} - \ln \frac{p}{p^{*}} \tag{6.144}$$

and hence their difference is

$$\varphi_{\mathrm{B}}^{*} - \varphi_{\mathrm{A}}^{*} = \left(1 + \frac{\beta}{2}\right) \ln \frac{\theta_{\mathrm{B}}}{\theta_{\mathrm{A}}} = \left(1 + \frac{\beta}{2}\right) \ln \frac{\varepsilon_{\mathrm{B}}}{\varepsilon_{\mathrm{A}}} > 0 \tag{6.145}$$

In brief, the standardized entropy per particle is greater in the system with the highest internal energy per particle. However, the relationship between probabilities is reversed, as they are inversely proportional to the internal energy per particle.

These characterize an *isobaric* state, where the pressure is constant everywhere. We note that in this case, while we used the principle of maximum entropy to derive the conditional energy per part, we did not use it for the compound system. The equality of pressure did not allow any degree of freedom for any quantity to be determined by maximization.

Apparently the system examined is not at equilibrium. If we leave the system to spontaneously evolve, it will reach at the equilibrium state of complete macroscopic uniformity, as described in section 6.14. In order to keep the system in steady state, as it appears in Figure 6.8, we should continually remove heat from A and supply it to B. Its rate q can be described by Fourier's law (Digression 3.E) which in this case takes the form:

$$q = \frac{\kappa A}{z} (\theta_{\mathrm{B}} - \theta_{\mathrm{A}}) \tag{6.146}$$

where κ denotes the thermal conductivity, A the area and z the length of the connector.

Since the system is at steady state, the total entropies $\Phi_{\mathrm{A}}^{*} = N_{\mathrm{A}}\varphi_{\mathrm{A}}^{*}, \Phi_{\mathrm{B}}^{*} = N_{\mathrm{B}}\varphi_{\mathrm{B}}^{*}$ should be constant in time ($\mathrm{d}\Phi_{\mathrm{A}}^{*}/\mathrm{d}t = \mathrm{d}\Phi_{\mathrm{B}}^{*}/\mathrm{d}t = 0$). Yet, entropy is continually generated in the system and exported, at an exactly equal rate, to the environment through the devices that keep the temperatures constant ($\theta_{\mathrm{A}}, \theta_{\mathrm{B}}$). According to equation (6.82) the entropies generated at each part and the compound system will be

$$\frac{\mathrm{d}\Phi_{\mathrm{A_G}}^{*}}{\mathrm{d}t} = \frac{q}{\theta_{\mathrm{A}}}, \qquad \frac{\mathrm{d}\Phi_{\mathrm{B_G}}^{*}}{\mathrm{d}t} = -\frac{q}{\theta_{\mathrm{B}}}, \qquad \frac{\mathrm{d}\Phi_{\mathrm{G}}^{*}}{\mathrm{d}t} = \frac{q}{\theta_{\mathrm{A}}} - \frac{q}{\theta_{\mathrm{B}}} = \frac{\kappa A}{z} \frac{(\theta_{\mathrm{B}} - \theta_{\mathrm{A}})^{2}}{\theta_{\mathrm{B}}\theta_{\mathrm{A}}} > 0 \tag{6.147}$$

This positive net entropy generation and export, while the system entropy remains constant, is the signature of a steady-state irreversible process. The generation and export of entropy is another case, additional to those of Chapter 5, where entropy, a purely probabilistic concept, behaves like a material concept.

6.18 Non-equilibrium open interaction of two systems far apart

Continuing section 6.17 on non-equilibrium cases, we reexamine the system of section 6.15 but now for known energies per particle ε_A and $\varepsilon_B > \varepsilon_A$. This inequality brings the system out of the equilibrium and materializing it requires energy to be pumped from the cooler system A to the hotter system B. The situation is seen in Figure 6.9 and is different from that in Figure 6.8 in that the two systems are far apart and thus the pressure is not the same in both. This entails a degree of freedom and gives room to entropy maximization to find the unknown probabilities P_A and P_B that a particle is found in part A and B, respectively.

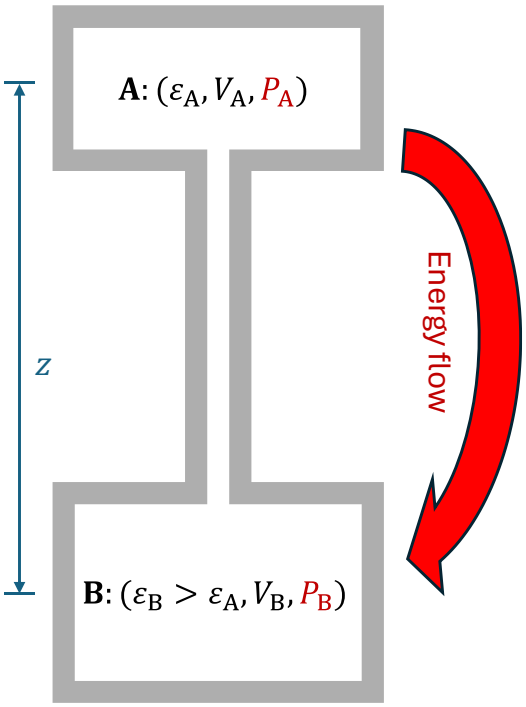

Figure 6.9 Explanation sketch for the non-equilibrium open interaction of two systems in contact, showing the characteristics in each system, where those typeset in black are known and those in red are determined by entropy maximization.

The quantities involved are similar to those in section 6.15 after replacing the energy constraint equation (6.102) with the known ε_A and ε_B. Therefore, Lagrangian in now formed by omitting the energy constraint from equation (6.103), thus getting

$$\begin{aligned} L = \varphi - \lambda_1(P_A + P_B - 1) = \\ = \frac{\beta P_A}{2}\ln\frac{\varepsilon_A}{\varepsilon^*} + \frac{\beta P_B}{2}\ln\frac{\varepsilon_B}{\varepsilon^*} + P_A \ln\frac{V_A}{v^*} + P_B \ln\frac{V_B}{v^*} - P_A \ln P_A - P_B \ln P_B \\ - \lambda_1(P_A + P_B - 1) \end{aligned} \tag{6.148}$$

Taking derivatives with respect to P_A and P_B we find

$$\begin{aligned} \frac{\partial L}{\partial P_A} = -1 - \lambda_1 - \frac{\beta}{2}\ln\frac{\varepsilon_A}{\varepsilon^*} + P_A \ln\frac{V_A}{v^*} - \ln P_A\,, \\ \frac{\partial L}{\partial P_B} = -1 - \lambda_1 + \frac{\beta}{2}\ln\frac{\varepsilon_B}{\varepsilon^*} + \ln\frac{V_B}{v^*} - \ln P_B \end{aligned} \tag{6.149}$$

Subtracting these two and equating to zero we find

$$-\ln P_{\mathrm{A}} + \ln P_{\mathrm{B}} + \frac{\beta}{2}\ln\frac{\varepsilon_{\mathrm{A}}}{\varepsilon^*} - \frac{\beta}{2}\ln\frac{\varepsilon_{\mathrm{B}}}{\varepsilon^*} + \ln\frac{V_{\mathrm{A}}}{v^*} - \ln\frac{V_{\mathrm{B}}}{v^*} = 0 \tag{6.150}$$

Exponentiating the latter equation, we find

$$\frac{P_{\mathrm{B}}}{P_{\mathrm{A}}} = \left(\frac{\varepsilon_{\mathrm{B}}}{\varepsilon_{\mathrm{A}}}\right)^{\frac{\beta}{2}}\frac{V_{\mathrm{B}}}{V_{\mathrm{A}}} \tag{6.151}$$

The temperatures and volumes per particle will be

$$\theta_{\mathrm{A}} = \frac{2\varepsilon_{\mathrm{A}}}{\beta}, \qquad \theta_{\mathrm{B}} = \frac{2\varepsilon_{\mathrm{B}}}{\beta}, \qquad v_{\mathrm{A}} = \frac{V_{\mathrm{A}}}{N_{\mathrm{A}}} = \frac{V_{\mathrm{A}}}{P_A N}, \qquad v_{\mathrm{B}} = \frac{V_{\mathrm{B}}}{N_{\mathrm{B}}} = \frac{V_{\mathrm{B}}}{P_B N} \tag{6.152}$$

Hence, from the ideal gas law we find that the pressures are

$$p_{\mathrm{A}} = \frac{\theta_{\mathrm{A}}}{v_{\mathrm{A}}} = \frac{2\varepsilon_{\mathrm{A}}/\beta}{V_{\mathrm{A}}/(P_A N)} = \frac{2N}{\beta}\frac{\varepsilon_{\mathrm{A}} P_{\mathrm{A}}}{V_{\mathrm{A}}}, \qquad p_{\mathrm{B}} = \frac{\theta_{\mathrm{B}}}{v_{\mathrm{B}}} = \frac{2N}{\beta}\frac{\varepsilon_{\mathrm{B}} P_{\mathrm{B}}}{V_{\mathrm{B}}} \tag{6.153}$$

By taking logarithms and subtracting we find

$$\ln p_{\mathrm{B}} - \ln p_{\mathrm{A}} = \left(1 + \frac{\beta}{2}\right)(\ln \varepsilon_{\mathrm{B}} - \ln \varepsilon_{\mathrm{A}}) = \left(1 + \frac{\beta}{2}\right)(\ln \theta_{\mathrm{B}} - \ln \theta_{\mathrm{A}}) \tag{6.154}$$

On the other hand, taking the differences of the standardized entropies per particle in the two parts, as given by equation (6.144), we find

$$\varphi^*_{\mathrm{B}} - \varphi^*_{\mathrm{A}} = \left(1 + \frac{\beta}{2}\right)(\ln \theta_{\mathrm{B}} - \ln \theta_{\mathrm{A}}) - (\ln p_{\mathrm{B}} - \ln p_{\mathrm{A}}) \tag{6.155}$$

and by virtue of equation (6.154), we conclude that

$$\varphi^*_{\mathrm{B}} - \varphi^*_{\mathrm{A}} = 0 \tag{6.156}$$

In other words, the two systems are now *isentropic*, rather than isobaric. The actual entropies per particle are as in section 6.17, while it is not meaningful to calculate their difference as these are not intensive properties.

The above analysis shows that all quantities do not depend on the altitude z. This may sound paradoxical, but it is not as the two subsystems have equal standardized entropies and fixed temperatures. However, as we will see in the next section 6.19, the altitude enters the scene when we examine the entire atmospheric column.

From the above equations it is easy to see that the following quantities are the same in both A and B:

$$\varphi^*, \qquad \theta^{\frac{\beta}{2}} v, \qquad \frac{\theta^{1+\frac{\beta}{2}}}{p}, \qquad p v^{1+\frac{2}{\beta}} \tag{6.157}$$

The same equations are characteristic of an *isentropic change* (or *reversible adiabatic change*) of a single system. A differential change of this type is characterized by $\mathrm{d}\Phi^* = 0$. Since $\mathrm{d}\Phi^* = \delta Q/\theta$ (equation (6.82)), we also have $\delta Q = 0$, i.e. the process involves zero heat transfer. We should stress, though, that an isentropic change of a single system is different from the coexistence of two different connected systems with the same

entropy. As we have shown in detail, that coexistence presupposes continuous removal transfer from the cold to the hot body and generates entropy at rates given again by equations (6.146) and (6.147), respectively.

We may observe that the required energy transfer to sustain a temperature difference at typical atmospheric ranges is small. Specifically, equation (6.146) can be written for the heat transfer per unit area, q, as $q = \kappa\Gamma_T$, where $\Gamma_T = (T_\mathrm{B} - T_\mathrm{A})/z$ (and in natural units $\Gamma_\theta = (\theta_\mathrm{B} - \theta_\mathrm{A})/z = k\Gamma_T$). Assuming a typical value of $\Gamma_T = 6.5$ K/km (see section 6.24) and still air, so that we only have conduction, with conductivity κ of the order of 0.02 W/(m K), we have $q = \kappa\Gamma_T = 0.02\,\mathrm{W/(m\,K)} \times 0.0065\,\mathrm{K/m} \approx 10^{-4}\mathrm{W/m^2}$. This value is exceptionally small and thus negligible, yet conceptually it is important to know that energy transfer exists and that without it the lapse rate would be $\Gamma_T = 0$ (isothermal state).

6.19 Non-equilibrium state of the air column

To investigate a non-equilibrium state in the air column we impose two additional constraints in those already set in section 6.16 that violate the isothermal state. Namely, we choose two levels $z = 0$ and $z = h$, and set their average kinetic energies to ε_0 and ε_h respectively, satisfying

$$\varepsilon_0 > \varepsilon > \varepsilon_h, \qquad \varepsilon = \frac{\beta}{\beta + 2}\varepsilon_\mathrm{tot} \tag{6.158}$$

where ε is the average kinetic energy per molecule in the column and ε_tot also includes the potential energy due to gravitation. By inspecting the derivations in section 6.16, we conclude that they still hold for this case, with the exception of abrupt changes in the two levels of known average kinetic energies. In other words entropy maximization results in the following profile:

$$\varepsilon(z) = \begin{cases} \varepsilon_0, & z = 0 \\ \varepsilon_h & z = h \\ \varepsilon & \text{otherwise} \end{cases} \tag{6.159}$$

While mathematically this (almost isothermal) profile is the one derived from maximization of entropy, it is not physically plausible. Physical reasoning implies that there will be diffusion at $z = 0$ and $z = h$, which will smooth out the transition $\varepsilon(z) = 0$ to $\varepsilon(z) = \varepsilon_h$ by making the gradient $\mathrm{d}\varepsilon/dz \propto \mathrm{d}\theta/\mathrm{d}z$ finite. This implies heat transfer, which, according to the Fourier's law (see Digression 3.E), will be (per unit area):

$$q(z) = -\kappa\frac{\mathrm{d}\theta}{\mathrm{d}z} \tag{6.160}$$

Empirical evidence suggests that κ is not constant but increases with temperature. For simplicity, we assume simple proportionality (cf. Kemp, 1962; Takhar et al., 2004):

$$\kappa = \frac{\kappa_0 T}{T_0} \tag{6.161}$$

where (κ_0, T_0) is an anchoring point. This results in

$$q(z) = -\frac{\kappa_0 \theta}{\theta_0}\frac{\mathrm{d}\theta}{\mathrm{d}z} \tag{6.162}$$

Now, this non-equilibrium situation can necessarily be accompanied by the absorption or removal of energy from the mass of the gas. Let $I(z)$ be energy inflow, i.e., the volumetric energy generation rate (in W/m^3) inside the gas at level z (or absorption if negative). We can decompose $I(z)$ into two components, $I(z) = E_{\mathrm{R}}(z) + E_{\mathrm{H}}(z)$, with the two components denoting radiation and heat, sensible or latent. For a unit area and height dz, energy conservation demands that:

$$q(z + \mathrm{d}z) = q(z) + \frac{\mathrm{d}q(z)}{\mathrm{d}z}\mathrm{d}z = q(z) + I(z)\mathrm{d}z \tag{6.163}$$

and hence

$$\frac{\mathrm{d}q(z)}{\mathrm{d}z} = I(z) \tag{6.164}$$

Combining this with equation (6.162) we find

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{\kappa_0 \theta}{\theta_0}\frac{\mathrm{d}\theta}{\mathrm{d}z}\right) = -I(z) \tag{6.165}$$

For simplicity we try a constant energy inflow independent of z, i.e. $I(z) = I$. Then

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{\kappa_0 \theta}{\theta_0}\frac{\mathrm{d}\theta}{\mathrm{d}z}\right) = -I \tag{6.166}$$

which implies

$$\frac{\kappa_0 \theta}{\theta_0}\frac{\mathrm{d}\theta}{\mathrm{d}z} = c_1 - Iz \tag{6.167}$$

where c_1 is an integration constant. This differential equation has a closed general solution, but here we adopt the following special solution, which is a linear function of z:

$$\theta = \theta_0 - \Gamma_\theta z \tag{6.168}$$

where T_0 is the temperature at $z = 0$ and Γ_θ is the minus temperature gradient (the lapse rate). Combining equations (6.167) and (6.168), we readily find

$$I = -\frac{\kappa_0 \Gamma_\theta^2}{\theta_0} < 0, \qquad c_1 = -\kappa_0 \Gamma_\theta \tag{6.169}$$

The negative value of I means net heat removal from the gas mass, which is possible through radiation emission (see Chapter 7)—or dominance of radiation emission over heat absorption.

The actual quantity of net energy removal per unit time is tiny. We assume, as in section 6.16, a standard value of $\Gamma_T = 6.5$ K/km and still air, so that we only have conduction, temperature $T_0 = 288$ K (the average temperature of Earth's surface) and conductivity κ of the order of 0.02 W/(m K). After conversion to SI units, we have $I = -\kappa_0 \Gamma_T^2 / T_0 = -0.02\ (\mathrm{W/(m\ K)}) \times 0.0065^2\,(\mathrm{K^2/m^2})/(288\ \mathrm{K}) = -3 \times 10^{-9}\ \mathrm{W/m^3}$. Multiplied by the height of the troposphere in the standard atmosphere, i.e., 11 km, this yields

a total net heat removal of $Ih \approx -3 \times 10^{-5}\,\mathrm{W/m^2}$. This becomes $-7 \times 10^{-5}\,\mathrm{W/m^2}$ for Γ_T of the order of 10 K/km. Both these are negligible quantities, yet important to sustain the temperature gradient Γ_T. The removed energy per unit time (i.e. power) from the mass of the atmosphere is counterbalanced by an equal amount of positive power imbalance warming the ground, so that overall the processes are balanced.

Combining the linear temperature profile (equation (6.168)), the ideal gas law (equation (6.59)), the relationship between density ρ and volume per particle v, i. e. , $\rho = m_0/v$ (equation (6.62)), and the hydrostatic law of pressure (equation (6.67)), and expressing all variables in terms of z and p, we find

$$\frac{dp}{dz} = \frac{m_0 g p}{\theta_0 - \Gamma_\theta z} \tag{6.170}$$

The solution of the differential equation is

$$p(z) = p_0 \left(1 - \frac{\Gamma_\theta z}{\theta_0}\right)^{\frac{m_0 g}{\Gamma_\theta}} \tag{6.171}$$

where p_0 is the pressure at $z = 0$. The other state variables are given as

$$\rho(z) = \frac{m_0 p_0}{\theta_0}\left(1 - \frac{\Gamma_\theta z}{\theta_0}\right)^{\frac{m_0 g}{\Gamma_\theta}-1}, \qquad v(z) = \frac{\theta_0}{p_0}\left(1 - \frac{\Gamma_\theta z}{\theta_0}\right)^{1-\frac{m_0 g}{\Gamma_\theta}} \tag{6.172}$$

Since the probability density of a particle being found at elevation z is proportional to the density $\rho(z)$, the former will be:

$$f(z) = \frac{g m_0}{\theta_0}\left(1 - \frac{\Gamma_\theta z}{\theta_0}\right)^{\frac{m_0 g}{\Gamma_\theta}-1} \tag{6.173}$$

The entropy per particle for specified volume V at elevation z is

$$\varphi_z(z) = \frac{\beta}{2}\ln\frac{\theta(z)}{\theta^*} + \ln\frac{V}{v^*} = \varphi_0 + \frac{\beta}{2}\ln\left(1 - \frac{\Gamma_\theta z}{\theta_0}\right), \qquad \varphi_0 \coloneqq \frac{\beta}{2}\ln\frac{\theta_0}{\theta^*} + \ln\frac{V}{v^*} \tag{6.174}$$

The standardized entropy per particle at elevation z is:

$$\begin{gathered}\varphi_z^*(z) = \frac{\beta}{2}\ln\frac{\theta(z)}{\theta^*} + \ln\frac{v(z)}{v^*} = \varphi_0^* + \left(\frac{\beta}{2} + 1 - \frac{m_0 g}{\Gamma_\theta}\right)\ln\left(1 - \frac{\Gamma_\theta z}{\theta_0}\right), \\ \varphi_0^* \coloneqq \frac{\beta}{2}\ln\frac{\theta}{\theta^*} + \ln\frac{v_0}{v^*}\end{gathered} \tag{6.175}$$

By taking the limits as $\Gamma_\theta \to 0$ in all above equations, we recover the equations for the isothermal atmospheric column, as given in section 6.16. In particular, in the limit the entropy becomes a linear function of z as in equation (6.140).

Multiplying $(\varphi_z^*(z) - \varphi_0^*)$ by (k/m_0) so as to convert it to SI units and making use of the equations in section 6.10, after algebraic manipulations we find that the classical entropy per unit mass at elevation z differs from that at zero elevation by

$$s(z) - s_0 = \left(c_p - \frac{g}{\Gamma_T}\right)\ln\left(1 - \frac{\Gamma_T z}{T_0}\right) \tag{6.176}$$

We observe that, if $\Gamma_T = g/c_p =: \Gamma_\mathrm{D} = (9.81\,\mathrm{m\,s^{-2}})/(1004\,\mathrm{J\,K^{-1}\,kg^{-1}}) = 0.0098\,\mathrm{K/m} = 9.8\,\mathrm{K/km}$ (where the value of c_p of dry air was found in Digression 6.G), the difference $s(z) - s_0$ becomes zero and the atmosphere is in *dry isentropic* state. Furthermore, from equation (6.172) we infer that if $\Gamma_\theta = m_0 g$ or $\Gamma_T = g/R$, the density becomes constant. This state, often mentioned as *homogenous*, is included here for completeness, noting that it never occurs in reality. However, states with lapse rate higher than that of the dry isotropic ($\Gamma_T > g/c_p = 9.8$ K/km) are observed, albeit rarely, and are known as *superadiabatic*. Another state, which can sometimes be met in the real atmosphere, is the *inversion*, in which the temperature increases with altitude. Graphical illustration of the different cases is provided in Figure 6.10.

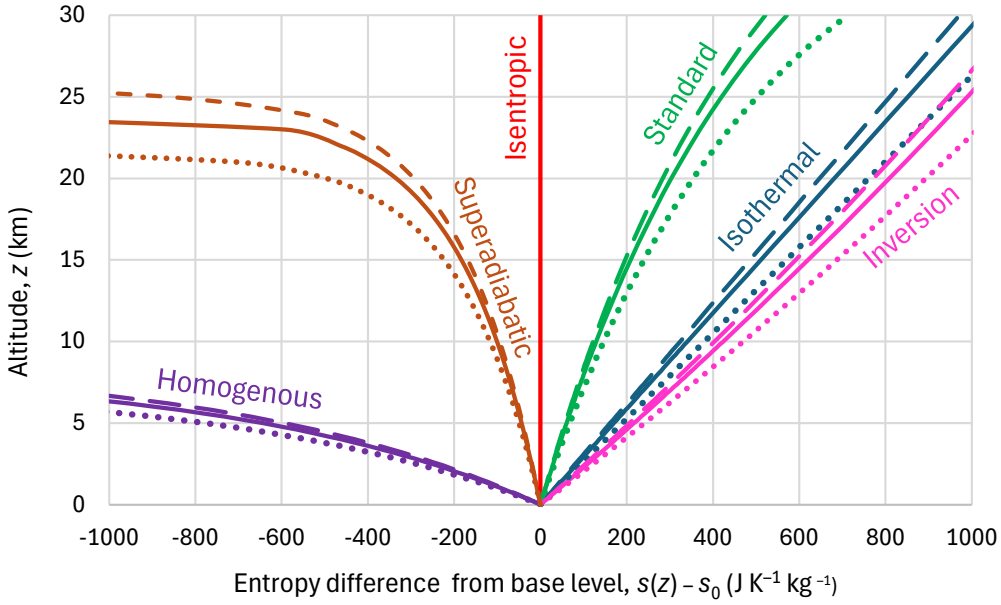

Figure 6.10 Difference of entropy per unit mass at elevation z from that at zero elevation for the indicated atmospheric states for dry atmosphere and for three temperatures T_0 at zero elevation: 288 K (continuous lines, 258 K (dotted lines) and 303 K (dashed lines). The lapse rates are: $\Gamma_T =$ for the isothermal state; $\Gamma_T = g/c_p = 9.8$ K/km for the isentropic state; $\Gamma_T = g/R = 34.2$ K/km for the homogenous state; and $\Gamma_T = 6.5$ K/km for the standard atmosphere. The specific rates chosen for illustration of the superadiabatic and inversion states are $\Gamma_T = 12$ K/km and –3 K/km, respectively.

It is stressed again that the non-equilibrium cases, i.e. all above except the isothermal, are accompanied by heat absorption or removal from the mass of the air. As we have illustrated, even tiny amounts of that heat inflow are accompanied by large departures from the isothermal (equilibrium) state. Hence, the actual profile of the atmosphere, including the lapse rate, is in perpetual change. An illustration is given in Figure 6.11 based on radiosonde data (see section 6.24 about radiosondes). We see that changes occur, even within hours, not only in the surface temperature but also in the lapse rate. For example, in the same day with a difference of 12 hours at the station Tucson, Arizona, the temperature increased from nighttime to daytime by 14 K and the profile from inverted ($\Gamma_T = -3$ K/km) became superadiabatic ($\Gamma_T = 10.5$ K/km). Near isothermal, or slightly inverted, is the nighttime temperature profile at Guimar-Tenerife (Spanish island). The profile shown for Athens is close to dry isotropic. The mechanisms

responsible for macroscopic changes, spatial and temporal, in temperature and its gradient are qualitatively discussed in Digression 6.H.

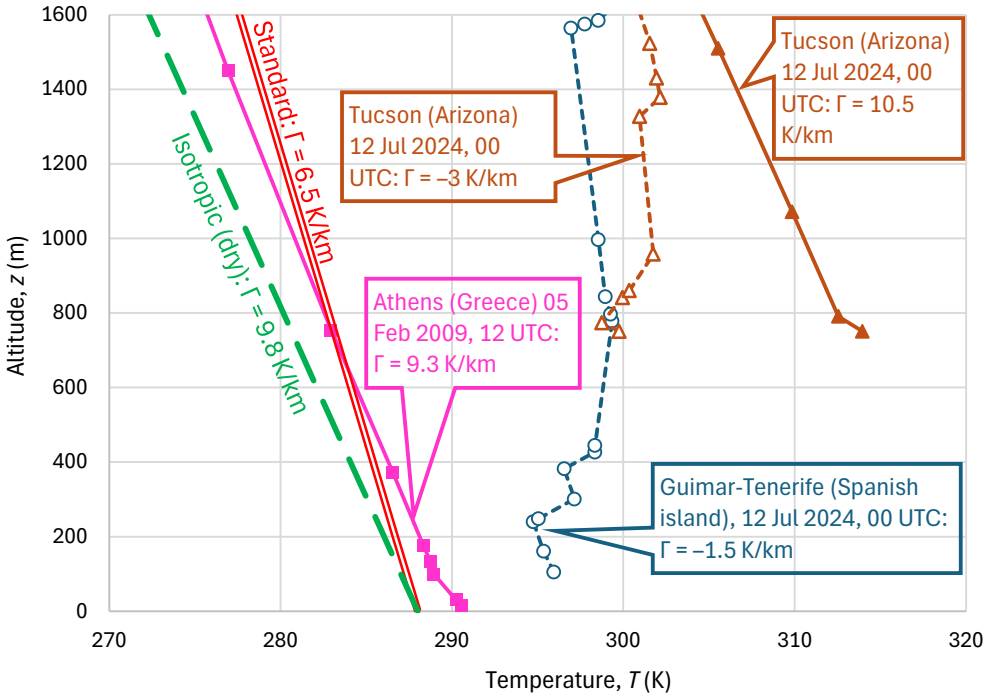

Figure 6.11 Lower troposphere profiles as observed by radiosondes at the indicated sites and dates. The slopes noted are calculated by linear regression on the points shown in the graph (not the entire range covered by radiosonde data). The profiles with continuous lines correspond to daytime and those with dashed lines to nighttime. For comparison, the standard atmospheric profile, as well as the slope of the isotropic profile are also shown. The data were retrieved from the University of Wyoming Atmospheric Science Radiosonde Archive (https://weather.uwyo.edu/upperair/sounding.shtml).

Digression 6.H: Mechanisms leading the temperature profile out of equilibrium

The mechanisms responsible for changes in the atmosphere, thus leading to departures from the equilibrium (i.e., the isothermal atmosphere) are:

- the warming of the soil and liquid water by the sunshine during the day and their cooling during the night;
- the water evaporation and transpiration at the surface level and condensation aloft;
- the convection, and the implied vertical transfer of sensible and latent heat;
- the winds caused by spatial temperature differences and influenced by Coriolis forces.

These are not static forcings, but processes, i.e., perpetual changes in the atmosphere. The processes occur on different time scales, some of which are too small to let the atmosphere evolve to the equilibrium (isothermal) state. These processes occur at a macroscopic level, with the motion of masses of air, typically referred to as parcels (see section 6.25). According to van Wijngaarden and Happer (2023), it takes a very long time for appreciable heat to flow into or out of a parcel of reasonable size because of the very small thermal conductivity of air. While this is true, we may note that temperature gradients also appear in the ocean, which can be much higher than the atmospheric for some layers in the tropics, even though the thermal conductivity is about 25 times higher. Also, deep temperature gradients appear in the atmospheres of other planets (cf. Koutsoyiannis and Tsakalias, 2025), including those with higher thermal conductivity (Jupiter and Saturn, where excessive presence of light hydrogen molecules facilitates rapid energy

transfer). In other words, non-equilibrium states are the rule and are not incompatible with the principle of maximum entropy, if this is viewed at multiple temporal and spatial scales (see sections 3.7 and 3.8).

The tendency to the equilibrium is continually distorted by numerus processes acting on different characteristic time scales. In the atmosphere, the driving mechanisms of these processes are the following and will be further discussed, with some quantification, in Chapter 7:

1. Clouds form and disappear, strongly affecting the shortwave and longwave radiation processes.
2. The Earth's surface is not homogeneous in terms of shortwave radiation absorption and reflection (spatially and temporally varying albedo).
3. Earth is round (not flat) and the sunrays come with different slopes at different places.
4. Earth rotates around its axis on a daily basis.
5. Earth rotates around the Sun on an annual basis.
6. Earth's orbit around the Sun is elliptical, resulting in changes in the distance between the two bodies.
7. The climatic system (see its definition in Chapter 8) is complex and is subject to irregular changes by internal processes and external forces (e.g. volcanos).
8. The solar radiation is subject to change due to Sun's dynamics.
9. At long scales, astronomical changes (Milanković cycles) affect the climatic system due to implied changes in radiation reaching the Earth.

6.20 Phase change

When two systems that are brought into contact are from the same substance (e.g. water) but in different phases, e.g. system A is gas (vapour) and system B is liquid (water) or solid (ice), then the equation of energy conservation needs to be adapted to include the *phase change energy*, i.e. the amount of energy per molecule ξ to break the bonds between molecules of the liquid or solid phase in order for the molecule to switch to the gaseous phase (evaporation). Also, the degrees of freedom in the two phases are different.

To derive the law describing the coexistence of gaseous and liquid phases, we study a single molecule (Figure 6.12) and maximize the combined uncertainty of its state related to:

(a) its phase (whether gaseous, denoted as A, or liquid, denoted as B);
(b) its position in space; and
(c) its kinetic state, i.e., its velocity and other coordinates corresponding to its degrees of freedom, all making up its thermal energy.

In a similar manner we can study the coexistence of gaseous and solid state (sublimation; see also section 6.21). The analysis that follows was first presented, in somewhat different formulation, for water vapour over liquid water by Koutsoyiannis (2014a) and over ice by Koutsoyiannis (2024).

The partial entropies of the two phases, i.e., the entropies conditional on the particle being in the gaseous (A) or liquid (B) phase, are:

$$\varphi_{\mathrm{A}} = \frac{\beta_{\mathrm{A}}}{2}\ln\frac{\varepsilon_{\mathrm{A}}}{\varepsilon_{\mathrm{A}}^*} + \ln\frac{V_{\mathrm{A}}}{v_{\mathrm{A}}^*}, \qquad \varphi_{\mathrm{B}} = c_{\mathrm{B}} + \frac{\beta_{\mathrm{B}}}{2}\ln\frac{\varepsilon_{\mathrm{B}}}{\varepsilon_{\mathrm{B}}^*} + \ln\frac{V_{\mathrm{B}}}{v_{\mathrm{B}}^*} \tag{6.177}$$

The total volume is V and the total number of particles is N. We further assume that the liquid phase is incompressible, so that the volume per particle

$$v_B = \frac{V_\mathrm{B}}{N_\mathrm{B}} = \frac{V_\mathrm{B}}{P_\mathrm{B} N} \tag{6.178}$$

is constant and hence the conditional entropies can be written as

$$\varphi_\mathrm{A} = \frac{\beta_\mathrm{A}}{2} \ln \frac{\varepsilon_\mathrm{A}}{\varepsilon_\mathrm{A}^*} + \ln \frac{V - v_\mathrm{B} P_\mathrm{B} N}{v_\mathrm{A}^*}, \qquad \varphi_\mathrm{B} = \frac{\beta_\mathrm{B}}{2} \ln \frac{\varepsilon_\mathrm{B}}{\varepsilon_\mathrm{B}^*} + \ln \frac{v_\mathrm{B} P_\mathrm{B} N}{v_\mathrm{B}^*} \tag{6.179}$$

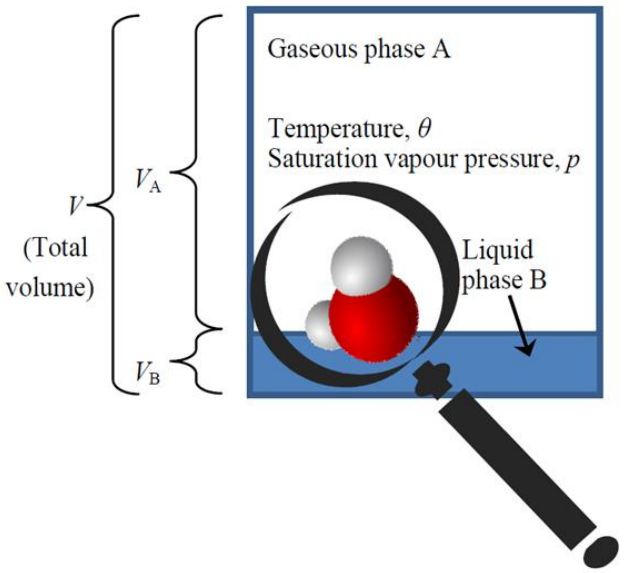

Figure 6.12 Explanatory sketch indicating the basic quantities involved in the equilibrium of water vapour with liquid water, zooming in on a single "shy" molecule which "tries to hide itself" by maximizing the combined uncertainty φ related to its phase (being either gaseous or liquid with probabilities P_A and P_B, respectively), position and kinetic state.

The unconditional entropy is:

$$\varphi = P_\mathrm{A} \varphi_\mathrm{A} + P_\mathrm{B} \varphi_\mathrm{B} + \varphi_P \tag{6.180}$$

with corresponding partition entropy:

$$\varphi_P \coloneqq -P_\mathrm{A} \ln P_\mathrm{A} - P_\mathrm{B} \ln P_\mathrm{B} \tag{6.181}$$

The two phases are in open interaction and the mass and energy conservation constraints are:

$$P_\mathrm{A} + P_\mathrm{B} = 1, \qquad P_\mathrm{A} \varepsilon_\mathrm{A} + P_\mathrm{B} (\varepsilon_\mathrm{B} - \xi) = \varepsilon \tag{6.182}$$

Hence the Lagrangian to be maximized is

$$L \coloneqq P_\mathrm{A} \varphi_\mathrm{A} + P_\mathrm{B} \varphi_\mathrm{B} - \lambda_1 (P_\mathrm{A} + P_\mathrm{B} - 1,) - \lambda_2 (P_\mathrm{A} \varepsilon_\mathrm{A} + P_\mathrm{B} (\varepsilon_\mathrm{B} - \xi) - \varepsilon) - P_\mathrm{A} \ln P_\mathrm{A} - P_\mathrm{B} \ln P_\mathrm{B} \tag{6.183}$$

Taking derivatives in (6.183) with respect to ε_A and ε_B we find

$$\frac{\partial L}{\partial \varepsilon_\mathrm{A}} = \frac{\beta_\mathrm{A} P_\mathrm{A}}{2 \varepsilon_\mathrm{A}} - \lambda_2 P_\mathrm{A}, \qquad \frac{\partial L}{\partial \varepsilon_\mathrm{B}} = \frac{\beta_\mathrm{B} P_\mathrm{B}}{2 \varepsilon_\mathrm{B}} - \lambda_2 P_\mathrm{B} \tag{6.184}$$

and equating them to zero we readily find

$$\frac{\beta_{\mathrm{A}}}{2\varepsilon_{\mathrm{A}}}=\frac{\beta_{\mathrm{B}}}{2\varepsilon_{\mathrm{B}}}=\lambda_2 \tag{6.185}$$

By virtue of equation (6.52), the temperatures will be equal in both phases:

$$\theta_{\mathrm{A}}=\theta_{\mathrm{A}}=\theta=\frac{2\varepsilon_{\mathrm{A}}}{\beta_{\mathrm{A}}}=\frac{2\varepsilon_{\mathrm{B}}}{\beta_{\mathrm{B}}} \tag{6.186}$$

Taking derivatives in (6.183) with respect to P_{A} and P_{B} we find

$$\begin{aligned}\frac{\partial L}{\partial P_{\mathrm{A}}}&=\varphi_{\mathrm{A}}-1-\lambda_1-\frac{\beta_{\mathrm{A}}}{2}-\ln P_{\mathrm{A}}\ ,\\ \frac{\partial L}{\partial P_{\mathrm{B}}}&=-\frac{v_{\mathrm{B}}P_{\mathrm{A}}N}{V-v_{\mathrm{B}}P_{\mathrm{B}}N}+\varphi_{\mathrm{B}}-\lambda_1-\frac{\beta_{\mathrm{B}}}{2}\left(1-\frac{\xi}{\varepsilon_{\mathrm{B}}}\right)-\ln P_{\mathrm{B}}\end{aligned} \tag{6.187}$$

The denominator of the first term of $\partial L/\partial P_{\mathrm{B}}$ can also be written as $v_{\mathrm{A}}p_{\mathrm{A}}N$, and hence the fraction equals $v_{\mathrm{B}}/v_{\mathrm{A}}$, which is negligible, because $v_{\mathrm{B}}\ll v_{\mathrm{A}}$. Equating the two derivatives to zero, and eliminating λ_1 we find

$$\varphi_{\mathrm{A}}-1-\frac{\beta_{\mathrm{A}}}{2}-\ln P_{\mathrm{A}}=\varphi_{\mathrm{B}}-\frac{\beta_{\mathrm{B}}}{2}\left(1-\frac{\varepsilon_{\mathrm{B}}}{\xi}\right)-\ln P_{\mathrm{B}} \tag{6.188}$$

Denoting $\varphi_{\mathrm{A}}^*=\varphi_{\mathrm{A}}-\ln P_{\mathrm{A}}$ (cf. Digression 6.C), this can be written as

$$\varphi_{\mathrm{A}}^*-\varphi_{\mathrm{B}}^*=1+\frac{\beta_{\mathrm{A}}}{2}-\frac{\beta_{\mathrm{B}}}{2}+\frac{\beta_{\mathrm{B}}}{2}\frac{\xi}{\varepsilon_{\mathrm{B}}}=\frac{\xi}{\theta}-\left(\frac{\beta_{\mathrm{B}}}{2}-\frac{\beta_{\mathrm{A}}}{2}-1\right) \tag{6.189}$$

It is useful to note here that, if we multiply both sides of equation (6.189) by θ, the resulting quantity

$$\lambda\coloneqq(\varphi_{\mathrm{A}}^*-\varphi_{\mathrm{B}}^*)\theta=\xi-\left(\frac{\beta_{\mathrm{B}}}{2}-\frac{\beta_{\mathrm{A}}}{2}-1\right)\theta \tag{6.190}$$

is the *latent heat of vaporization* per molecule.

The standardized entropies per particle in the two phases expressed in terms of $\theta, p, v_{\mathrm{B}}$ are (from equation (6.68)):

$$\varphi_{\mathrm{A}}^*=\left(1+\frac{\beta_A}{2}\right)\ln\frac{\theta}{\theta_{\mathrm{A}}^*}-\ln\frac{p}{p_{\mathrm{A}}^*},\qquad \varphi_{\mathrm{B}}^*=\frac{\beta_{\mathrm{B}}}{2}\ln\frac{\theta}{\theta_{\mathrm{B}}^*}+\ln\frac{v_{\mathrm{B}}}{v_{\mathrm{B}}^*} \tag{6.191}$$

The entropy difference is

$$\varphi_{\mathrm{A}}^*-\varphi_{\mathrm{B}}^*=-c_1-\left(\frac{\beta_{\mathrm{B}}}{2}-\frac{\beta_{\mathrm{A}}}{2}-1\right)\ln\frac{\theta}{\theta_{\mathrm{A}}^*}-\ln\frac{p}{p_{\mathrm{A}}^*},\qquad c_1\coloneqq\ln\frac{v_{\mathrm{B}}}{v_{\mathrm{B}}^*}+\frac{\beta_{\mathrm{B}}}{2}\ln\frac{\theta_{\mathrm{A}}^*}{\theta_{\mathrm{B}}^*} \tag{6.192}$$

Hence

$$\frac{\xi}{\theta}-\left(\frac{\beta_{\mathrm{B}}}{2}-\frac{\beta_{\mathrm{A}}}{2}-1\right)=-c_1-\left(\frac{\beta_{\mathrm{B}}}{2}-\frac{\beta_{\mathrm{A}}}{2}-1\right)\ln\frac{\theta}{\theta_{\mathrm{A}}^*}-\ln\frac{p}{p_{\mathrm{A}}^*} \tag{6.193}$$

Solving (6.193) for $\ln p$ and substituting to (6.191) after algebraic manipulations we find

$$\varphi_{\mathrm{A}}^{*}=c_1-\left(\frac{\beta_{\mathrm{B}}}{2}-\frac{\beta_{\mathrm{A}}}{2}-1\right)+\frac{\beta_{\mathrm{B}}}{2}\ln\frac{\theta}{\theta_{\mathrm{A}}^{*}}+\frac{\xi}{\theta},\qquad \varphi_{\mathrm{B}}^{*}=c_1+\frac{\beta_{\mathrm{B}}}{2}\ln\frac{\theta}{\theta_{\mathrm{A}}^{*}} \tag{6.194}$$

Notice the term $\beta_{\mathrm{B}}/2$ (rather $\beta_{\mathrm{A}}/2$) in φ_{A}^{*}, which is a necessity for (6.189) to hold true.

Now, exponentiating (6.193) we find

$$p=\frac{p_{\mathrm{A}}^{*}}{\mathrm{e}^{c_2}}\exp\left(-\frac{\xi}{\theta}\right)\left(\frac{\theta}{\theta_{\mathrm{A}}^{*}}\right)^{-(\beta_{\mathrm{B}}/2-\beta_{\mathrm{A}}/2-1)} \tag{6.195}$$

Assuming that at some temperature θ_0, $p(\theta_0)=p_0$, we write (6.195) in a more convenient manner as:

$$p=p_0\exp\left(\frac{\xi}{\theta_0}\left(1-\frac{\theta_0}{\theta}\right)\right)\left(\frac{\theta_0}{\theta}\right)^{\beta_{\mathrm{B}}/2-\beta_{\mathrm{A}}/2-1} \tag{6.196}$$

In the classical thermodynamical formalism, this takes the form:

$$p=p_0\exp\left(\alpha_1\left(1-\frac{T_0}{T}\right)\right)\left(\frac{T_0}{T}\right)^{\alpha_2} \tag{6.197}$$

where

$$\alpha_1\coloneqq\frac{\xi}{kT_0},\qquad \alpha_2\coloneqq\frac{\beta_{\mathrm{B}}}{2}-\frac{\beta_{\mathrm{A}}}{2}-1=\frac{c_{\mathrm{L}}-c_p}{R} \tag{6.198}$$

and c_{L} is the specific heat of the liquid phase.

A detailed application of the above framework to the phase change of water is given in section 6.21, combined with proof of its accuracy. Note that, once the constants in (6.196)–(6.197) are known, their numerical evaluation is direct as the derived equation is closed for p as well as for θ or T. Specifically, the inversion of (6.197) yields

$$T=T_0\frac{\alpha_1}{\alpha_2}\frac{1}{-W_{-1}\left(-\frac{\alpha_1}{\alpha_2}\exp\left(-\frac{\alpha_1}{\alpha_2}\right)\left(\frac{p}{p_0}\right)^{\frac{1}{\alpha_2}}\right)} \tag{6.199}$$

where $W_{-1}(z)$ is the Lambert W function of z (non-principal real branch; see Appendix 6-I).

Now, if we take the differential in equation (6.189) we write

$$\mathrm{d}(\varphi_{\mathrm{A}}^{*}-\varphi_{\mathrm{B}}^{*})=-\frac{\xi}{\theta^2}d\theta=-\frac{\xi}{\theta}\frac{d\theta}{\theta}=\left(-(\varphi_{\mathrm{A}}^{*}-\varphi_{\mathrm{B}}^{*})-\left(\frac{\beta_{\mathrm{B}}}{2}-\frac{\beta_{\mathrm{A}}}{2}-1\right)\right)\frac{d\theta}{\theta} \tag{6.200}$$

Likewise, from equation (6.192) we have

$$\mathrm{d}(\varphi_{\mathrm{A}}^{*}-\varphi_{\mathrm{B}}^{*})=-\left(\frac{\beta_{\mathrm{B}}}{2}-\frac{\beta_{\mathrm{A}}}{2}-1\right)\frac{d\theta}{\theta}-\frac{dp}{p} \tag{6.201}$$

Equating the right-hand sides of the two, after algebraic manipulations we find

$$(\varphi_{\mathrm{A}}^{*}-\varphi_{\mathrm{B}}^{*})\frac{d\theta}{\theta}=\frac{dp}{p} \tag{6.202}$$

or

$$\frac{dp}{d\theta} = (\varphi_{\mathrm{A}}^{*} - \varphi_{\mathrm{B}}^{*})\frac{p}{\theta} \tag{6.203}$$

After conversion to classical units, this is the well-known *Clausius-Clapeyron equation*, a differential equation whose solution, under the assumption stated, is obviously (6.195). Here we derived it as a result of entropy maximization, just for the completeness of the presentation. In fact, the differential form is not necessary because in application only the closed solution is actually needed.

As a cautionary note, it is useful to mention that in several classical and statistical thermodynamics books, the Clausius-Clapeyron equation is typically integrated using an incorrect assumption of a constant latent heat, $\lambda = (\varphi_{\mathrm{A}}^{*} - \varphi_{\mathrm{B}}^{*})\theta = \xi$. This results in the quite common solution:

$$p = c\exp\left(-\frac{\xi}{\theta}\right) = c\exp\left(-\frac{L}{kT}\right) \tag{6.204}$$

which however is inaccurate.

In closing this section, it is useful to emphasize the importance of the results presented. They determine how much water can be evaporated (and subsequently condensed in the atmosphere), thus providing the physical basis of the hydrological cycle. And as has been shown, the scientific foundation of the characterization of evaporation relies on the combination of entropy maximization (equation (6.180)) with energy availability (equation (6.182)). The water availability is another factor determining the quantity that is actually evaporated. Further information useful for the characterization of atmospheric water is given in section 6.21

6.21 Phase change of water

As we have discussed in Digression 6.E, the water molecule, as a triatomic molecule with nonlinear structure, has $\beta_{\mathrm{A}} = 6$ degrees of freedom in its gaseous phase. The number of degrees of freedom in the liquid phase is greater than β_{A} because of the "social behaviour" of water molecules. Specifically, in addition to the translational and rotational degrees of freedom of individual molecules, there are local clusters with low energy vibrational modes that can be thermally excited. The average number of degrees of freedom per molecule (individual and collective involving more than one water molecules) is very high, $\beta_{\mathrm{L}} = 18$ (e.g., Fraundorf, 2003). Hence the constant $\alpha_2 \coloneqq \beta_{\mathrm{B}}/2 - \beta_{\mathrm{A}}/2 - 1 = 5$. We slightly modify this value to $\alpha_2 = 5.06$ to agree with the experimental quantity $(c_{\mathrm{L}} - c_p)/R$.

To avoid ambiguity, while in section 6.20 we used the symbol p for the (partial) pressure of the substance coexisting in gaseous and liquid (or solid) phase, in what follows we will replace it with e (which is typically used in meteorology) and reserve p for the total pressure.

To anchor our relationships on the phase transition of water, we choose as reference point the triple point of water, which is experimentally defined with accuracy as $T_0 =$

273.16 K (= 0.01°C) and $e_0 = 6.11657$ hPa (Wagner and Pruss, 2002). It is recalled that the triple point of a substance is the state (temperature and pressure) at which the three phases (gas, liquid, and solid) of that substance coexist in thermodynamic equilibrium. The specific heat of water vapour for constant pressure, again determined at the triple point, is $c_p = 1884.4$ J kg^{-1} K^{-1} and that of liquid water is $c_\mathrm{L} = 4219.9$ J kg^{-1} K^{-1} (Wagner and Pruss, 2002), so that $c_\mathrm{L} - c_p = 2335.5$ J kg^{-1} K^{-1} and $(c_\mathrm{L} - c_p)/R = 5.06$, where the specific gas constant of water vapour is $R = 461.5$ J kg^{-1} K^{-1} (Digression 6.E).

The latent heat of vaporization is determined by equation (6.190), which after conversion into the classical formalism (noting that $k/m_0 = R$) and per unit mass will be

$$L \coloneqq \frac{\lambda}{m_0} = \frac{\lambda R}{k} = \frac{\xi}{m_0} - \left(\frac{\beta_\mathrm{B}}{2} - \frac{\beta_\mathrm{A}}{2} - 1\right)\frac{k}{m_0}T = \frac{\xi}{m_0} - \left(\frac{\beta_\mathrm{B}}{2} - \frac{\beta_\mathrm{A}}{2} - 1\right)RT \tag{6.205}$$

or

$$L = \frac{\xi}{m_0} - (c_\mathrm{L} - c_p)T \tag{6.206}$$

The experimental value of L_0 at T_0 is $L_0 = 2.501 \times 10^6$ J kg^{-1} so that:

$$\frac{\xi}{m_0} = L_0 + (c_L - c_p)T_0 = 3.139 \times 10^6 \text{ J kg}^{-1}, \quad \alpha_1 \coloneqq \frac{\xi}{kT_0} = \frac{\xi}{m_0}\frac{1}{RT_0} = 24.921 \tag{6.207}$$

where we have slightly modified the last two decimal digits of the constant α_1 to optimize its fit to the data (see below), and

$$L/(\text{J kg}^{-1}) = 3.139 \times 10^6 - 2336\, T/\text{K} = 2.501 \times 10^6 - 2336\, T/°\text{C} \tag{6.208}$$

In Figure 6.13 it is verified that equation (6.208) is very close to tabulated data from Smithsonian Meteorological Tables (List, 1951), as well as a commonly suggested empirical linear equation for latent heat (Shuttleworth, 1993). It is important to know that the entropic framework which gives the saturation vapour pressure is the same framework that predicts the relationship of the latent heat of vaporization with temperature.

With the above values of thermodynamic properties, equation (6.197) becomes

$$e = e_0 \exp\left(24.921\left(1 - \frac{T_0}{T}\right)\right)\left(\frac{T_0}{T}\right)^{5.06}, \qquad T_0 = 273.16 \text{ K}, \;\; e_0 = 6.11657 \text{ hPa} \tag{6.209}$$

For comparison, equation (6.204) (based on the inconsistent integration of the Clausius-Clapeyron equation for constant L) is

$$e = e_0 \exp\left(19.84\left(1 - \frac{T_0}{T}\right)\right) \tag{6.210}$$

This can also be derived by using the Newton-Raphson numerical method of approximation (notice that 19.84 = 24.92 – 5.06) at an initial condition $T_0/T = 1$.

Empirical equations based on observations are in common use. These are regarded to be the most accurate—but in fact are less accurate than (6.209) (see Koutsoyiannis, 2012). Figure 6.14 compares the two theoretical equations (6.209) and (6.210) also with

one empirical. All seem indistinguishable. However, Figure 6.15, which compares relative differences from measurements, clearly indicates the inappropriateness of (6.210).

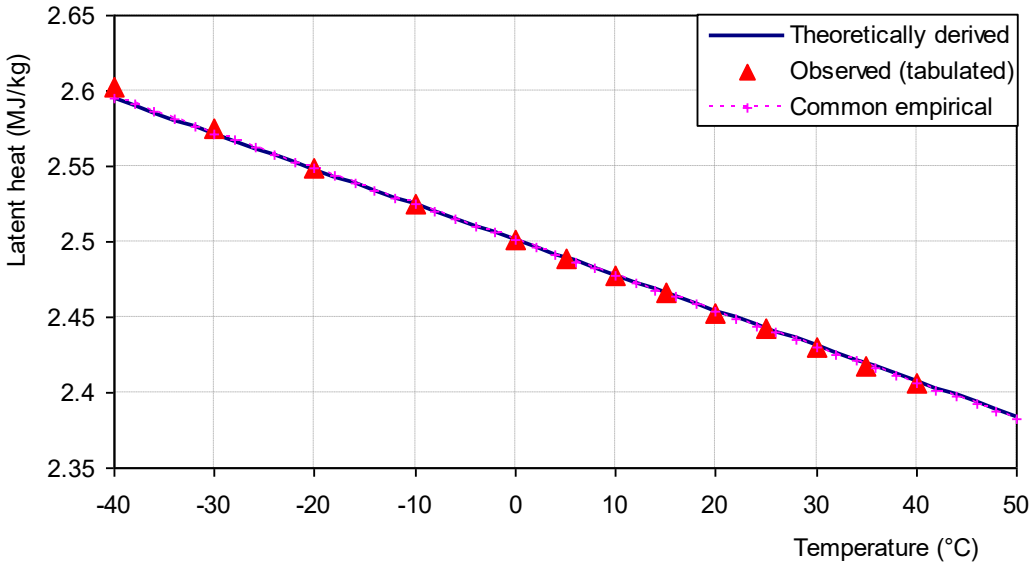

Figure 6.13 Comparison of latent heat of water as given by equation (6.208) and with standard tabulated data from List (1951), as well as with a common empirical equation (Shuttleworth, 1993), which is $L/(\mathrm{J\,kg^{-1}}) = 2.501 \times 10^6 - 2361\, T/°\mathrm{C}$. (Source: Koutsoyiannis, 2012.)

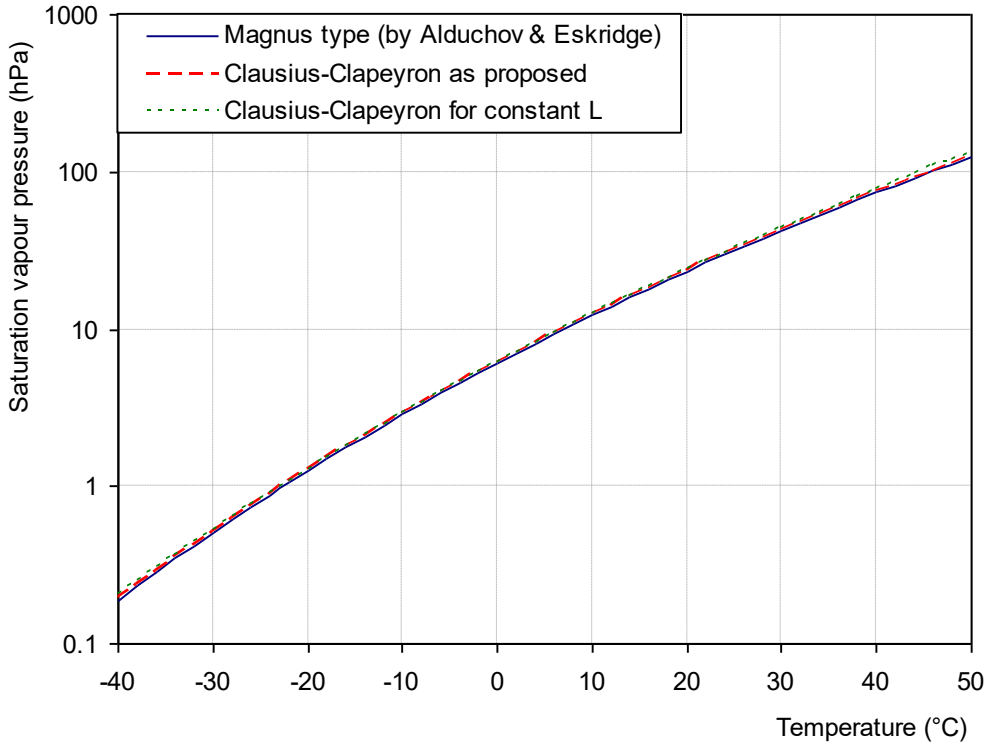

Figure 6.14 Comparison of the saturation water vapour pressure obtained by the consistent equation (6.209) as well as by the standard but inconsistent equation (6.210). In addition, an empirical equation, $e/\mathrm{hPa} = 6.1094 \exp((17.625\, T/°C)/(243.04 + T/°C))$ from Alduchov & Eskridge (1996) is also compared. (Source: Koutsoyiannis, 2012.)

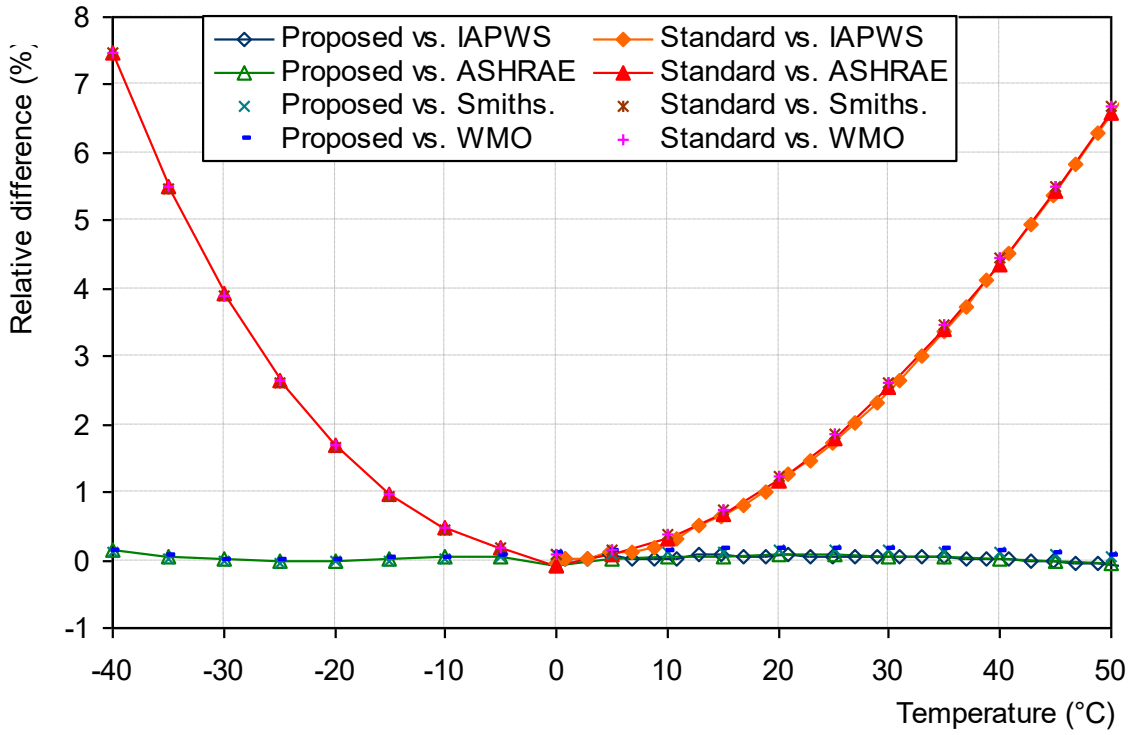

Figure 6.15 . Comparison of relative differences of the saturation water vapour pressure obtained by the consistent equation (6.209) as well as by the inconsistent equation (6.210), with observed data. Four reference data sets are used, which are given in tabulated form from different origins: (a) the International Association for the Properties of Water and Steam(IAPWS), (b) the Smithsonian Meteorological Tables (Smiths.), (c) the World Meteorological Organization (WMO) meteorological tables, and (d) the American Society of Heating, Refrigerating and Air-conditioning Engineers (ASHRAE). (Source: Koutsoyiannis, 2012.)

For the saturation water pressure over ice, where the equilibrium is between gaseous and solid phases, it suffices to replace in equations (6.197)-(6.198) the specific heat of liquid water c_{L} with that of ice, c_{I}, and the latent heat of vaporization with that of sublimation, resulting in a constant a_{I} to substitute for a_1. Following Ambaum (2020), we adopt the value $c_{\mathrm{I}} = 2097\ \mathrm{J\,kg^{-1}\,K^{-1}}$ and hence $(c_{\mathrm{I}} - c_p)/R = 0.461$. Optimizing the average relative square error from benchmark values provided by Murphy and Koop (2005; Appendix C) for temperatures 150 to 273.16 K, we find $\alpha_{1_{\mathrm{I}}} = 22.812$. Hence, the equation for the saturation water pressure over ice becomes:

$$e_{\mathrm{I}}(T) = e_0 \exp\left(22.812\left(1 - \frac{T_0}{T}\right)\right)\left(\frac{T_0}{T}\right)^{0.461}, \quad T_0 = 273.16\ \mathrm{K}, e_0 = 6.11657\ \mathrm{hPa} \tag{6.211}$$

6.22 Quantification of water vapour in the atmosphere

In all the above derivations and applications on the phase change of water we did not take into account the fact that the atmosphere contains other gases. And indeed, the presence of other gases does not have a role in the process. The pressure e calculated by equation (6.209) is known as the *saturation water vapour pressure*. It is the partial pressure of the water molecules alone. The latter is a function of the temperature, $e = e(T)$, while, as a

result of entropy maximization, the temperature T is the same for the molecules of all gases.

As we will see below (section 6.24), in the troposphere the atmospheric temperature is decreasing with the increase of the altitude. As $e(T)$ is an increasing function (equation (6.209) and Figure 6.15), the saturation water pressure is also decreased at high altitudes. Therefore, at high altitudes the water availability often exceeds that implied by the decreased $e(T)$ and in this case the water vapour condenses in liquid or solid phases, forming clouds. In other cases, the water availability is lower than implied by $e(T)$. In this case the actual water pressure, which we denote as e_a, is lower than $e(T)$.

The ratio

$$U \coloneqq \frac{e_\mathrm{a}}{e(T)} \tag{6.212}$$

is called relative humidity. The temperature at which a given air mass will reach saturation is called *dew point*, T_d. This means that

$$T_\mathrm{d} \coloneqq e^{-1}(e_\mathrm{a}) \Leftrightarrow e_\mathrm{a} = e(T_\mathrm{d}) \tag{6.213}$$

where e^{-1} denotes the inverse of function $e(T)$. As the relative humidity is the ratio of vapour pressures, it is often useful to quantify the presence of humidity in terms of the ratio of masses or densities. This quantification is done by the *specific humidity*, q, and the mixing ratio r, defined as

$$q \coloneqq \frac{\rho_\mathrm{V}}{\rho_\mathrm{D} + \rho_\mathrm{V}}, \qquad r \coloneqq \frac{\rho_\mathrm{V}}{\rho_\mathrm{D}}, \qquad \left(q = \frac{r}{1+r}, \qquad r = \frac{q}{1-q}\right) \tag{6.214}$$

where ρ_V and ρ_D are the densities of water vapour and dry air, respectively.

It is easily shown that U, q and r are one-to-one related through

$$q = \frac{\epsilon e_\mathrm{a}}{p - (1-\epsilon)e_\mathrm{a}} = \frac{\epsilon U e}{p - (1-\epsilon)Ue} \Leftrightarrow U = \frac{p}{e}\,\frac{q}{\epsilon + (1-\epsilon)q} \tag{6.215}$$

where ϵ is the ratio of the molecular mass of water to that of the mixture of gases in the dry air. From Digression 6.E, the former is 18.015 kg/kmol, and from Digression 6.G the latter is 28.96 kg/kmol, so that $\epsilon = 0.622$.

If we know the temperature and the dew point, then the relative humidity is calculated as:

$$U = \frac{e(T_\mathrm{d})}{e(T)} = \exp\left(a_1\left(\frac{T_0}{T} - \frac{T_0}{T_\mathrm{d}}\right)\right)\left(\frac{T}{T_\mathrm{d}}\right)^{a_2} \tag{6.216}$$

where, as we have seen in section 6.21, $T_0 = 273.16$ K, $a_1 = 24.921$ and $a_2 = 5.06$. The inverse relationship is

$$T_\mathrm{d} = T_0\frac{a_1}{a_2}\frac{1}{-W_{-1}\left(-\frac{a_1}{a_2}\frac{T_0}{T}\exp\left(-\frac{a_1}{a_2}\frac{T_0}{T}\right)U^{\frac{1}{a_2}}\right)} \tag{6.217}$$

where $W_{-1}(z)$ is the Lambert W function of z (non-principal real branch; see Appendix 6-I).

6.23 Moist air and the moist isentropic profile

As seen in Digression 6.E, the characteristics of the air as a gas mixture depend on those of nitrogen, oxygen, argon and water vapour (Table 6.3). The first three are well mixed in the atmosphere, but the mole fraction of water vapour, while averaging to 0.4% globally, varies considerably in space and time. In the full absence of water vapour, we have the dry atmosphere, whose standardized entropy per particle, according to equation (6.68), is

$$\varphi_{\mathrm{D}}^{*}=\left(1+\frac{\beta_{\mathrm{D}}}{2}\right)\ln\frac{\theta}{\theta_{\mathrm{D}}^{*}}-\ln\frac{p}{p_{\mathrm{D}}^{*}}, \tag{6.218}$$

However, when there is moisture in the atmosphere, whose partial pressure is e, the standardized entropy per dry-air particle is

$$\varphi_{\mathrm{D}}^{*}=\left(1+\frac{\beta_{\mathrm{D}}}{2}\right)\ln\frac{\theta}{\theta_{\mathrm{D}}^{*}}-\ln\frac{p-e}{p_{\mathrm{D}}^{*}}, \tag{6.219}$$

To find the total standardized entropy for a mixture consisting of N_{D} particles of dry air (treated as a mixture of the well mixed gases; see section 6.12), N_{A} particles of water vapour and N_{B} particles of water in liquid phase (droplets in clouds), we consider equation (6.194) and write

$$\begin{aligned}\Phi^{*}=N_{\mathrm{D}}&\left(\left(1+\frac{\beta_{\mathrm{D}}}{2}\right)\ln\frac{\theta}{\theta_{\mathrm{D}}^{*}}-\ln\frac{(p-e)}{p_{\mathrm{D}}^{*}}\right)\\&+N_{\mathrm{A}}\left(c_1-\left(\frac{\beta_{\mathrm{B}}}{2}-\frac{\beta_{\mathrm{A}}}{2}-1\right)+\frac{\beta_{\mathrm{B}}}{2}\ln\frac{\theta}{\theta_{\mathrm{A}}^{*}}+\frac{\xi}{\theta}\right)+N_{\mathrm{B}}\left(c_1+\frac{\beta_{\mathrm{B}}}{2}\ln\frac{\theta}{\theta_{\mathrm{A}}^{*}}\right)\end{aligned} \tag{6.220}$$

Setting $N_{\mathrm{A}}+N_{\mathrm{B}}=:N_{\mathrm{W}}$ we get

$$\begin{aligned}\Phi^{*}=N_{\mathrm{D}}&\left(\left(1+\frac{\beta_{\mathrm{D}}}{2}\right)\ln\frac{\theta}{\theta_{\mathrm{D}}^{*}}-\ln\frac{(p-e)}{p_{\mathrm{D}}^{*}}\right)+N_{\mathrm{W}}\left(c_1+\frac{\beta_{\mathrm{B}}}{2}\ln\frac{\theta}{\theta_{\mathrm{A}}^{*}}\right)\\&+N_{\mathrm{A}}\left(\frac{\xi}{\theta}-\left(\frac{\beta_{\mathrm{B}}}{2}-\frac{\beta_{\mathrm{A}}}{2}-1\right)\right)\end{aligned} \tag{6.221}$$

In an isentropic change from state 1 to state 2 we will have $\Phi_2^{*}-\Phi_1^{*}=0$, while N_{D} and N_{W} are constant and $N_{\mathrm{A}}, \theta, p, e$ vary. Hence, considering equation (6.190), we write:

$$0=N_{\mathrm{D}}\left(\left(1+\frac{\beta_{\mathrm{D}}}{2}\right)\ln\frac{\theta_2}{\theta_1}-\ln\frac{p_2-e_2}{p_1-e_1}\right)+N_{\mathrm{W}}\frac{\beta_{\mathrm{B}}}{2}\ln\frac{\theta_2}{\theta_1}+N_{\mathrm{A}_2}\frac{\lambda_2}{\theta_2}-N_{\mathrm{A}_1}\frac{\lambda_1}{\theta_1} \tag{6.222}$$

Dividing by N_{D}, we get

$$0=\left(1+\frac{\beta_{\mathrm{D}}}{2}\right)\ln\frac{\theta_2}{\theta_1}-\ln\frac{p_2-e_2}{p_1-e_1}+\frac{N_{\mathrm{W}}}{N_{\mathrm{D}}}\frac{\beta_{\mathrm{B}}}{2}\ln\frac{\theta_2}{\theta_1}+\frac{N_{\mathrm{A}_2}}{N_{\mathrm{D}}}\frac{\lambda_2}{\theta_2}-\frac{N_{\mathrm{A}_1}}{N_{\mathrm{D}}}\frac{\lambda_1}{\theta_1} \tag{6.223}$$

To convert Φ^{*}/N_{D} to the classical units, we multiply by R_{D} and use the mixing ratio, so that, considering equations (6.63) and (6.214), we have

$$r \coloneqq \frac{\rho_{\mathrm{V}}}{\rho_{\mathrm{D}}} = \frac{N_{\mathrm{A}} m_{0_{\mathrm{A}}}}{N_{\mathrm{D}} m_{0_{\mathrm{D}}}} = \frac{N_{\mathrm{A}} R_{\mathrm{D}}}{N_{\mathrm{D}} R_{\mathrm{A}}} \Longrightarrow \frac{N_{\mathrm{A}} R_{\mathrm{D}}}{N_{\mathrm{D}}} = r R_{\mathrm{A}} \tag{6.224}$$

and considering equation (6.205) we obtain

$$0 = \left(c_{p_{\mathrm{D}}} + r_{\mathrm{W}} c_{\mathrm{B}}\right) \ln \frac{T_2}{T_1} - R_{\mathrm{D}} \ln \frac{p_2 - e_2}{p_1 - e_1} + \frac{r_2 L_2}{T_2} - \frac{r_1 L_1}{T_1} \tag{6.225}$$

where r_{W} is the total mixing ratio of water in gaseous and liquid phases. Equation (6.225) does not have a closed solution but needs to be solved numerically. We assume that $T_2 < T_1$ (so that saturated air at state 1 remains saturated also at state 2) and designate as state 1 the one in which all water molecules are at gaseous phase, so that $N_{\mathrm{W}} = N_{\mathrm{A}_1}$ $r_{\mathrm{W}} = r_1$ (and hence $N_{\mathrm{W}} R_{\mathrm{D}} / N_{\mathrm{D}} = r_1 R_{\mathrm{A}} = r_1 R_{\mathrm{B}}$, because both water phases have the same constant R). Depending on particular conditions of the occurrence of a change, we have the following cases:

1. *Moist isentropic* (also known as *moist adiabatic*): The total mixing ratio r_{W} in equation (6.225) is constant (we do not have leak of water). This remains the same at state 2 ($r_{\mathrm{W}_2} = r_{\mathrm{W}_1} = r_{\mathrm{W}} = r_1$).
2. *Pseudo-isentropic* (also known as *pseudoadiabatic*): The water that is converted from gaseous to liquid phase does not remain in the air but is removed through precipitation. If the difference $\Delta p \coloneqq p_1 - p_2$ is small, we can set $r_{\mathrm{W}} = r_1$, and once we find T_2 from equation (6.225), we can calculate r_2 for T_2 and assume that a water quantity proportional to $r_2 - r_1$ has precipitated. If Δp is large, we subdivide it to smaller steps and apply the same procedure iteratively, updating r_{W} with r_2 for the next iteration.
3. *Simplified pseudo-isentropic*: This is an approximation, used very often, for both previous cases, in which we disregard the details. Specifically, we drop $r_1 c_{\mathrm{B}}$ in the first term in equation (6.225). The approximation works better for the pseudo-isentropic case, but the error is not large even for the moist isentropic case.

Numerical illustration and graphical comparison of the three cases are shown in Figure 6.16, with initial conditions $p_1 = 1010$ hPa (close to the standard atmosphere of 1013 hPa) and $T_1 = 288$ K (close to the average Earth's temperature). The calculations were done for steps of $\Delta p = 10$ hPa. It is seen that the curves for the pseudoadiabatic and the simplified cases are virtually indistinguishable, while that of the moist adiabatic gives slightly higher temperatures. For additional comparison, the dry isotropic case is also shown the figures. In this, all quantities related to the presence of water are zero, so that equation (6.225) becomes

$$0 = c_{p_{\mathrm{D}}} \ln \frac{T_2}{T_1} - R_{\mathrm{D}} \ln \frac{p_2}{p_1} \tag{6.226}$$

hence being identical to equation (6.157) (standard isentropic change). Additionally, Figure 6.16 shows the pseudo-isentropic curves for temperatures at level $p_1 = 1010$ hPa of 15 K higher ($T_1 = 303$ K) and 30 K lower ($T_1 = 258$ K) than in the main case ($T_1 = 288$ K).

Now we assume that the pressure p, calculated as above, represents that of atmospheric column in isentropic state, and varies with the elevation z according to the hydrostatic law of equation (6.67). `We can thus convert the relationship $T(p)$, shown in the upper panel of Figure 6.16 into $T(z)$, through numerical integration, of ρg where $\rho = p/RT$. With this conversion, we can calculate a moist isentropic lapse rate, which is no longer constant, but depends of temperature and pressure (or temperature and altitude).

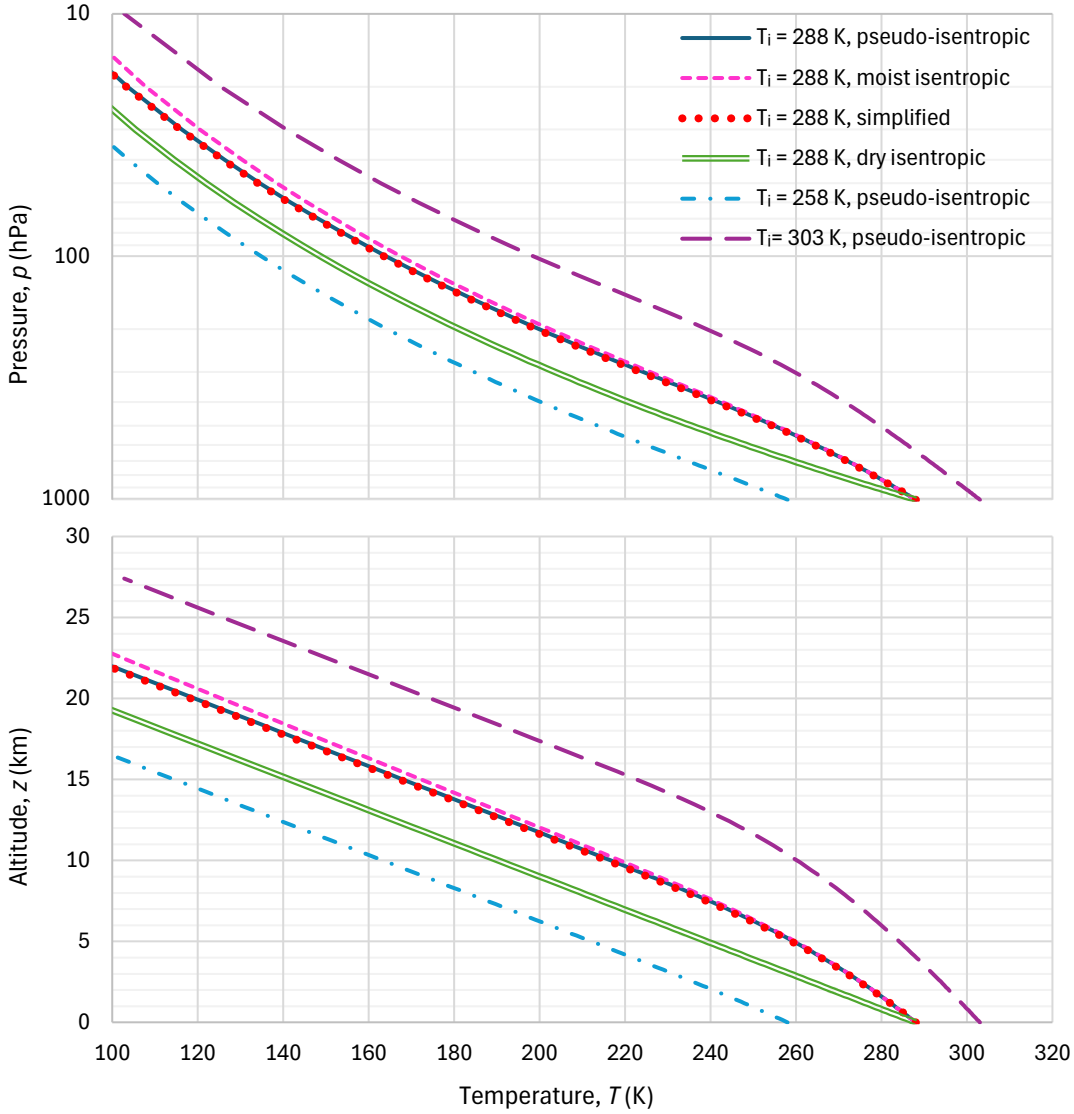

Figure 6.16 (**upper**) Illustration of isentropic changes as functions between total pressure p and temperature T that have the same total entropy per unit mass, for the indicated initial conditions defined as $p_\mathrm{i} = 1010$ hPa, T_i as shown. (**lower**) Vertical temperature profiles derived from the isentropic curves of the upper panel, after translating pressure into altitude using the hydrostatic equation. For the calculations, the saturation vapour pressure was determined from equation (6.209) for temperatures higher than 0 °C = 273.15 K and (6.211) for lower temperatures.

The results are shown in the lower panel of Figure 6.16, from which we can make the following observations:

- For low temperatures, the lapse rate $-\mathrm{d}T/\mathrm{d}z$ tends to the constant dry isotropic rate ($\Gamma_\mathrm{D} = g/c_p = 9.8$ K/km) in all cases. This is explained by the fact that low

temperatures (corresponding to high altitudes in the lower panel) do not allow water vapour to remain in gaseous phase. However, in the moist isotropic case (where the liquid water remains in the atmosphere) the lapse rate at high altitudes is slightly decreased to 9.2 K/km.

- At zero altitude, the lapse rate for the moist isotropic, pseudo-isotropic and simplified cases are the same, Γ_{M}, for a fixed surface temperature, but depend on that temperature. It can be much lower than the dry isotropic lapse rate (namely it is 4.8, 3.5 and 8.2 K/km for surface temperature of 288, 258 and 303 K, respectively).
- For low surface temperature the moist isotropic profile becomes practically indistinguishable from the dry isotropic profile, again because the low temperature does not allow much water to sublimate from the solid (ice) to gaseous phase.

6.24 Vertical profile of Earth's atmosphere

As already explained, the equilibrium temperature profile of the atmospheric is isothermal (section 6.16), but several mechanisms of change drive it outside of the equilibrium (Digression 6.H). In particular, in the troposphere the temperature drops as the altitude increases. There is a common misleading perception that this is spontaneously driven by gravitation, but, as has been shown (section 6.16), gravitation has no impact on temperature. The following thought experiment would perhaps be enough to convince anyone that temperature drop cannot be spontaneous—if it were, then we could not produce energy from this drop. The thought experiment (which can be materialized by an incredulous reader not convinced just by thought) goes like this. We start from the foot of a mountain, where the temperature is 10 °C. We store in an open container water, whose temperature will be equal to that of the environment, i.e. 10 °C. We assume that the peak of the mountain is at 2 km or more above the foot. We climb the mountain carrying a thermally insulated container. Near the peak of the mountain, the temperature will be lower , i.e. $10 - 6.5 \times 2 = -3$ °C. There the water will be frozen. We fill with ice or snow the insulated container and seal it. Returning to the base, we exploit the temperature difference of 13 °C between water in the two containers to produce work or energy—for example we may use a thermoelectric generator (TEG) to produce electricity directly from heat, without any moving parts.

In the real Earth's atmosphere, the temperature decreases with altitude in the troposphere, in the lowest layer of the stratosphere it may (or may not) remain constant, and continues to change at even higher altitudes. The typical atmospheric temperature profile is represented by the standard atmosphere adopted by the International Civil Aviation Organization (ICAO, 1993). According to the U.S. Standard Atmosphere (1976), the standard atmosphere is

> *a hypothetical vertical distribution of atmospheric temperature, pressure and density which, by international agreement, is roughly representative of year-round, midlatitude conditions.*

It is based on large inventories of observational data and is widely used for meteorological and engineering applications. The vertical profile of temperature vs. altitude and atmospheric pressure is shown in Figure 6.17, with the temperature being 15 °C or 288.15 K at the surface, linearly decreasing in the troposphere, with the (minus) gradient (lapse rate) $\Gamma := -\mathrm{d}T/\mathrm{d}z$ being 6.5 K/km up to the altitude of 11 km, and taking a constant value of 216.65 K above this up to 20 km. Then it increases with altitude in the middle and upper stratosphere, and decreases again in the mesosphere.

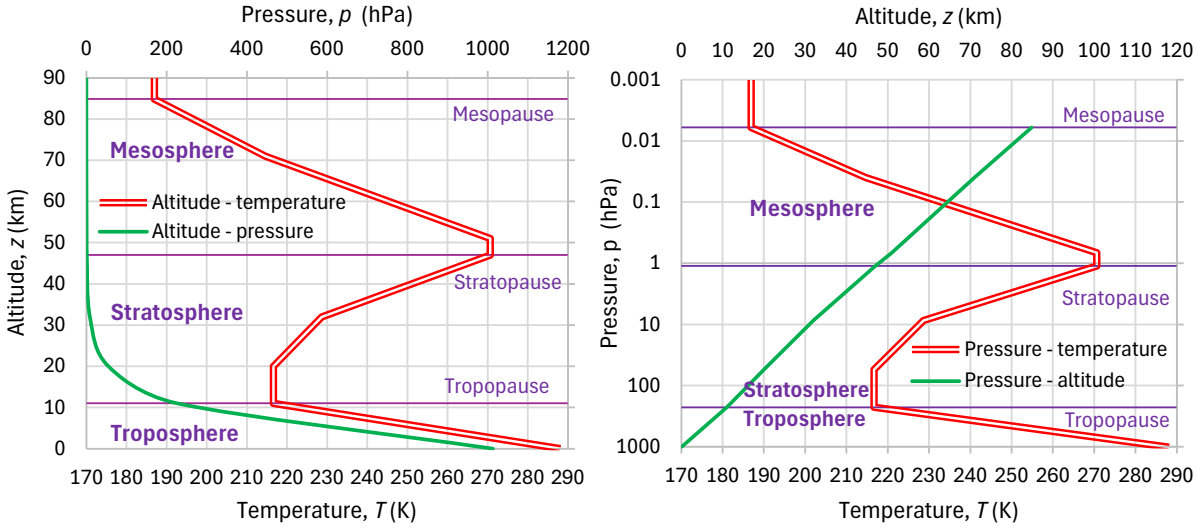

Figure 6.17 Vertical profiles of the ICAO (1993) standard atmosphere: (**left**) temperature and pressure as functions of altitude; (**right**) temperature and altitude as functions of pressure. (Source: Koutsoyiannis and Tsakalias, 2025.)

As we have seen in section 6.19 (in particular in Figure 6.11), the actual temperature profile changes with time and space and the standard atmosphere is just a representative temporal and spatial average. There have been particular local profiles for different zones of the Earth, which are in use in several applications. Of these, here we include the profiles compiled for MODTRAN (MODerate resolution atmospheric TRANsmission model), a code that performs detailed modelling of radiation in the atmosphere (Berk et al., 1987, 2008, 2014). The specific implementation we use is that of the University of Chicago, readily provided as an interactive web application. MODTRAN implements five different locality profiles, which differ in their temperature, H_2O and O_3 profiles. The profiles most relevant to this chapter are depicted in Figure 6.18.*

As inferred by Koutsoyiannis (2024), the tropical profile corresponds roughly to the equator, but is representative for the entire torrid zone (between 23.4° N or S), the midlatitude profiles (different from winter and summer) correspond to a latitude at about 45° N or S, and the subarctic profiles (again different from winter and summer) at about the latitude of polar circles (66.6° N or S).

It is useful to compare these profiles to observational data. Such data come from radiosondes, typically launched twice a day (at 00:00 UTC and 12:00 UTC) from 1300

* Additional information can be found in "The 6 model atmospheres in MODTRAN", http://modtran.spectral.com/static/modtran6/html/help_atmosphere_model.html.

upper-air stations, which are part of the Global Observing System. The point measurements of this sparce network are assimilated into meteorological models that produce the so-called reanalysis data by integrating various sources of information. These data concern a lot of meteorological variables and are given at a three-dimensional gridded basis. Here we use temperature data from the ERA5 Reanalysis on a monthly scale. ERA5 stands for the fifth generation atmospheric reanalysis of the European Centre for Medium-Range Weather Forecasts (ECMWF; ECMWF ReAnalysis). Its data are publicly available for the period 1940 onwards at a spatial resolution of 0.5°. The data sets used here were retrieved from the Physical Sciences Laboratory platform of the US National Oceanic and Atmospheric Administration (NOAA).*

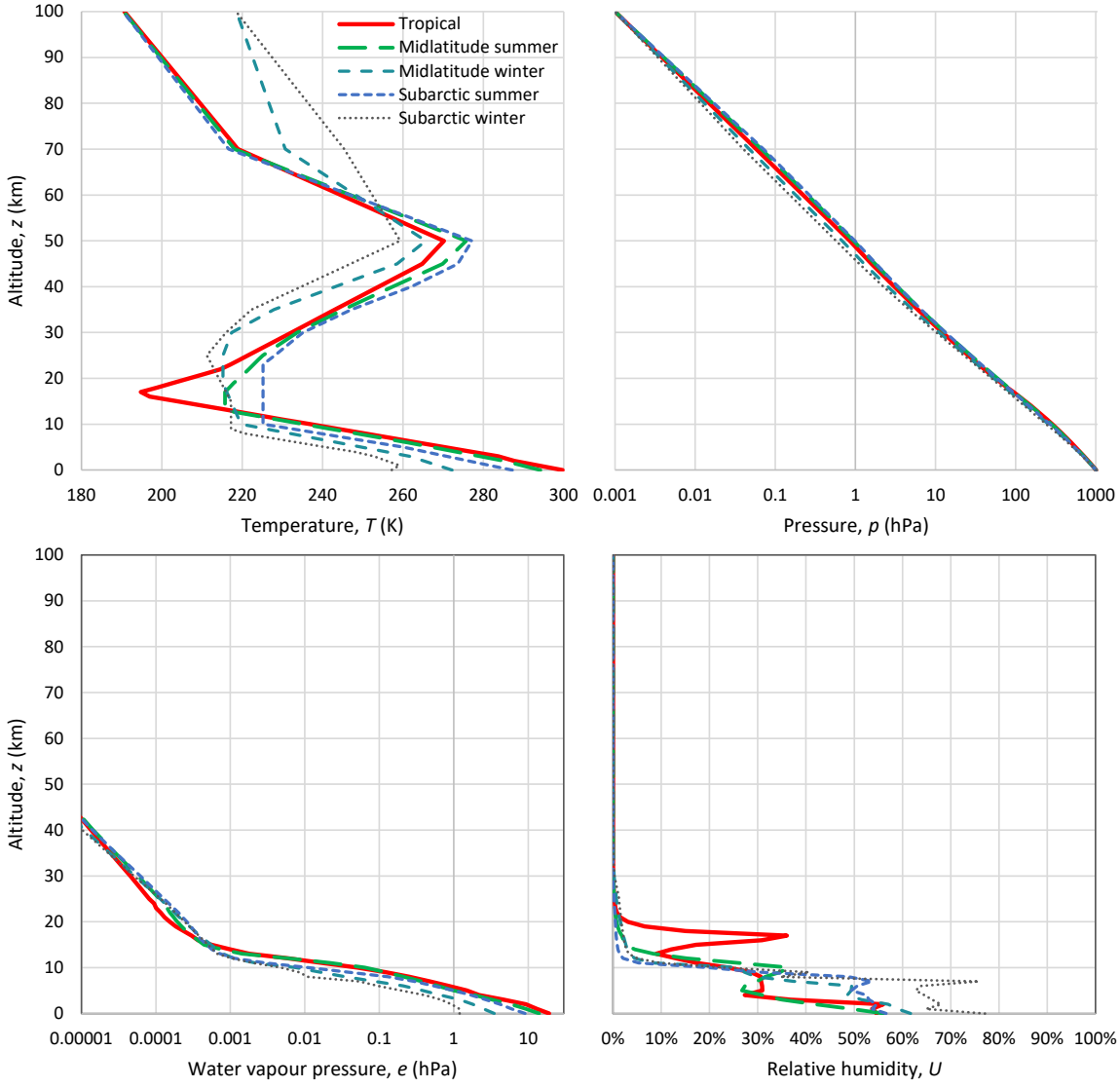

Figure 6.18 Vertical profiles of the indicated variables used in MODTRAN (Source: Koutsoyiannis, 2024).

* Platform WRIT, Monthly Timeseries, NOAA Physical Sciences Laboratory, https://psl.noaa.gov/cgi-bin/data/atmoswrit/timeseries.pl). For the period 1950 onwards, they are also accessible from the Climate Explorer (CLIMEXP) platform (https://climexp.knmi.nl/) which in addition allows extended processing of the data.

Figure 6.19 shows vertically distributed ERA5 temperature data spatially integrated over two zones, namely a tropical zone extending 7.5° N and 7.5° S, and a polar zone of width 7.5° (from 82.5° N to 90° N). The polar zone extends only over ocean and in the tropical zone the integration was made over the ocean grid points only. The vertical coordinate is the geopotential height, i.e., the vertical coordinate representing a point's height above mean sea level, adjusted for gravity variations. This is also provided by ERA5 reanalysis for each pressure level, for which temperature data are available. These observations are compared to the profile of the standard atmosphere, shifted so as to match the surface temperature (at $z = 0$), and the closest MODTRAN profile, without a temperature shift. In addition, the dry and moist isentropic profiles are plotted in the graphs.

In addition to these profiles, the isothermal profile for the average temperature is also plotted in each of the panels of Figure 6.19. To find the average temperature, we need first to define it. In absence of a standard definition, here we define it as follows: Average temperature, $\overline{T}$, of a certain body with mass m and spatially distributed density and temperature, is the temperature of a hypothetical (abstract) body with same mass and characteristics as the actual body but uniform temperature, which has heat content H equal to that of the actual body (see additional information about this in Digression 6.I).

To simplify the calculations, we make them for dry atmospheric conditions. The heat content is assumed to equal enthalpy per unit mass, as given in equation (6.86), times the mass. The total mass of the air column per unit area is

$$m = \int_0^\infty \rho(z)\mathrm{d}z \tag{6.227}$$

and the total enthalpies per unit area for the spatially distributed and the average temperature are

$$H = \int_0^\infty c_p T(z)\rho(z)\mathrm{d}z\,, \qquad H = mc_p\overline{T} \tag{6.228}$$

Thus, eliminating H we find:

$$\overline{T} = \frac{\int_0^\infty T(z)\rho(z)\mathrm{d}z}{\int_0^\infty \rho(z)\mathrm{d}z} \tag{6.229}$$

Using the hydrostatic equation ($\mathrm{d}p = \rho g d\mathrm{z}$, equation (6.67)) and replacing the infinity in the upper limit of the integral with a finite elevation (top) for which the pressure p_T is known we simplify this to

$$\overline{T} = \frac{1}{p_\mathrm{B} - p_\mathrm{T}} \int_{p_\mathrm{T}}^{p_\mathrm{B}} T(p)\mathrm{d}p \tag{6.230}$$

where p_B is the surface (bottom) pressure and for simplification we assumed constant g. The integral is numerically evaluated from the observed data.

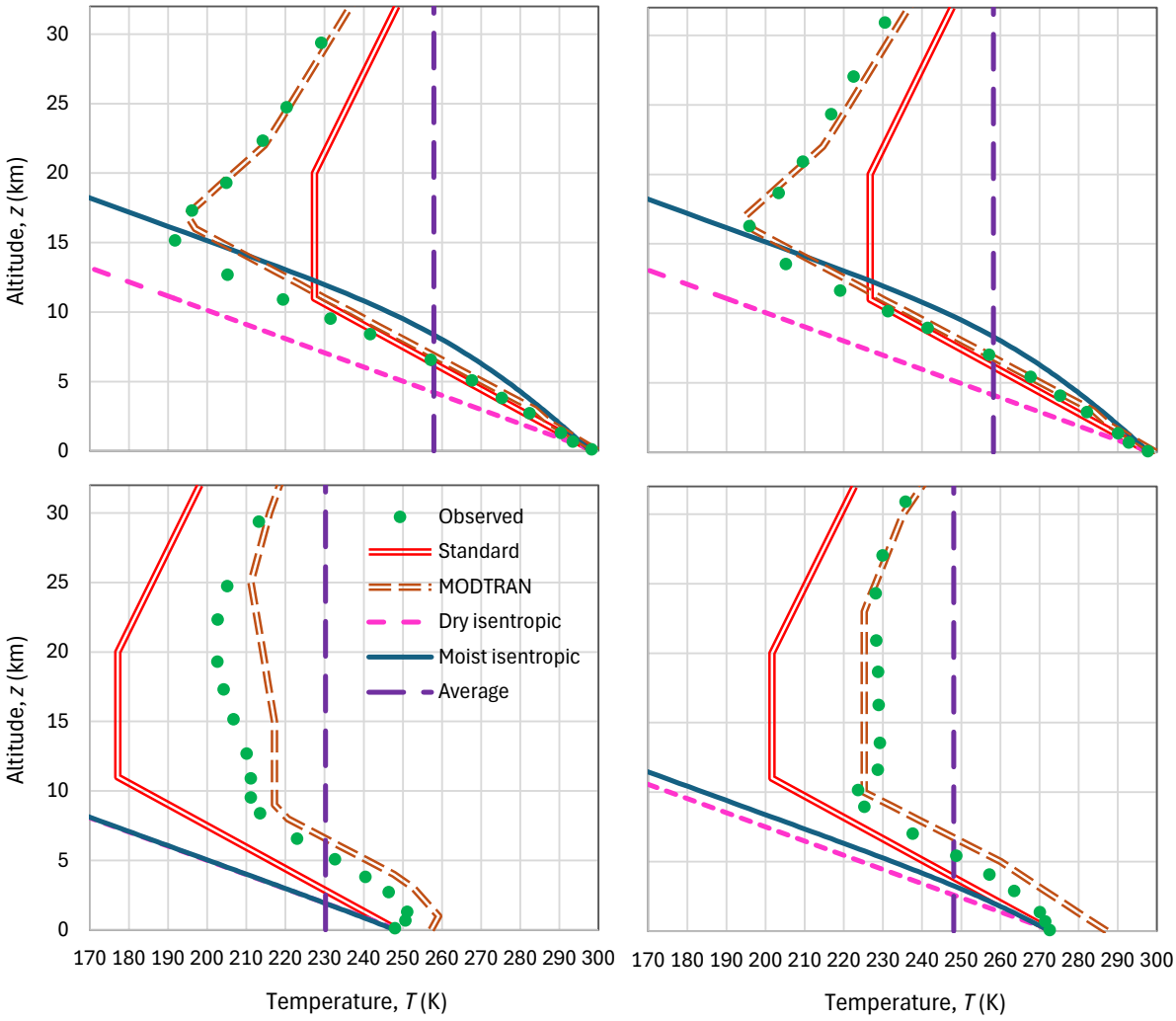

Figure 6.19 ERA5 temperature (marked as "observed") vs. ERA5 geopotential height, spatially integrated over ocean grid points at (**upper row**) a tropical zone extending ±7.5° around the equator and (**lower row**) a polar zone of width 7.5° south of the north pole, in comparison to (a) the profile of the standard atmosphere (shifted to match the temperature at $z = 0$), (b) the closest MODTRAN profile (namely the tropical profile for the upper row, and the subarctic winter and summer profile for the left and right panels of the lower row, respectively), (c, d) the dry and moist isentropic profiles, and (e) the isothermal profile with the average temperature. The temperature plots are for mean monthly scale and specifically for the months of (**left column**) February and (**right column**) August. are plotted in the graphs.Vertical profiles of the ICAO (1993) standard atmosphere: temperature and pressure as functions of altitude; (**right**) temperature and altitude as functions of pressure.

Comparison of all three figures presented above allows us to make several important observations.

1. The Earth's atmosphere is not isothermal, and the profile appears not to follow a single law.
2. The MODIS local profiles are in much better agreement with the observational data, compared to the standard atmosphere.
3. In the mesosphere, the temperature decreases with the increase of altitude, which, as explained in section 6.19, means that there is removal of heat from the

atmosphere (cooling) due to emission of longwave radiation that is not balanced by convective heat transfer from below.

4. In the stratosphere, the pattern is opposite (inversion), which implies absorption of heat by the atmosphere. This is caused by the ozone layer, which peaks at the stratosphere and absorbs the ultraviolet radiation.
5. The standard atmosphere suggests constant temperature in the lower stratosphere up to 20 km, but in fact this only happens in polar areas and midlatitudes in winter, not in the tropics and in midlatitudes during summers. The explanation for this should be sought in (a) the smaller ozone concentration below 20 km and (b) the tendency of the profile to the isothermal when the temperature is low, as explained in point 10 below.
6. The troposphere is characterized by temperature decrease with the increase of altitude, which again means removal of heat from the atmosphere (cooling). On the other hand, the troposphere is characterized by convection by ascending air, which conveys both sensible and latent heat upward. The net effect of supplied minus removed energy (radiation and heat) must be negative (net removal), as explained in section 6.19. However, this does not suffice as an explanation because the net effect is a tiny amount if compared to the huge energy fluxes occurring in the troposphere. The additional reason, which will be further discussed in section 6.25, is the tendency of convection more toward the isentropic than the isothermal profile.
7. While there is a tendency toward the isentropic profile in the troposphere, we cannot maintain that the troposphere is isentropic—whether dry or moist. This becomes clear from the polar areas where the dry and moist isentropic profiles almost coincide (because of the low presence of water vapour) with a lapse rate of 9.8 K/km. However, there, the actual rate is less than even that of the standard atmosphere of 6.5 K/km (Figure 6.19). In the tropics the actual lapse rate is close to the standard and between the lower moist isentropic rate and the higher dry isentropic rate.
8. Hence, a common conviction that it is the moist isentropic state that determines the lapse rate in the troposphere is incorrect. An additional argument showing this incorrectness is the fact that most of the time the atmosphere is not saturated and therefore the moist isentropic profile is not applicable. As seen in the humidity profile in Figure 6.18, the average relative humidity close to the surface is about 50%, far from saturated, and at higher altitudes it mostly decreases.
9. Another common conviction is that the lapse rate has a critical upper value. This goes back to Manabe and Wetherald (1967) who stated that "Free and forced convection, and mixing by the large-scale eddies, prevent the lapse rate from exceeding a critical lapse rate equal to 6.5 °C km^{-1}", but without providing a proof for their claim. This claim is also wrong, as lapse rate > 6.5 K/km are quite common and, as we have seen (Figure 6.11) even superadiabatic lapse rates (>10 K/km) are possible (albeit rare). Also in terms of long-term (climatic) averages, Figure 6.19 shows for the tropics that there are observations (at altitudes > 7 km)

lying below the curve of the standard atmosphere, which imply lapse rates > 6.5 K/km.

10. In the polar areas and even in midlatitudes in winter, the atmosphere is closer to the isothermal profile than in tropics and midlatitudes during summer. This is explained by the lower level of convection, whose complete absence would necessarily lead to an isothermal atmosphere.
11. In polar winters there is inversion in the lowest atmospheric layer, more than 1 km thick. This is explained by advection, i.e. horizontal heat transfer by wind from warmer areas.
12. The tropopause, defined to be the boundary where the air stops cooling with altitude (the troposphere) and either stabilizes or begins to warm (the stratosphere)*, is not constant at 11 km as suggested by the standard atmosphere. Rather, it is about 10 km in the polar summer and lower than that in the polar winter, and becomes 15-16 km in the tropics. Explanation for this is provided in point 10 above.

While the above points provide qualitative explanations on what is observed in vertical temperature profiles in the atmosphere, they do not suffice for quantitative modelling as they cannot support deduction, unlike the analyses of previous sections. Rather, they necessitate induction, based on data, which is the subject of Chapter 8.

We stress that these points refer to overyear annual or monthly profiles and the actual profiles at a specific time vary substantially, both geographically and temporarily, as already explained in section 6.19 (Figure 6.11). There are also long-term variations as explained in Digression 6.J.

Digression 6.I: Is average temperature meaningful?

Criticism has been expressed on the concept of global mean surface temperature on the grounds that averaging an intensive variable, such as temperature, is not physically meaningful (e.g. Essex et al., 2007). However, here we have already defined and used the column average temperature, assigning it a physical meaning. To do so, we invoked the heat contain of a gas mass, which is determined by enthalpy. This is an extensive quantity, which we converted to an intensive quantity, namely average temperature, by dividing by the mass and the specific heat. This is similar to what we have done in other intensive quantities (such as density from mass and volume, entropy per particle or unit mass from total entropy and number of particles or mass, etc.). We can hardly find anything problematic in doing so with temperature.

The same logic could be extended to globally averaged surface temperature. If we consider a layer, say, 2 m thick, of air near the surface, we can find the total enthalpy at this level and define the average temperature in a similar manner. Actually, this could be even simpler if we consider constant air density. Then it suffices to integrate the surface temperature field over the globe and divide by the globe's area. Yet is would not be too difficult to consider the variation of density (smaller in mountains etc.) and adapt the integral accordingly.

If we did not accept the legitimacy (in terms of physical meaning) of averaging, even in local regions (in terms of space, time or both), then we would be unable to use terms such as "room temperature" or "hourly temperature", or to say that a winter day is cooler than a summer day.

* A more specific operational definition by WMO (1957) is the lowest level at which the lapse rate decreases to 2 K/km or less, provided that the average lapse rate between that level and all higher levels within 2 km does not exceed 2 K/km..

And if we accept local averaging, there is no obstacle in defining a global average. Temperature measurements are instantaneous and at a point basis, and we always need to do integration for finite scales. The reason integration is needed is that there is heterogeneity, like in most quantities on Earth. The small or large heterogeneity makes no difference. For example, in hydrology we deal with quantities that vary across three or more orders of magnitude.

The scientific manner to deal with heterogeneity, whether small or large, is to use statistical indicators such as averages, variances, expectations of maxima or minima, and other statistics, most of which are either expectations or time averages, connected to each other through the notion of ergodicity (see Chapter 2 through Chapter 4). We only need to have in mind that all these are abstract quantities, not real world values, as are most concepts used in mathematics and physics.

As we will see in Chapter 8, the climate per se is by definition a stochastic concept, typically based on averaging. Without accepting the legitimacy of averaging, we would be unable to deal with climate scientifically.

All in all, constructing proper abstract concepts, pertinent to the phenomena we examine, is the only way of doing science. But their construction should be accompanied with proper definitions and fighting of ambiguity and conceptual fuzziness.

Digression 6.J: Long-term temporal changes in temperature profile

The extended (86-year long) data set used in section 6.24, allows us to explore whether or not long-term (climatic) changes appear in the temperature profiles. To investigate this we split the available records into two parts, each 43-year long. Figure 6.20 compares observed profiles for the two periods on an annual average basis, in comparison to theoretical ones.

What is more remarkable is the similarity, rather than change, of the profiles in the two periods. Yet there are some visible systematic differences, namely:

- The temperature decrease in the lower stratosphere, which is in agreement with a central assumption of Manabe and Wetherald (1967) for a warming climate.
- The warming of the surface which is small in the tropics (0.51 K) and much larger in the arctic (1.87 K).
- The small increase of the column average temperature, 0.43 K in the tropics and 0.30 K in the arctic.

The difference between the surface and the column average temperature in the arctic (1.87 – 0.43 = 1.44 K), clearly implies an increase in the temperature gradient, as also shown by Koutsoyiannis and Tsakalias (2025).

There is a general conviction in the community of climate modellers that the observed global warming would be accompanied by increase of the tropopause height, decrease of its temperature and constant lapse rate in the troposphere. However, a recent extensive analysis by Zou et al. (2023) reported that, from 2006 to 2021, ERA5 shows a warming tropical tropopause (0.10 ± 0.11 K/decade) along with a very small rise in tropopause height (50 ± 20 m/decade, which is statistically insignificant, even using classical statistics). And the recent study by Koutsoyiannis and Tsakalias (2025) showed that the lapse rate is not constant, but in both frigid zones has been increasing.

Therefore, the most remarkable long-term change is the increase of the lapse rate in the frigid zones, as reported by Koutsoyiannis and Tsakalias (2025) and confirmed in this Digression. This increase affects both the radiation regime, as it enhances the atmospheric radiation effect (see Chapter 7) and the heat transport regime as it reduces wind speed. Specifically, the increase of the minus vertical temperature gradient in the poles combined with a rather stable gradient in the tropics, weakens the surface equator-to pole temperature gradients thus resulting in weaker winds and hence weaker evaporation (Ma et al., 2025).

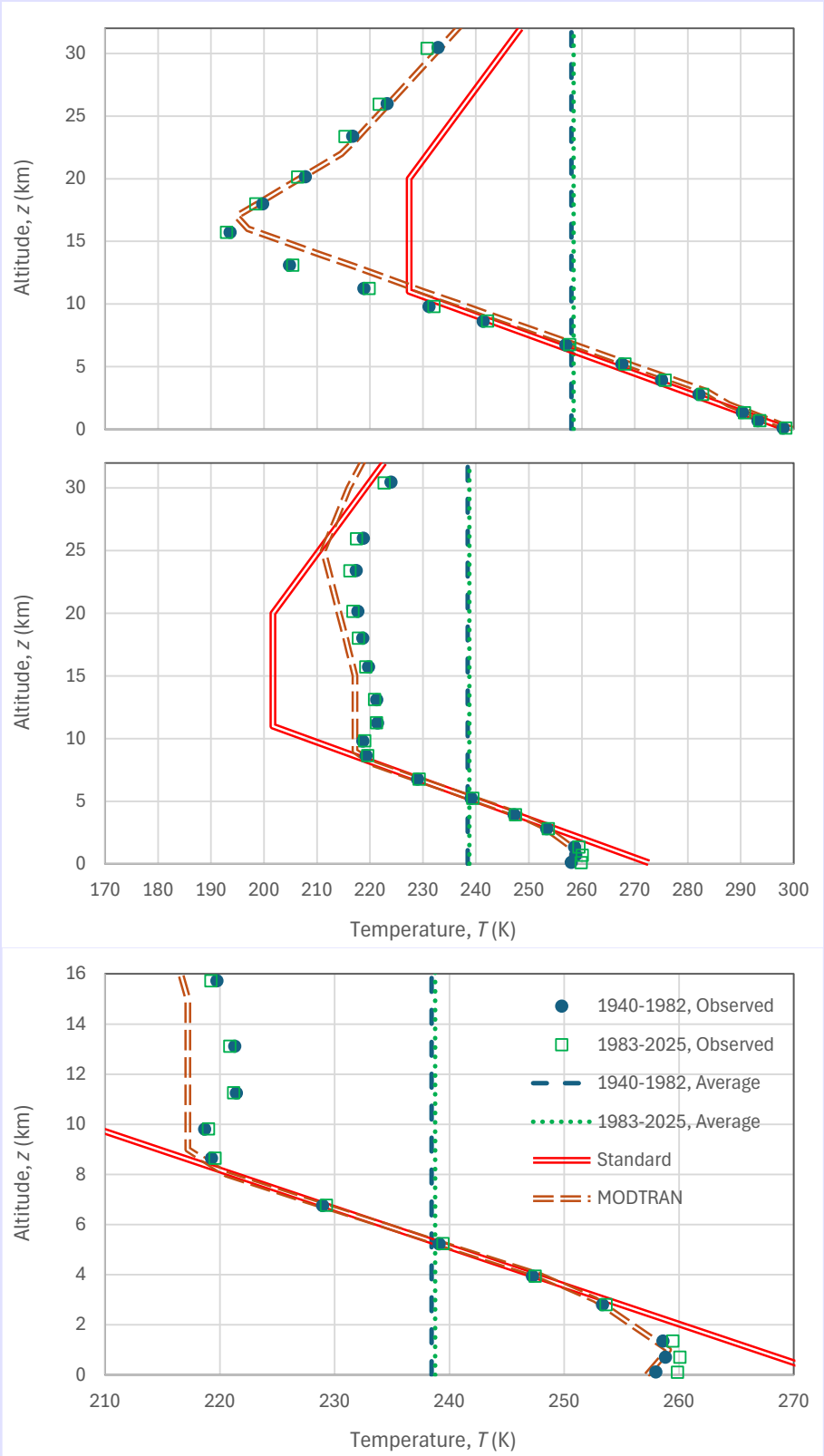

Figure 6.20 ERA5 temperature (marked as "observed") vs. ERA5 geopotential height, as in Figure 6.19, but averaged over the entire year for two periods of equal length of 43 years as indicated in the figure legend: (**upper**) for the tropical zone extending ±7.5° around the equator (**middle**) the polar zone of width 7.5° south of the north pole, and (**lower**) zoom in of the middle panel. For the polar profile (middle and lower) the shift of the standard profile was chosen so as to better fit the observed profile, rather than to match the surface temperature.

6.25 Air cells and parcels, and macroscopic stability

As our analyses showed, if we only considered the molecular motion, we would expect an atmospheric profile close to isothermal. However, there are macroscopic atmospheric structures, named cells or parcels, that move as structures, while being continually transformed in shape and physical characteristics. Their scale is much larger than the molecular and can be huge, covering large parts of the planet. They are caused by factors such as differential heating, buoyancy, and Earth's rotation. They are characterized by continuous macroscopic changes, the responsible mechanisms of which are qualitatively discussed in Digression 6.H. They are typically treated via continuum fluid dynamics, rather than molecular-level thermodynamics.

Contrary to molecules which are precisely the same for each substance, these structures differ and expand in a hierarchy of horizontal, vertical, and temporal scales. This hierarchy includes the following types of cells, where smaller cells can be components of larger scale ones:

- *Rayleigh-Bénard cells*, with horizontal and vertical scales of mm to m, and temporal scale of min. They are formed over heated fluid layers imposing buoyancy instability.
- *Small thermals* or *plumes*, with horizontal scale of 10 m to 2 km, vertical scale of 100 m to 2 km (boundary layer depth), and temporal scale of s to min. They are buoyant rising parcel of warm air from surface heating.
- *Mesoscale convection cells*, with horizontal scale of 1–50 km, vertical scale of 0.5 to 2 km and temporal scale of min to h. Again these are driven by surface heating, followed by atmospheric cooling. They are often combined with cloud formation (stratocumulus or boundary-layer clouds).
- *Deep convective cells*, with horizontal scale of 5–20 km, vertically extending up to the tropopause and with time scale of half to a few hours. They are caused by strong buoyant updrafts fuelled by latent heat release and can produce heavy rain.
- *Mesoscale convective systems*, with horizontal scale of 100 to 1000 km or more, vertically extending up to the tropopause and with time scale of hours to a day. They are clusters of deep cells acting as one system.
- *Global circulation cells*, i.e. three distinct, large-scale atmospheric circulation patterns in each hemisphere—the Hadley (equator to ~30°), Ferrel (~30 to ~60°), and polar (~60 to 90°) cells covering the entire troposphere and with time scales reaching seasonal. They are driven by solar heating and Earth's rotation, and redistribute heat from the equator to the poles, creating global wind belts and permanent pressure zones that define Earth's climatic zones.

The coexistence of molecular and macroscopic motion at multiple scales is the main reason for the departure of the actual vertical temperature profile from the isothermal. For the macroscopic motion will not necessarily retain constant temperature. To see this let us examine an air parcel that, having been warmed at the surface, moves upward. If the process is fast enough, then no heat is removed from, or added to, the parcel and hence

the process is reversible and adiabatic. Consequently, the change will be isentropic, as described by equations (6.156)-(6.157), rather than isothermal. The lapse rate Γ will then be equal to the isentropic lapse rate, the dry one, Γ_{D}, if the atmosphere is not saturated, or the moist one, Γ_{M}, if it is saturated. This isentropic lapse rate defines what we call *neutral macroscopic stability*. In comparison, microscopic (molecular) stability we have when the lapse rate is zero.

Let us now examine the macroscopic stability and instability for more general conditions. We assume that (a) the actual (environment) lapse rate is Γ, (b) at an elevation z_0 the temperature is T_0, (c) the isentropic lapse rate is that elevation is Γ_{I} and (d) an air parcel is moved slightly upward, from z_0 to $z > z_0$, in an isentropic process (adiabatically). At elevation z the parcel will have temperature $T_{\mathrm{I}} = T_0 - \Gamma_{\mathrm{I}}(z - z_0)$, while the environment will have temperature $T = T_0 - \Gamma(z - z_0)$. If $\Gamma < \Gamma_{\mathrm{I}}$, then $T > T_{\mathrm{I}}$ and, due to the ideal gas law, $\rho < \rho_{\mathrm{I}}$. Hydrostatically, the parcel, having larger density than the environment, will sink and return to its initial position. Conversely, if we moved the parcel down at $z < z_0$, it would have lower density than the environment and would rise to return again to its initial position. Hence, the case $\Gamma < \Gamma_{\mathrm{I}}$ signifies (macroscopic) stability, the opposite, $\Gamma > \Gamma_{\mathrm{I}}$ signifies instability, with equality, $\Gamma = \Gamma_{\mathrm{I}}$ signifying neutral stability. In instability, a parcel that was initially moved up will continue its motion, rather than return. If it contains moisture, at some elevation it will become saturated, and beyond this the vapour will condense, giving rise to cloud formation and perhaps precipitation.

Since we have two cases of isentropic processes, dry with lapse rate Γ_{D} and moist with lapse rate $\Gamma_{\mathrm{M}} < \Gamma_{\mathrm{D}}$, we can distinguish the following cases, if we compare these to the environmental lapse rare Γ:

- $\Gamma > \Gamma_{\mathrm{D}}$: unconditional macroscopic instability;
- $\Gamma = \Gamma_{\mathrm{D}}$: dry neutral macroscopic stability;
- $\Gamma_{\mathrm{M}} < \Gamma < \Gamma_{\mathrm{D}}$: conditional macroscopic instability;
- $\Gamma = \Gamma_{\mathrm{M}}$: saturated neutral macroscopic stability;
- $0 < \Gamma < \Gamma_{\mathrm{M}}$: unconditional macroscopic stability;
- $\Gamma = 0$: macroscopic and microscopic stability; and
- $\Gamma > 0$: macroscopic stability with inversion.

In principle, regarding entropy as quantification of uncertainty, we could maximize entropy not only at a molecular level but also at a macroscopic level. As clarified in section 2.3, entropy is a function of the probabilities of the events that form a partition. Many different partitions could be formed, which have different entropies and, when these are maximized, different values of the variables involved will be obtained. Hence, we would expect that the maximization of a coarser partition would give a different result from that of the finest partition. In our case, we would expect that entropy maximization at a level of an air parcel would result in an isentropic, rather than an isothermal, microscopic state.

Since the entropy maximization at the microscopic (molecular) level results in $\Gamma = 0$ and that in macroscopic level in $\Gamma = \Gamma_{\mathrm{I}}$, it is reasonable to expect that the actual Γ would be in between Γ and Γ_{I} and hence the standard lapse rate $\Gamma = 6.5$ K/km does not come as a surprise. But as we have already explained, we do not expect to find this by deduction,

due to the complexity of the entire climatic system and the fact that this standard value is the result of an integration over the globe. Hence it is necessary to rely on induction. Yet, in Digression 6.K we discuss a toy model, based on deduction, to illustrate the case that entropy could be defined on several scales, and its maximization gives different results depending on the scale. The toy model assumes that there are patterns shaped by different molecules, as happens in the formation of parcels, and, as explained in Digression 3.B, the typical stochastic representation of patterns is through dependence (and correlation).

Digression 6.K: Dependence and multiscale entropy maximization

Here we present the most simplified example (a toy model) for better illustration and intuition development. We consider two monatomic molecules in one-dimensional motion (so that each one has only one degree of freedom) and we are only interested in the uncertainty about the velocities $\underline{u}_i, i = 1,2$, assumed to have zero mean, and not about their position. If the mean kinetic energy is ε and the mass of each molecule is m_0, then, in a way similar to section 6.2, we infer that the expected value of the squared velocity is

$$\sigma^2 \coloneqq \mathrm{E}\left[\left\|\underline{u}_i\right\|^2\right] = \frac{2\varepsilon}{m_0}$$

and the marginal density function of the velocity is normal:

$$f(u_i) = \frac{\mathrm{e}^{-\frac{u_i^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$$

If the two velocities are independent, then their joint density function will be bivariate normal

$$f(u_1, u_2) = \frac{\mathrm{e}^{-\frac{u_1^2+u_2^2}{2\sigma^2}}}{2\pi\sigma^2}$$

The entropy of each molecule is

$$\varphi_i = -\int_{-\infty}^{\infty} \ln\left(f(u_i)u^*\right) f(u_i)\mathrm{d}u_i = \int_{-\infty}^{\infty} \left(\frac{u_i^2}{2\sigma^2} + \ln\left(\sqrt{2\pi}\frac{\sigma}{u^*}\right)\right)\frac{\mathrm{e}^{-\frac{u_i^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}\mathrm{d}u_i = \frac{1}{2} + \ln\left(\sqrt{2\pi}\frac{\sigma}{u^*}\right)$$
$$= \frac{1}{2} + \frac{1}{2}\ln\left(2\pi\left(\frac{\sigma}{u^*}\right)^2\right) = \frac{1}{2} + \frac{1}{2}\ln\frac{4\pi\varepsilon}{m_0 u^{*2}} = \frac{1}{2}\ln\frac{\varepsilon}{\varepsilon^*}$$

where the u^* is a constant with units of velocity (a constant Lebesgue density), necessary to reinstate dimensional consistency and $\varepsilon^* \coloneqq m_0 u^{*2}/4\pi\mathrm{e}$. The entropy of the two molecules is

$$\varphi = -\iint_{-\infty}^{\infty} \ln\left(f(u_1, u_2)u^{*2}\right) f(u_1, u_2)\mathrm{d}u_1\mathrm{d}u_2 = \iint_{-\infty}^{\infty} \left(\frac{u_1^2 + u_2^2}{2\sigma^2} + \ln\left(2\pi\left(\frac{\sigma}{u^*}\right)^2\right)\right)\frac{\mathrm{e}^{-\frac{u_1^2+u_2^2}{2\sigma^2}}}{2\pi\sigma^2}\mathrm{d}u_1\mathrm{d}u_2$$
$$= 1 + \ln\left(2\pi\left(\frac{\sigma}{u^*}\right)^2\right) = 1 + \ln\frac{4\pi\varepsilon}{m_0 u^{*2}} = \ln\frac{\varepsilon}{\varepsilon^*} = 2\varphi_i$$

This complies with the additivity property and is similar to what we have done in all previous analyses. The entropy per particle remains φ_i.

Now we make a key assumption, differing from all previous analyses, that there is dependence between the velocities of the two molecules, expressed by a correlation coefficient $\rho \neq 0$. The joint distribution function becomes

$$f(u_1,u_2)=\frac{\mathrm{e}^{-\frac{u_1^2+u_2^2-2\rho u_1u_2}{2(1-\rho^2)\sigma^2}}}{2\pi\sqrt{1-\rho^2}\,\sigma^2}$$

and the entropy is

$$\begin{aligned}\varphi&=-\iint_{-\infty}^{\infty}\ln\big(f(u_1,u_2)u^{*2}\big)f(u_1,u_2)\mathrm{d}u_1\mathrm{d}u_2\\&=\iint_{-\infty}^{\infty}\left(\frac{u_1^2+u_2^2-2\rho u_1u_2}{2(1-\rho^2)\sigma^2}+\ln\left(2\pi\sqrt{1-\rho^2}\left(\frac{\sigma}{u^*}\right)^2\right)\right)\frac{\mathrm{e}^{-\frac{u_1^2+u_2^2-2\rho u_1u_2}{2(1-\rho^2)\sigma^2}}}{2\pi\sqrt{1-\rho^2}\,\sigma^2}\mathrm{d}u_1\mathrm{d}u_2\\&=1+\ln\left(2\pi\sqrt{1-\rho^2}\left(\frac{\sigma}{u^*}\right)^2\right)=1+\ln\frac{4\pi\varepsilon}{m_0u^{*2}}+\frac{1}{2}\ln(1-\rho^2)\end{aligned}$$

which after algebraic manipulations becomes

$$\varphi=\ln\frac{\varepsilon}{\varepsilon^*}+\frac{1}{2}\ln(1-\rho^2)$$

Because $0\le 1-\rho^2\le 1$, the last term is negative and hence in this case $\varphi\le\varphi_1+\varphi_2$ with the equality holding when $\rho=0$. Thus, we confirm that entropy is maximized when the motions of the two molecules are independent. The entropy per molecule is now smaller than φ_i, i.e.

$$\varphi^{(1)}=\frac{1}{2}\ln\frac{\varepsilon}{\varepsilon^*}+\frac{1}{4}\ln(1-\rho^2)=\frac{1}{2}\ln\frac{\varepsilon}{\varepsilon^*}+\frac{1}{4}\ln(1+\rho)+\frac{1}{4}\ln(1-\rho)$$

But let us take another step and consider as a "unit" or "scale" in which to calculate entropy not a single molecule, but the pair. A representative quantity for the "unit" of 2 is the average velocity

$$\underline{u}^{(2)}=\frac{\underline{u}_1+\underline{u}_2}{2}$$

While in the "unit" of a molecule, the partition we work on is described as the Cartesian product of intervals $(u_1,u_1+\mathrm{d}u_1)$ and $(u_2+\mathrm{d}u_2)$), in the "unit" of the pair of molecules we have a coarsened partition of intervals $\big(u_1+u_2,u_1+u_2+\mathrm{d}(u_1+u_2)\big)$. The variable $\underline{u}^{(2)}$ has variance

$$\sigma_2^2\coloneqq\mathrm{E}\left[\left\|\underline{u}^{(2)}\right\|^2\right]=\frac{(1+\rho)\sigma^2}{2}=\frac{(1+\rho)\varepsilon}{2m_0}$$

normal distribution with density function:

$$f\big(u^{(2)}\big)=\frac{\mathrm{e}^{-\frac{u^{(2)}}{(1+\rho)\sigma^2}}}{\sqrt{\pi(1+\rho)}\sigma}$$

and entropy

$$\begin{aligned}\varphi^{(2)}\coloneqq\Phi[\underline{u}^{(2)}]&=\frac{1}{2}+\frac{1}{2}\ln\left(2\pi\left(\frac{\sigma_2}{u^*}\right)^2\right)=\frac{1}{2}+\frac{1}{2}\ln\left(2\pi\left(\frac{\sigma_2}{u^*}\right)^2\right)\\&=\frac{1}{2}+\frac{1}{2}\left(\ln\frac{4\pi\varepsilon}{m_0u^{*2}}-\ln 4+\ln(1+\rho)\right)=\frac{1}{2}+\frac{1}{2}\ln\frac{4\pi\varepsilon}{m_0u^{*2}}+\frac{1}{2}\ln\frac{1+\rho}{4}\end{aligned}$$

This finally yields

$$\varphi^{(2)}=\frac{1}{2}\ln\frac{\varepsilon}{\varepsilon^*}+\frac{1}{2}\ln\frac{1+\rho}{4}$$

In summary for dependent particles the entropy per a "unit" (or "scale") of 1 and 2 particles will be, respectively

$$\varphi^{(1)} = \frac{1}{2}\ln\frac{\varepsilon}{\varepsilon^*} + \frac{1}{4}\ln(1+\rho) + \frac{1}{4}\ln(1-\rho), \qquad \varphi^{(2)} = \frac{1}{2}\ln\frac{\varepsilon}{\varepsilon^*} + \frac{1}{2}\ln\frac{1+\rho}{4}$$

Clearly, the former is maximized when $\rho = 0$ and the latter when ρ = 1. The two become equal for $\rho = 15/17$.

If, trying to find a balance between the two "scales", we may take a weighted average for the two assuming weights w and $1-w$, respectively. The contribution of the correlation to the weighted average entropy (actually a negative quantity) is

$$\varphi_\rho := w\varphi^{(1)} + (1-w)\varphi^{(2)} = \left(\frac{1}{2} - \frac{w}{4}\right)\ln(1+\rho) + \frac{w}{4}\ln(1-\rho) - (1-w)\ln 2$$

This is maximized when $\mathrm{d}\varphi_\rho/\mathrm{d}\rho = 0$, which results in

$$\rho = 1 - w$$

and maximum contribution to entropy

$$\varphi_\rho^{\max} = \frac{1+\rho}{4}\ln\frac{1+\rho}{4} + \frac{1-\rho}{4}\ln\frac{1-\rho}{4} + (1-\rho)\ln 2$$

Graphical illustration of the above results is provided in Figure 6.21.

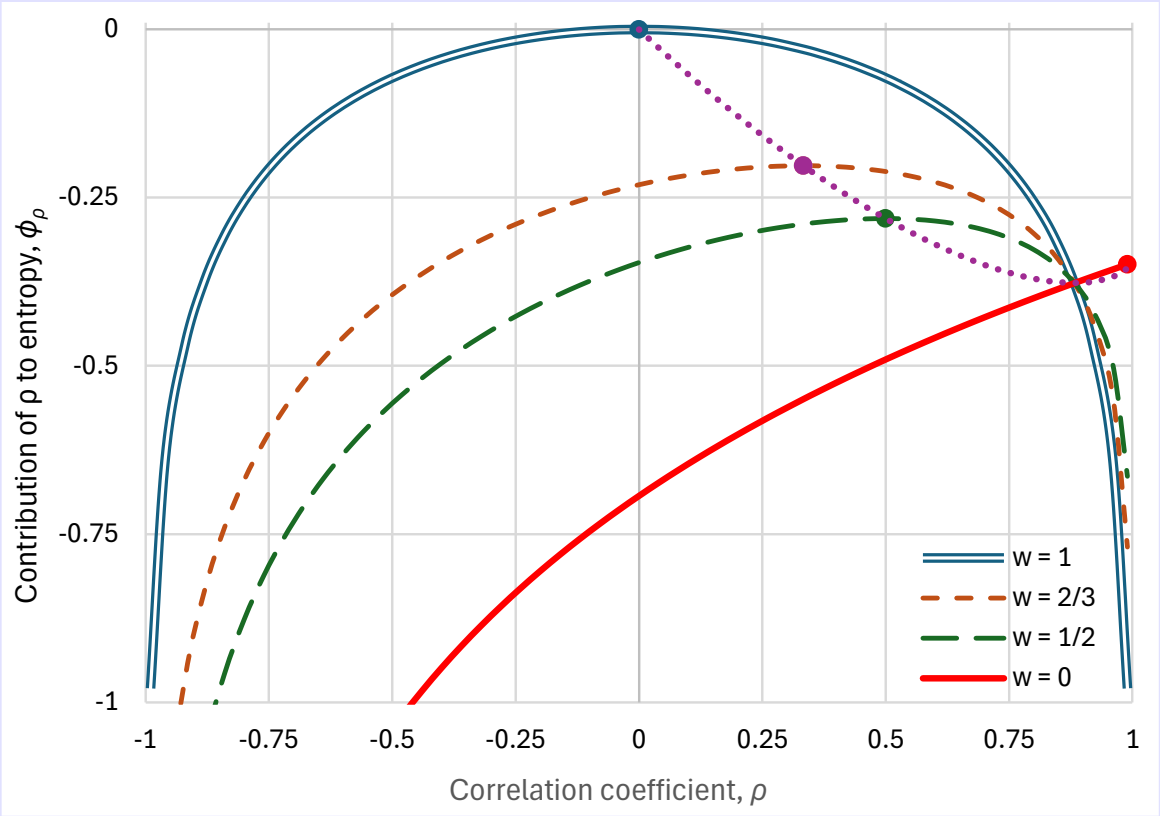

Figure 6.21 Contribution to the entropy, φ_ρ, of the correlation between to particles, ρ, as a function of ρ and the weight w given to the entropy of a single particle. The dots represent the points in which φ_ρ is maximized and the dotted line, depicting $\varphi_\rho^{\max}$, connects these points.

The above toy model is provided for exploratory purposes and does not aim to support modelling of the atmosphere. The main idea illustrated by the toy model is that there is no unique "unit" or "scale" in Nature. When the physical mechanisms act on multiple scales, or the system composition includes units of different sizes, it would not be prudent to adhere to a unique scale or unit, on which to maximize entropy. As the above example shows, maximizing entropy on one scale reduces it at another scale. There cannot be simultaneous maximization at all scales. Rather a balance on many scales would better be sought. It is reminded that entropy means uncertainty,

and entropy maximization results in maximum uncertainty. If this was applied to a small scale, then we would have small uncertainty at larger scales. But this does not seem to be the way Nature works. The uncertainty tends to be as high as possible on any scale, small or large. Hence, Nature seems to "sacrifice" some uncertainty on one scale in order to increase uncertainty on other scales.

While this idea does not serve modelling in our context, there have been modelling applications based on it in different contexts. Koutsoyiannis (2005b) successfully explained the Hurst-Kolmogorov behaviour of hydrometeorological processes (including the long-range dependence seen in temperature time series) by maximizing the averaged entropy on multiple time scales, giving equal importance to each of the time scales. In a similar approach, Koutsoyiannis (2006) modelled the intermittent character of rainfall, determining entropy at each time scale from the probability of rainfall occurrence at this scale. Maximizing entropy at all scales, again with equal importance to each one, he was able to find the probability of occurrence at any time scale. In the model calibration, the probabilities of rainfall occurrence at only two time scales were used, while for all other scales they were predicted by the multi-scale entropy maximization framework. The agreement with data was impressive, as can be seen in Figure 6.22, reproduced from that study.

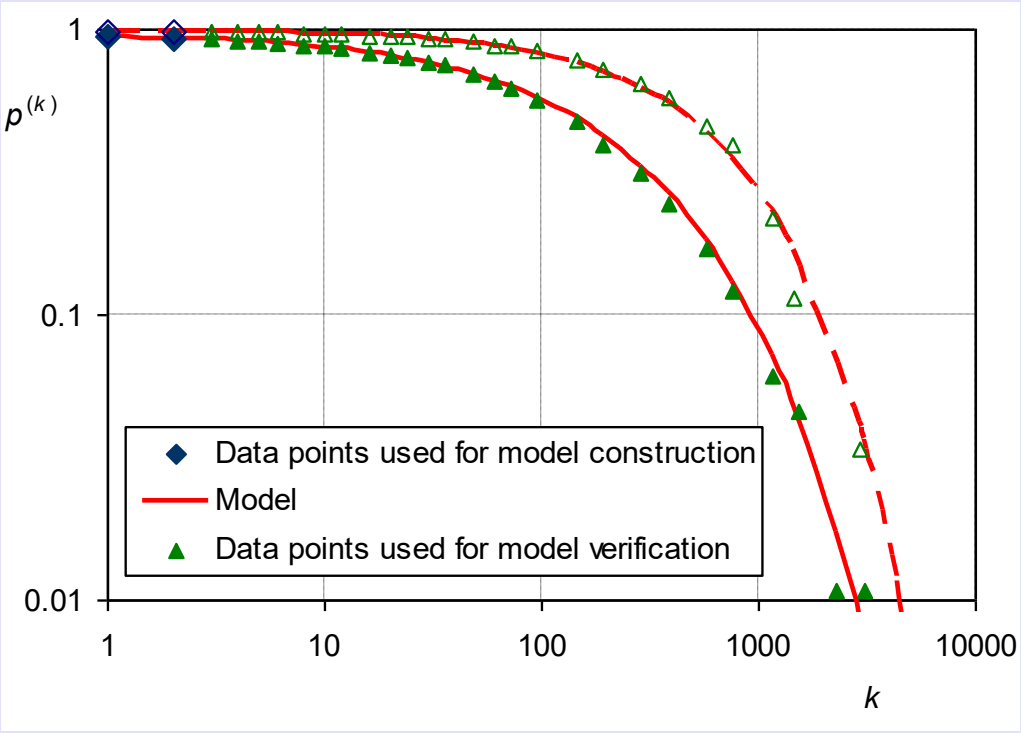

Figure 6.22 Probability dry $p^{(k)}$ vs. scale k (in hours), as estimated from the Athens rainfall data set and predicted by a multi-scale entropy maximization framework for the entire year (full triangles and full line) and the dry season (empty triangles and dashed line). (Source: Koutsoyiannis, 2006.)

6.26 Principles of thermodynamics

In the above presentation we have followed the Aristotelian suggestion for parsimony (section 1.3), i.e., to use "fewer postulates or hypotheses or propositions". Namely, to derive the basic results of atmospheric thermodynamics in Chapter 6 we have only used (a) Newton's laws, (b) the principles of conservation of mass, momentum and energy, and (c) the stochastic principle of maximum entropy. We did not use the so-called four principles (or laws) of thermodynamics. Nor did we use other assumptions typically postulated in statistical physics, such as the equiprobability of microstates (e.g., Moore,

2002) or the principle of equipartition of energy into the available degrees of freedom (often referred to as a postulate). Rather, we saw that all these are results of entropy maximization.

For the completeness of the presentation, in this section we summarize some information about the four laws of thermodynamics and discuss their relationship with our results.

The *Zeroth Law* states that if two systems are in thermal equilibrium with a third system, then they are in thermal equilibrium with each other. This law defines the notion of thermal equilibrium. In turn, this is necessary to define temperature as two systems that are in equilibrium have the same temperature.

In our framework this law is unnecessary. Two systems are in equilibrium if they are put in contact and the entropy of the compound system has been maximized. Besides, the temperature is defined through equation (6.50). As we have seen, equation (6.99) implies that two systems put in contact, in which entropy has been maximized, will have the same temperature. This is a consequence of entropy maximization and does not presuppose an axiomatic introduction of the Zeroth Law.

Moreover, the usefulness of the law is problematic. As we have seen, there appear gradients in temperature (see sections 6.19 and 6.24) and therefore, different layers of the atmosphere, however thin, do not have the same temperature, even though they are in contact. This does not violate the law per se, because the gradients imply non-equilibrium conditions, but reduce its usefulness as most natural systems are not in equilibrium.

The *First Law* is related to the conservation of energy. We have already used several times the principle of conservation of energy, according to which energy can be neither created nor destroyed but can only change forms. In classical thermodynamics, this principle is typically stated as: the heat supplied to a system (δQ) equals the increase in internal energy of the system ($\mathrm{d}E$) plus the work done by the system (δW). This version of the principle of conservation of energy has also been produced by our framework in section 6.11.

In classical thermodynamics, the *Second Law* states that in the process of reaching a thermodynamic equilibrium, the total entropy of a system increases, or at least does not decrease. In this respect, the Second Law is none other than the principle of maximum entropy applied to thermodynamic systems. Starting from any condition and approaching the equilibrium, the entropy change $\mathrm{d}\Phi^*$ of the total system cannot be negative.

Finally, the *Third Law* of thermodynamics states that the entropy of a system approaches zero as the temperature approaches absolute zero. Here we did not need this law at all. Yet, we were able to replicate it in Digression 6.B within a negligible difference for the value of energy in which the entropy becomes zero. In our framework the entropy-zeroing energy is not zero but a super-tiny quantity, smaller than the quantum zero-point energy that is imposed by Heisenberg's uncertainty principle. Thus, our stochastic framework resulted in a version of the Third Law that is in better agreement with quantum physics than the classical version which implies that zero energy is in principle feasible.

Appendix 6-I: The Lambert W function

The Lambert W function (or ω-function) is defined to be the inverse of the function $z = f(w) \coloneqq we^{w}$. This function denoted as $W_k(z) \coloneqq f^{-1}(z)$, which thus satisfies $W_k(z)\mathrm{e}^{W_k(z)} = z$, is a multivalued function on the complex plane if the integer k is not specified, or single valued if k is specified. For real z and $-1/\mathrm{e} \leq z < 0$, there are two possible real values of $W_k(z)$, denoted as $W_0(z)$, and $W_{-1}(z)$. If $z \geq 0$, there is a single real value $W_0(z)$. The real-valued branch $W_0(z)$ of $W_k(z)$ satisfying $W_0(z) \geq -1$ is called the principal branch of the W function, and the other real-valued branch, $W_{-1}(z)$, which satisfies $W_{-1}(z) \leq -1$, is the non-principal real branch.

Here we deal with the non-principal real branch $W_{-1}(z)$ only. Approximations for this, even for one-shot evaluation, can be found in Chapeau-Blondeau and Monir (2002), Barry et al. (2004) and Chatzigeorgiou (2013).

The function $W_k(z)$ is available for direct use in most computational environments.* For values of z that are relevant to our particular problem (calculations related to the quantification of water vapour in the atmosphere), in addition to the approximations found in literature, we propose the following, which is very accurate and fast:

$$-W_{-1}(z) = 1.285(-\ln(-z))^{0.933} + 0.872(\ln(-\ln(-z)))^{0.612} \tag{6.231}$$

The relative error is negligible, smaller than 3×10^{-5} for the values relevant to our calculations, i.e., those corresponding to the range of temperature shown in the figures of section 6.21 (for which $-0.07 \leq z \leq -0.015$). Notice the minus sign in $-W_{-1}(z)$, which makes this quantity positive.

* In Mathematica and Maple, which perform both symbolic and numerical calculations, the function is named ProductLog and LambertW, respectively. With the latter name, it is also available in R (https://cran.r-project.org/web/packages/LambertW/index.html), MATLAB (https://www.mathworks.com/help/symbolic/lambertw.html), Python (https://docs.scipy.org/doc/scipy/reference/generated/scipy.special.lambertw.html), etc., while several functions implementing it are available online; e.g. for Excel (https://www.vbforums.com/attachment.php?attachmentid=89337&d=1341009088) and for LibreOffice (https://gist.github.com/m93a/a0199c4f40b43bb8116810daa46dd92d).

Chapter 7. Radiation in the atmosphere

7.1 The Planck blackbody radiation formula

[to be written in a next version]

7.2 The Stefan-Boltzmann law

[to be written in a next version]

7.3 Factors affecting atmospheric radiation

[to be written in a next version]

7.4 Linking atmospheric radiation with atmospheric thermodynamics

[to be written in a next version]

7.5 Is the atmosphere a greenhouse?

[to be written in a next version]

Digression 7.A: Quantification of the radiation changes modifying the temperature profile

7.6 Relative importance of radiatively active gases and clouds

[to be written in a next version]

7.7 ???

[to be written in a next version]

Chapter 8. Geophysical processes and stochastic induction

8.1 Introduction

[to be written in a next version]

8.2 The stochastic definition of climate

[to be written in a next version]

8.3 Multi-scale analysis of time series

[to be written in a next version

8.4 Inspecting causality in geophysical processes

[to be written in a next version]

8.5 Carbon cycle and residence times

[to be written in a next version]

8.6 Carbon isotopes and their relevance in the atmosphere and climate

[to be written in a next version]

8.7 Evaporation and hydrological cycle

[to be written in a next version]

8.8 ???

[to be written in a next version]

Chapter 9. Epilogue

9.1 Physics is stochastics

[to be written in a next version]

9.2 Entropy in fields other than physics

[to be written in a next version]

9.3 Physics is not enough: The decisive role of the biosphere

[to be written in a next version]

References

Adams, H. 1918. *The Education of Henry Adams*. Houghton Mifflin Company, Boston, MA, USA.

Alduchov, O.A., and Eskridge. R.E., 1996. Improved Magnus form approximation of saturation vapour pressure. *J. Appl. Meteor.*, 35, 601–609.

Ambaum M.H.P., 2020. Accurate, simple equation for saturated vapor pressure over water and ice. *Q. J. R. Meteorol. Soc.*, 146, 4252–4258. doi:10.1002/qj.3899

Ananthaswamy, A., 2019. *Through Two Doors at Once: The Elegant Experiment That Captures the Enigma of Our Quantum Reality*. Dutton, New York, USA.

Apostol, T.M., 1967. One-variable calculus, with an introduction to linear algebra (Vol. 1). Waltham, Toronto, London: Blaisdell.

Atkins, P., 2003. *Galileo's Finger: The Ten Great Ideas of Science*, Oxford University Press: New York, NY, USA.

Atkins, P. 2007. *Four Laws that Drive the Universe*, Oxford Univ. Press, Oxford, 131 pp.

Bailey, A.L., 1929. A summary of advanced statistical methods. US Statistical Bureau, https://babel.hathitrust.org/cgi/pt?id=mdp.39015067991326&view=1up&seq=7.

Bailey, K.D. 2009. *Entropy Systems Theory: Systems Science and Cybernetics*. Eolss Publishers, Oxford, UK, 169 pp.

Bar-Hillel, M. and Falk, R., 1982. Some teasers concerning conditional probabilities. *Cognition*, 11(2), 109-122.

Barnes, F.B., 1954. Storage required for a city water supply. *J. Inst. Eng., Australia*, 26, 198-203.

Barnett, V., 2006. Chancing an interpretation: Slutsky's random cycles revisited. *Euro. J. History of Economic Thought*, 13(3), 411-432, doi: 10.1080/09672560600875596.

Barry, D.A., Li, L., and Jeng, D.S., 2004. Comments on "Numerical evaluation of the Lambert W function and application to generation of generalized Gaussian noise with exponent ½". *IEEE Transactions on Signal Processing*, 52 (5), 1456-1457.

Beard, L.R., 1965. Use of interrelated records to simulate streamflow. *Proc. Am. Soc. Civil Eng., J. Hydraul. Div.*, 91(HY5), 13-22.

Beck, M., 2012. *Quantum Mechanics: Theory and Experiment*. Oxford University Press, New York, USA.

Ben-Naim, A., 2008. *A Farewell to Entropy: Statistical Thermodynamics Based on Information*, World Scientific Pub., Singapore, 384 pp.

Berk, A., Bernstein, L.S., and Robertson, D.C., 1987. *MODTRAN: A Moderate Resolution Model for LOWTRAN*. Scientific Report No. 1; Air Force Geophysics Laboratory, Air Force Systems Command, United States Air Force: Hanscom Air Force Base, Massachusetts, USA, https://apps.dtic.mil/sti/pdfs/ADA185384.pdf

Berk, A., Acharya, P.K., Bernstein, L.S., Anderson, G.P., Lewis, P., Chetwynd, J.H., and Hoke, M.L., 2008. *Band model method for modeling atmospheric propagation at arbitrarily fine spectral resolution*, U.S. Patent #7433806.

Berk, A., Conforti, P., Kennett. R., Perkins, T., Hawes, F., and van den Bosch, J., 2014. *MODTRAN6: A Major Upgrade of the MODTRAN Radiative Transfer Code*. Proc. SPIE, 9088, 90880H, doi:10.1117/12.2050433.

Bernoulli, J., 1713. Ars Conjectandi, Opus Posthumum. Accedit Tractatus de Seriebus Infinitis, et Epistola Gallicé Scripta de Ludo Pilae Reticularis. Basileae, Impensis Thurnisiorum, Fratrum [English translation: *The art Of Conjecturing, Together With Letter to a Friend on Sets in Court Tennis*. Translated with an Introduction and Notes by Edith Dudley Sylla; The Johns Hopkins University Press, Baltimore, Maryland, USA, 2006.]

Birkhoff, G. D., 1931. Proof of the ergodic theorem. *Proc. Nat. Acad. Sci.*, 17, 656–660.

Boltzmann, L., 1872. Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen, in Sitzungsberichte Akad. Wiss., Vienna, part II, 66, 275-370; reprinted in Boltzmann's Wissenschaftliche Abhandlungen, Vol. I, Leipzig, J. A. Barth, 1909, 316-402.

Boltzmann, L., 1884/85. Über die Eigenschaften monocyklischer und anderer damit verwandter Systeme. *Crelles J.*, 98, 68–94.

Boltzmann, L. 1877. Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht. *Wien. Ber.*, 76, 373-435.

Boltzmann, L., 1896/1898. Vorlesungen über Gastheorie, J.A. Barth, Leipzig, Germany. (English translation: Lectures on Gas Theory, University of California Press, Berkeley, Ca. USA, 1964.)

Boltzmann, L., 1897. On the indispensability of atomism in natural science. *Annal. Phys. Chem.*, 60, 231, doi: 10.1007/978-94-010-2091-6_5.

Boltzmann, L., 1901. On the necessity of atomic theories in physics. *Monist*, *12,* 65–79. https://www.jstor.org/stable/27899285.

Boltzmann, L., 2003. Further studies on the thermal equilibrium of gas molecules. In The kinetic theory of gases: an anthology of classic papers with historical commentary (pp. 262-349).

Bortkiewicz, L., 1917. *Die Iterationen — Ein Beitrag zur Wahrscheinlichkeitstheorie*. Springer, Berlin.

Box, G.E., and Jenkins, G.M., 1970. *Time Series Models for Forecasting and Control*. Holden Day, San Francisco, USA.

Brillouin, L., 1949. Life, thermodynamics, and cybernetics. In *Maxwell's Demon. Entropy, Information, Computing*, Princeton University Press: Princeton, NJ, USA, 89–103, doi: 10.1515/9781400861521.

Brillouin, L., 1950. Thermodynamics and information theory. *Am. Sci.*, 38, 594–599. https://www.jstor.org/stable/27826339.

Brissaud, J.-B., 2005. The meanings of entropy. *Entropy*, 7, 68-96, doi: 10.3390/e7010068.

Bureau International des Poids et Mesures, 2019. *The International System of Units (SI)*. 9th edition, https://www.bipm.org/documents/20126/41483022/SI-Brochure-9-EN.pdf.

Chaitin, G.J., 2002. Computers, paradoxes and the foundations of mathematics. *Am. Sci.*, *90*, 164–171.

Chapeau-Blondeau, F. and Monir, A., 2002. Numerical evaluation of the Lambert W function and application to generation of generalized Gaussian noise with exponent 1/2. *IEEE Transactions on Signal Processing*, 50 (9), 2160-2165.

Chatzigeorgiou, I., 2013. Bounds on the Lambert function and their application to the outage analysis of user cooperation. *IEEE Communications Letters*, 17 (8), 1505-1508.

Chow, V.T., 1969. Stochastic analysis of hydrologic systems. Final Report, University of Illinois. Urbana, Illinois, https://www.ideals.illinois.edu/bitstream/handle/2142/90345/Chow_1969.pdf?sequence=2.

Chow, V.T., and Kareliotis, S.J., 1970. Analysis of stochastic hydrologic systems. *Water Resources Research*, 6(6), 1569-1582.

Clausius, R., 1865. Ueber verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie, Annalen der Physik, 201 (7), 353–400, doi: 10.1002/andp.18652010702.

Clausius, R., 1867. *The Mechanical Theory of Heat: With Its Applications to the Steam-Engine and To the Physical Properties of Bodies*. J. van Voorst, London, http://books.google.gr/books?id=8LIEAAAAYAAJ.

Clausius, R., 1872. A contribution to the history of the mechanical theory of heat. *Phil. Mag.*, 43, 106–115.

Courtney, A., and Courtney, M. 2008. Comments Regarding "On the Nature of Science". *Physics in Canada*, 64(3), 7-8.

Curry, J.A. and Webster, P.J., 2011. Climate science and the uncertainty monster. *Bull. Am. Meteorol. Soc.*, 92, 1667–1682.

Darrow, K.K., 1944. The concept of entropy. *Am. J. Phys.*, *12*, 183, doi: 10.1119/1.1990592.

Davison, A.C., and Huser, R., 2015. Statistics of Extremes. *Annual Review of Statistics and Its Application* 2 (1), 203-235.

Dechant, A., and Lutz, E., 2015. Wiener-Khinchin theorem for nonstationary scale-invariant processes. *Physical Review Letters* 115 (8), 080603.

Dimitriadis, P., and Koutsoyiannis, D., 2015. Climacogram versus autocovariance and power spectrum in stochastic modelling for Markovian and Hurst–Kolmogorov processes. *Stochastic Environmental Research & Risk Assessment*, 29 (6), 1649–1669, doi: 10.1007/s00477-015-1023-7.

Dimitriadis, P., Koutsoyiannis, D., and Tzouka, K., 2016. Predictability in dice motion: how does it differ from hydrometeorological processes?. *Hydrological Sciences Journal*, 61 (9), 1611–1622, doi:10.1080/02626667.2015.1034128.

Driessen, A., 2019. Aristotle and the foundation of quantum mechanics. *philsci-archive*, http://philsci-archive.pitt.edu/16265/.

Essex, C., McKitrick, R. and Andresen, B., 2007. Does a global temperature exist?. *Journal of Non-Equilibrium Thermodynamics*, 32 (1), 1-27.

Ewing, J., 1920. LXII. The specific heat of saturated vapour and the entropy-temperature diagrams of certain fluids. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 39 (234), 633–646.

Feynman, R., 1985. *The Character of Physical Law*. 12th printing, MIT press, Cambridge, Mass., USA.

Fischer, H., 2010. *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. Springer, New York, USA.

Fisher, R.A., and Tippett, L.H.C., 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Math. Proc. Camb. Phil. Soc.*, 24(2), 180–190, doi: 10.1017/S0305004100015681.

Fraundorf, P., 2003. Heat capacity in bits. *American Journal of Physics*, 71, 1142–1151

Fréchet, M., 1927. Sur la loi de probabilité de l'écart maximum, *Ann. Soc. Polon. Math.*, 6, 93.

Gaertner, H.-M., 2003. Huygens' Principle: A case against optimality. *Behavioral and Brain Sciences*, 26, 779–781.

Galton, F., 1890. Dice for statistical experiments. *Nature*, 42, 13–14.

Gauch, H.G., Jr., 2003. *Scientific Method in Practice*. Cambridge University Press, Cambridge.
Gibbs, J.W., 1902. *Elementary Principles in Statistical Mechanics*. Yale University Press, New Haven, Connecticut. (Reprinted by Dover, New York, 1960 and made available online: https://www.gutenberg.org/ebooks/50992).
Gnedenko, B., 1943. Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.*, 44(3), 423–453, doi: 10.2307/1968974.
Gnedenko, B.V., and Kolmogorov, A.N., 1949. Predelnye raspredelniya dlya summ nezavisimykh sluchainykh velichin (Limit distributions for sums of independent random variables), Gostekhizdat, Moscow-Leningrad, 1949 (in Russian).
Goldstein, H., Poole, C., and Safko, J.L., 2002. *Classical Mechanics*. Addison Wesley, San Francisco, USA.
Graham, L., 2011, The Power of Names. *Theology and Science*, 9 (1), 157-164, doi: 10.1080/14746700.2011.547020.
Graham, L. and Kantor, J.-M., 2009. *Naming Infinity: A True Story of Religious Mysticism and Mathematical Creativity*. Harvard University Press.
Gumbel, E.J., 1958. *Statistics of Extremes*, Columbia Univ. Press, New York, USA.
Guye, C.-E., 1922. *L'Évolution Physico-Chimique*. Etienne Chiron, Paris, France.
Havrda J., and Charvát F., 1967. Concept of structural a-entropy. *Kybernetika*, 3, 30–35.
Hazen, A., 1914. The storage to be provided in impounding reservoirs for municipal water supply. *Transactions of the American Society of Civil Engineers*, 77, 1539-1669.
Hecht, H., 2016. Gottfried Wilhelm Leibniz and the origin of the principle of least action – a never ending story. *Annalen der Physik*, 528 (9-10), 623–635, doi: 10.1002/andp.201600239.
Heisenberg, W., 1962. The development of the interpretation of the quantum theory. In *Niels Bohr and the Development of Physics, Essays Dedicated to Niels Bohr on the Occasion of his Seventieth Birthday*, edited by W. Pauli, 2nd edition, Pergamon Press, New York, 12-29, https://archive.org/details/nielsbohrdevelop0000paul/.
Hemelrijk, J., 1966. Underlining random variables. *Statistica Neerlandica*, 20 (1), 1-7.
Herbertson, A.J., 1907. *Outlines of Physiography, an Introduction to the Study of the Earth*, Arnold, London, UK.
Hill, G., and Holman, J., 1986. *Entropy–The Driving Force of Change*; Royal Society of Chemistry, London, UK, 17 pp., ISBN: 0-15186-967-X.
Hosking, J.R.M.,1990. L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society, Series B*, 52, 105–124.
Hosking, J.R.M., 1994. The four-parameter kappa distribution. *IBM Journal of Research and Development*, 38(3), 251–258, doi: 10.1147/rd.383.0251.
Hosking, J.R., Wallis, J.R. and Wood, E.F., 1985a. Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3), 251-261.
Hosking, J.R., Wallis, J.R. and Wood, E.F., 1985b. An appraisal of the regional flood frequency procedure in the UK Flood Studies Report. *Hydrological Sciences Journal*, 30(1), 85-109.
Hurst, H.E., 1951. Long term storage capacities of reservoirs. *Trans. Am. Soc. Civil Eng.*, 116, 776–808.
ICAO (International Civil Aviation Organization) 1993. *Manual of the ICAO Standard Atmosphere: Extended to 80 Kilometres / 262 500 Feet*. 3rd Ed., Doc 7488/3, ICAO, Montreal, Canada, 304 pp.
Iliopoulou, T., and Koutsoyiannis, D., 2019. Revealing hidden persistence in maximum rainfall records. *Hydrological Sciences Journal*, 64 (14), 1673–1689, doi:10.1080/02626667.2019.1657578.
Iliopoulou, T., and Koutsoyiannis, D., 2020. Projecting the future of rainfall extremes: better classic than trendy. *Journal of Hydrology*, 588, 125005, doi:10.1016/j.jhydrol.2020.125005.
Jaeger, G., 2017. Quantum potentiality revisited. *Phil. Trans. R. Soc. A*, 375, 20160390, doi: 10.1098/rsta.2016.0390.
Jaeger, G., 2018. Developments in quantum probability and the Copenhagen approach. *Entropy*, 20, 420.
Jaynes, E.T., 1957. Information theory and statistical mechanics. *Phys. Rev.*, 106(4), 620–630.
Jaynes, E.T. 1992. The Gibbs paradox. In *Maximum entropy and Bayesian methods: proceedings of the Eleventh International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis* (Seattle, 1991), 1-21, Kluwer, Dordrech,.
Jaynes, E.T., 2003. *Probability Theory: The Logic of Science*, Cambridge Univ. Press, Cambridge, UK, 728 pp.
Kastner, R.E., Kauffman, S. and Epperson, M., 2018. Taking Heisenberg's potentia seriously. *International Journal of Quantum Foundations*, 4, 158 – 172, https://www.ijqf.org/archives/4643.
Kemp, N.H., 1962. *Calculation of Heat Transfer from Similarity Boundary Layer Equations by a Single Integral Method*. Research Report No. 137, Armed Services Technical Information Agency (U.S. Ballistic Systems Division), Arlington, Virginia, USA, https://apps.dtic.mil/sti/tr/pdf/AD0285506.pdf.

Kendall, M.G., and Stuart, A., 1963. *The Advanced Theory of Statistics*, Vol. 1, Distribution Theory, 2nd edition, Charles Griffin & Co., London, 434 pp.

Keshner, M.S., 1982. 1/f noise. *Proc. IEEE*, 70, 212–218.

Khintchine, A., 1933. Zu Birkhoffs Lösung des Ergodenproblems. *Math. Ann.*, 107, 485-488.

Khintchine, A., 1934. Korrelationstheorie der stationären stochastischen Prozesse. *Mathematische Annalen*, 109 (1), 604–615.

Khinchin, A.Y., 1957. *Mathematical Foundations of Information Theory.* Dover, USA.

Klemeš, V., 1987. One hundred years of applied storage reservoir theory. *Water Resources Management*, 1(3), 159-175.

Kolmogorov, A.N., 1931. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Math. Ann.*, 104, 415-458. (English translation: On analytical methods in probability theory, In: Kolmogorov, A.N., *Selected Works of A. N. Kolmogorov - Volume 2, Probability Theory and Mathematical Statistics*, ed. by A.N. Shiryayev, Kluwer, Dordrecht, The Netherlands, 62-108, 1992).

Kolmogorov, A.N., 1933. Grundbegriffe der Wahrscheinlichkeitsrechnung, *Ergebnisse der Math.* (2), Berlin (2nd English Edition: Foundations of the Theory of Probability, 84 pp. Chelsea Publishing Company, New York, 1956).

Kolmogorov, A.N., 1938. A simplified proof of the Birkhoff-Khinchin ergodic theorem. *Uspekhi Mat. Nauk*, 5, 52-56. (English edition: Kolmogorov, A.N., 1991, *Selected Works of A. N. Kolmogorov - Volume 1, Mathematics and Mechanics*, ed. by Tikhomirov, V.M., Kluwer, Dordrecht, The Netherlands, 271-276).

Kolmogorov, A.N., 1940. Wiener spirals and some other interesting curves in a Hilbert space. *Dokl. Akad. Nauk SSSR*, 26, 115-118. (English edition: Kolmogorov, A.N., 1991, *Selected Works of A. N. Kolmogorov - Volume 1, Mathematics and Mechanics*, ed. by Tikhomirov, V.M., Kluwer, Dordrecht, The Netherlands,, 324-326).

Kolmogorov, A.N., 1947. Statistical theory of oscillations with continuous spectrum. *Collected papers on the 30th anniversary of the Great October Socialist Revolution*, Vol. 1, *Akad. Nauk SSSR*, Moscow-Leningrad, 242-252. (English edition: Kolmogorov, A.N., 1992. *Selected Works of A. N. Kolmogorov - Volume 2, Probability Theory and Mathematical Statistics*, ed. by Shiryayev, A.N., Kluwer, Dordrecht, The Netherlands, 321-330).

Kolokoltsov, V.N., 2021. On a probabilistic derivation of the basic particle statistics (Bose–Einstein, Fermi–Dirac, canonical, grand-canonical, intermediate) and related distributions. *Transactions of the Moscow Mathematical Society*, 2021, 77–87, doi: 10.1090/mosc/316.

Koutsoyiannis, D. 1997. *Statistical Hydrology*. Edition 4 (in Greek), National Technical University of Athens, Athens, 312 pp., doi: 10.13140/RG.2.1.5118.2325.

Koutsoyiannis, D., 2003. Climate change, the Hurst phenomenon, and hydrological statistics. *Hydrological Sciences Journal*, 48 (1), 3–24, doi: 10.1623/hysj.48.1.3.43481.

Koutsoyiannis, D., 2004a. Statistics of extremes and estimation of extreme rainfall, 1, Theoretical investigation. *Hydrological Sciences Journal*, 49 (4), 575–590, doi: 10.1623/hysj.49.4.575.54430.

Koutsoyiannis, D., 2004b. Statistics of extremes and estimation of extreme rainfall, 2, Empirical investigation of long rainfall records. *Hydrological Sciences Journal*, 49 (4), 591–610, doi: 10.1623/hysj.49.4.591.54424.

Koutsoyiannis, D., 2005a. Uncertainty, entropy, scaling and hydrological stochastics, 1, Marginal distributional properties of hydrological processes and state scaling. *Hydrological Sciences Journal*, 50 (3), 381–404, doi: 10.1623/hysj.50.3.381.65031.

Koutsoyiannis, D., 2005b. Uncertainty, entropy, scaling and hydrological stochastics, 2, Time dependence of hydrological processes and time scaling. *Hydrological Sciences Journal*, 50 (3), 405–426, doi: 10.1623/hysj.50.3.405.65028.

Koutsoyiannis, D., 2005c. Internal report, Supplement to "Uncertainty, entropy, scaling and hydrological stochastics, 1, Marginal distributional properties of hydrological processes and state scaling", http://www.itia.ntua.gr/641/.

Koutsoyiannis, D., 2006. An entropic-stochastic representation of rainfall intermittency: The origin of clustering and persistence. *Water Resources Research*, 42 (1), W01401, doi:10.1029/2005WR004175.

Koutsoyiannis, D., 2007. A critical review of probability of extreme rainfall: principles and models, *Advances in Urban Flood Management*, ed. by Ashley, R., Garvin, S., Pasche, E., Vassilopoulos, A., and Zevenbergen, C., 139–166, Taylor and Francis, London. doi: 10.1201/9780203945988.ch7.

Koutsoyiannis, D., 2009. Seeking parsimony in hydrology and water resources technology. European Geosciences Union General Assembly 2009, Geophysical Research Abstracts, Vol. 11, Vienna, 11469, doi: 10.13140/RG.2.2.20511.97443.

Koutsoyiannis, D., 2010. A random walk on water. *Hydrology and Earth System Sciences*, 14, 585–601, doi: 10.5194/hess-14-585-2010.

Koutsoyiannis, D., 2011a. Hurst-Kolmogorov dynamics and uncertainty, *Journal of the American Water Resources Association*, 47 (3), 481–495, doi: 10.1111/j.1752-1688.2011.00543.x.

Koutsoyiannis, D., 2011b. Hurst-Kolmogorov dynamics as a result of extremal entropy production. *Physica A: Statistical Mechanics and its Applications*, 390 (8), 1424–1432.

Koutsoyiannis, D., 2012. Clausius-Clapeyron equation and saturation vapour pressure: simple theory reconciled with practice *European Journal of Physics*, 33 (2), 295–305, doi:10.1088/0143-0807/33/2/295.

Koutsoyiannis, D., 2013a. Physics of uncertainty, the Gibbs paradox and indistinguishable particles. *Studies in History and Philosophy of Modern Physics*, 44, 480–489, doi: 10.1016/j.shpsb.2013.08.007.

Koutsoyiannis, D., 2013b. Hydrology and Change. *Hydrological Sciences Journal*, 58 (6), 1177–1197, doi: 10.1080/02626667.2013.804626.

Koutsoyiannis, D., 2014a. Entropy: from thermodynamics to hydrology. *Entropy*, 16 (3), 1287–1314, doi: 10.3390/e16031287.

Koutsoyiannis, D., 2014b. Random musings on stochastics (Lorenz Lecture), *AGU 2014 Fall Meeting*, San Francisco, USA, American Geophysical Union, doi: 10.13140/RG.2.1.2852.8804.

Koutsoyiannis, D., 2016. Generic and parsimonious stochastic modelling for hydrology and beyond. *Hydrological Sciences Journal*, 61 (2), 225–244, doi: 10.1080/02626667.2015.1016950.

Koutsoyiannis, D., 2017. Entropy production in stochastics. *Entropy*, 19 (11), 581, doi: 10.3390/e19110581.

Koutsoyiannis, D., 2019a. Knowable moments for high-order stochastic characterization and modelling of hydrological processes. *Hydrological Sciences Journal*, 64 (1), 19–33, doi: 10.1080/02626667.2018.1556794.

Koutsoyiannis, D., 2019b. Time's arrow in stochastic characterization and simulation of atmospheric and hydrological processes. *Hydrological Sciences Journal*, 64 (9), 1013–1037, doi: 10.1080/02626667.2019.1600700.

Koutsoyiannis, D., 2020. Simple stochastic simulation of time irreversible and reversible processes. *Hydrological Sciences Journal*, 65 (4), 536–551, doi: 10.1080/02626667.2019.1705302 (also: http://www.itia.ntua.gr/1975/).

Koutsoyiannis, D., 2023. Knowable moments in stochastics: Knowing their advantages. *Axioms*, 12 (6), 590, doi:10.3390/axioms12060590.

Koutsoyiannis, D., 2024. Relative importance of carbon dioxide and water in the greenhouse effect: Does the tail wag the dog?. *Science of Climate Change*, 4 (2), 36–78, doi:10.53234/scc202411/01.

Koutsoyiannis, D., 2025. *Stochastics of Hydroclimatic Extremes - A Cool Look at Risk*. Edition 5, ISBN: 978-618-85370-0-2, 420 pages, doi:10.57713/kallipos-1, Kallipos Open Academic Editions, Athens, http://www.itia.ntua.gr/en/docinfo/2000/..

Koutsoyiannis, D., and Dimitriadis, P., 2021. Towards generic simulation for demanding stochastic processes. *Sci*, 3, 34, doi: 10.3390/sci3030034.

Koutsoyiannis, D., and Georgakakos, A., 2006. Lessons from the long flow records of the Nile: determinism vs indeterminism and maximum entropy. *20 Years of Nonlinear Dynamics in Geosciences*, Rhodes, Greece, doi: 10.13140/RG.2.2.10996.14727.

Koutsoyiannis, D., and Iliopoulou, T., 2024. *Understanding Climate: Gifts from the Nile*, 60 pages, SR 301, The Heritage Foundation, Washington, DC, USA.

Koutsoyiannis, D., and Mamassis, N., 2021. From mythology to science: the development of scientific hydrological concepts in the Greek antiquity and its relevance to modern hydrology. *Hydrology and Earth System Sciences*, 25, 2419–2444, doi:10.5194/hess-25-2419-2021.

Koutsoyiannis, D., and Montanari, A., 2007. Statistical analysis of hydroclimatic time series: Uncertainty and insights. *Water Resources Research*, 43 (5), W05429, doi: 10.1029/2006WR005592.

Koutsoyiannis, D., and Montanari, A., 2015. Negligent killing of scientific concepts: the stationarity case. *Hydrological Sciences Journal*, 60 (7-8), 1174–1183, doi: 10.1080/02626667.2014.959959.

Koutsoyiannis, D., and Sargentis, G.-F., 2021. Entropy and wealth. *Entropy*, 23 (10), 1356, doi: 10.3390/e23101356.

Koutsoyiannis, D., and Tsakalias, G., 2025. Unsettling the settled: Simple musings on the complex climatic system. *Frontiers in Complex Systems*, 3, 1617092, doi: 10.3389/fcpxs.2025.1617092.

Koutsoyiannis, D., Makropoulos, C., Langousis, A., Baki, S., Efstratiadis, A., Christofides, C., Karavokiros, G., and Mamassis, N., 2008. Interactive comment on "HESS Opinions, Climate, hydrology, energy, water: Recognizing uncertainty and seeking sustainability" by Koutsoyiannis, D. et al. *Hydrol. Earth Syst. Sci. Discuss.*, *5*, S1761–S1774, http://www.hydrol-earth-syst-sci-discuss.net/5/S1761/2008/.

Koutsoyiannis, D., Dimitriadis, P., Lombardo, F., and Stevens, S., 2018. From fractals to stochastics: Seeking theoretical consistency in analysis of geophysical data. *Advances in Nonlinear Geosciences*; ed. by Tsonis, A., Springer, New York, NY, USA, doi: 10.1007/978-3-319-58895-7_14.

Krumbein, W.C., 1968. Statistical models in sedimentology. *Sedimentology*, 10 (1), 7-23.
Kyun, K. and Kim, K., 2006. *Equilibrium Business Cycle Theory in Historical Perspective*. Cambridge University Press, Cambridge, UK.
Lago-Fernández, L.F., and Corbacho, F., 2009. Using the negentropy increment to determine the number of clusters. In I*nternational Work-Conference on Artificial Neural Networks*; Springer: Berlin/Heidelberg, Germany, 448–455.
Larouche, L.H., Jr., 1993. On Larouche's Discovery. http://www.archive.schillerinstitute.com/fidelio_archive/1994/fidv03n01-1994Sp/fidv03n01-1994Sp_037-on_larouches_discovery-lar.pdf.
Lasota, A., and Mackey, M.C., 1994. *Chaos, Fractals and Noise*, Springer-Verlag, New York, USA.
Leadbetter, M.R., 1983. Extremes and local dependence in stationary sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65, 291–306.
Leff, H.S., 2012. Removing the mystery of entropy and thermodynamics—Part V. *Phys. Teach*., 50, 274–276, doi: 10.1119/1.3703541.
Lesher, J.H., 2010. Saphêneia in Aristotle:'Clarity','Precision', and 'Knowledge'. *Apeiron*, 43 (4), 143-156.
List R.J., 1951. *Smithsonian Meteorological Tables*. Smithsonian Institution, Washington, DC, USA, 527 pp.
Livadiotis, G., and McComas, D.J., 2013. Understanding kappa distributions: A toolbox for space science and astrophysics. *Space Science Reviews*, 175 (1-4), 183-214.
Lombardo, F., Napolitano, F., Russo, F., and Koutsoyiannis, D., 2019. On the exact distribution of correlated extremes in hydrology. *Water Resources Research*, 55 (12), 10405–10423, doi: 10.1029/2019WR025547.
Lombardo, F., Volpi, E., Koutsoyiannis, D., and Papalexiou, S.M., 2014. Just two moments! A cautionary note against use of high-order moments in multifractal models in hydrology. *Hydrology and Earth System Sciences*, 18, 243–255, doi: 10.5194/hess-18-243-2014.
Ma, N., Zhang, Y., and Yang, Y., 2025. Recent decline in global ocean evaporation due to wind stilling. *Geophys. Res. Lett.* 52, e2024GL114256, doi: 10.1029/2024GL114256.
Mackey, M.C., 2003. *Time's Arrow: The Origins of Thermodynamic Behavior*, Dover, Mineola, NY, USA, 175 pp.
Manabe, S. and Wetherald, R.T., 1967. Thermal equilibrium of the atmosphere with a given distribution of relative humidity. *J. Atmos. Sci.*, 24 (3), 241-259.
Mandelbrot, B.B., 1999. *Multifractals and 1/f Noise: Wild Self-Affinity in Physics (1963-1976)*. Springer, New York, NY, USA.
Mandelbrot, B.B., and van Ness, J.W., 1968. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10(4), 422-437.
Mandelbrot, B.B. and Wallis, J.R., 1968. Noah, Joseph, and operational hydrology. *Water Resources Research*, 4(5), 909-918.
Markonis, Y., and Koutsoyiannis, D., 2016. Scale-dependence of persistence in precipitation records, *Nature Climate Change*, 6, 399–401, doi: 10.1038/nclimate2894.
Matalas, N.C., 1967. Mathematical assessment of synthetic hydrology. *Water Resour. Res.,* 3(4), 937-945.
Mathieu, M., 1988. On the origin of the notion 'Ergodic Theory'. *Expositiones Mathematicae*, 6, 373-377.
Mazliak, L., 2018. The beginnings of the Soviet Encyclopedia. The utopia and misery of mathematics in the political turmoil of the 1920s. *Centaurus*, 60(1-2), 25-51.
Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E., and Houston, T.G., 2012. An overview of the Global Historical Climatology Network-Daily Database. *J. Atmos. Oceanic Technol.*, 29, 897-910, doi: 10.1175/JTECH-D-11-00103.1.
Mermin, N.D., 1985. Is the moon there when nobody looks? Reality and the quantum theory. *Physics Today*, 38 (4), 38-47.
Metropolis, N., and Ulam, S., 1949. The Monte Carlo method. *Journal of the American Statistical Association*, 44(247), 335-341.
Mielke, P.W. Jr., 1973. Another family of distributions for describing and analyzing precipitation data. *Journal* of *Applied Meteorology,* 12 (2), 275-280.
Mielke P.W. Jr., and Johnson, E.S., 1973. Three-parameter kappa distribution maximum likelihood estimates and likelihood ratio tests. *Monthly Weather Review, 101(9)*, 701-707.
Milly, P.C.D., Betancourt, J. Falkenmark, M. Hirsch, R.M. Kundzewicz, Z.W. Lettenmaier, D.P. and Stouffer, R.J., 2008. Stationarity Is Dead: Whither Water Management?. *Science*, 319, 573-574.
Montanari, A., and Koutsoyiannis, D., 2014. Modeling and mitigating natural hazards: Stationarity is immortal!. *Water Resources Research*, 50 (12), 9748–9756, doi: 10.1002/2014WR016092.
Moore, T.A., 2002. *Six Ideas that Shaped Physics. Unit T–Thermodynamics*. McGraw-Hill, New York, USA.
Murphy, D.M. and Koop, T., 2005. Review of the vapor pressures of ice and supercooled water for atmospheric applications, *Q. J. R. Meteorol. Soc.*, 131, 1539–1565.

Neumann, J. and Flohn, H., 1987. Great historical events that were significantly affected by the weather: Part 8, Germany's war on the Soviet Union, 1941–45. I. Long-range weather forecasts for 1941–42 and climatological studies. *Bulletin of the American Meteorological Society*, 68(6), 620-630.

Niederreiter, H., 1992. *Random Number Generation and Quasi-Monte Carlo Methods*, Society for Industrial and Applied Mathematics, Philadelphia, USA.

O'Connell, P.E., Koutsoyiannis, D., Lins, H.F., Markonis, Y., Montanari, A., and Cohn, T.A., 2016. The scientific legacy of Harold Edwin Hurst (1880 – 1978). *Hydrological Sciences Journal*, 61 (9), 1571–1590, doi: 10.1080/02626667.2015.1125998.

Olbert, S., 1968. Summary of Experimental Results from M.I.T. Detector on IMP-1. *Physics of the Magnetosphere, Astrophysics and Space Science Library*, ed. by Carovillano R.L., McClay J.F., Radoski H.R., vol 10. Springer, Dordrecht, The Netherlands.

Papalexiou, S.M., and Koutsoyiannis, D. 2012. Entropy based derivation of probability distributions: A case study to daily rainfall. *Advances in Water Resources*, 45, 51–57, doi: 10.1016/j.advwatres.2011.11.007.

Papoulis, A., 1990. *Probability and Statistics*, Prentice-Hall, New Jersey, USA.

Papoulis, A., 1991. *Probability, Random Variables and Stochastic Processes* (3rd edn.). McGraw-Hill, New York.

Perry, J., 1903. The Thermodynamics of Heat Engines. *Nature*, 67, 602–605, doi: 10.1038/067602a0.

Persons, W.M. 1919. *Measuring and Forecasting General Business Conditions*. Harvard University & American Institute of Finance, https://babel.hathitrust.org/cgi/pt?id=hvd.32044018749697.

Planck, M., 1906. *Vorlesungen über die Theorie der Wärmestrahlung*. J. A. Barth, Leipzig, https://archive.org/details/vorlesungenberd03plangoog.

Planck, M., 1914. *The Theory of Heat Radiation*. Blakiston, Philadelphia, Pennsylvania, USA, https://www.gutenberg.org/ebooks/40030.

Popper, K., 1982. *Quantum Physics and the Schism in Physics*. Unwin Hyman, London, 229 pp.

Prutchi, D., 2012. *Exploring Quantum Physics Through Hands-On Projects*. Wiley, Hoboken, New Jersey, USA.

Rényi, A., 1961. On measures of information and entropy. *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability* 1960, 547–561.

Ripley, B.D., 1987. *Stochastic Simulation*, Wiley, New York, USA.

Robertson, H.S., 1993. *Statistical Thermophysics*. Prentice Hall, Englewood Cliffs, NJ, 582 pp.

Salas, J.D., 1993. Analysis and modeling of hydrologic time series. *Handbook of Hydrology*, ed. by Maidment, D., Ch. 19, 19.1-19.72, McGraw-Hill, New York, USA.

Sanders, G., 2018. An Aristotelian approach to quantum mechanics. *Academia*, http://www.academia.edu/35229710/.

Sargentis, G.-F., Iliopoulou, T., Sigourou, S., Dimitriadis, P., and Koutsoyiannis, D., 2020. Evolution of clustering quantified by a stochastic method — Case studies on natural and human social structures. *Sustainability*, 12 (19), 7972, doi: 10.3390/su12197972.

Saridis, G.N., 2004. Entropy as a philosophy. In *Proceedings of the 2004 Performance Metrics for Intelligent Systems Workshop*; National Institute of Standards & Technology (NIST), Manufacturing Engineering Lab, Gaithersburg, MD, USA, http://apps.dtic.mil/sti/citations/ADA515701.

Schoemaker, P.J.H., 2003. Huygens versus Fermat: No clear winner. *Behavioral and Brain Sciences*, 26, 781–783.

Schrödinger, E., 1944. *What is Life? The Physical Aspect of the Living Cell*. Cambridge University Press, Cambridge, UK.

Shaked, M. and Shanthikumar, J.G., 2007. *Stochastic Orders*. Springer, New York, NY, USA.

Shannon, C.E., 1948. A mathematical theory of communication. *Bell Systems Tech. J.*, 27(379), 623-656.

Shore, J. and Johnson, R., 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26 (1), 26-37.

Shuttleworth W.J., 1993. Evaporation. Ch.4 in *Handbook of Hydrology*, ed. by Maidment, D.R., McGraw-Hill, New York, pp. 4.1-4.53.

Singh, V.P., and Rajagopal, A.K., 1986. A new method of parameter estimation for hydrologic frequency analysis. *Hydrological Science and Technology*, 2 (3) 33-44.

Slutsky, E.,1925, Über stochastische Asymptoten und Grenzwerte, *Metron*, 5(3), 3-89.

Slutsky, E., 1927. Slozhenie sluchainykh prichin, kak istochnik tsiklicheskikh protsessov. *Voprosy kon''yunktury*, 3, 34 – 64. 1927 (English edition: Slutzky, E., 1937. The summation of random causes as the source of cyclic processes. *Econometrica: Journal of the Econometric Society*, 105-146).

Slutsky, E., 1928a. Sur ùn critérium de la convergence stochastique des ensembles des valeurs éventuelles). *Comptes rendus Acad. Sci.*, 187, 370.

Slutsky, E., 1928b. Sur les fonctions éventuelles continues, intégrables et dérivables dans le sens stochastiques. *Comptes rendus des séances de l'Académie des Sciences,* 187, 878.

Slutsky, E.,1929. Quelques propositions sur les limites stochastiques éventuelles. *Comptes rendus des séances de l'Académie des Sciences*, 189, 384.

Smith, P. J., 1995. A recursive formulation of the old problem of obtaining moments from cumulants and vice versa. *The American Statistician*, 49 (2), 217–218.

Stigler, S.M., 2002. *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, Mass., USA.

Stowe, K., 2007. *Thermodynamics and Statistical Mechanics* (2nd edn.). Cambridge Univ. Press, Cambridge, 556 pp.

Strauss, W.A., 2007. Partial Differential Equations: An Introduction. John Wiley & Sons.

Styer, D., 2019. Entropy as disorder: History of a misconception. *Phys. Teach.*, 57, 454, doi: 10.1119/1.5126822.

Sudler, C., 1927. Storage required for the regulation of streamflow. *Trans. Am. Soc. Civ. Eng.*, 91, 622-660.

Swendsen, R.H., 2006. Statistical mechanics of colloids and Boltzmann's definition of the entropy. *American Journal of Physics*, 74 (3), 187-190.

Swendsen, R.H., 2008. Gibbs' paradox and the definition of entropy. *Entropy*, 10(1), 15-18.

Swendsen, R.H., 2011. How physicists disagree on the meaning of entropy. *Am. J. Phys.*, 79 (4), 342-348.

Swinburne, J., 1904. Entropy. *Nature*, 70, 54–55, doi: 10.1038/070054b0.

Szilard, L., 1929. Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen. *Zeitschrift für Physik*, 53 (11), 840-856.

Takhar, H.S., Chamkha, A.J. and Nath, G., 2004. Effect of thermophysical quantities on the natural convection flow of gases over a vertical cone. *International Journal of Engineering Science*, 42 (3-4), 243-256.

Thomas, H.A., and Fiering, M.B. 1962. Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. *Design of Water Resource Systems*, ed. by Maass, A., Hufschmidt, M.M., Dorfman, R., Thomas, H.A., Jr., Marglin, S.A., and Fair, G.M., Harvard University Press, Cambridge, Mass.

Tippett, L.H.C., 1927. *Random Sampling Numbers*. Cambridge University Press, Cambridge, UK.

Todorovic, P., and Zelenhasic, E., 1970. A stochastic model for flood analysis. *Water Resources Research*, 6(6), 1641–1648.

Tsallis, C., 1988. Possible generalization of Boltzmann-Gibbs statistics. *J. Statist. Phys.*, 52, 479–487.

Tsallis, C., 2022. Entropy. *Encyclopedia*, 2 (1), 264-300, doi: 10.3390/encyclopedia2010018

Tyralis, H., and Koutsoyiannis, D., 2011. Simultaneous estimation of the parameters of the Hurst-Kolmogorov stochastic process. *Stochastic Environmental Research & Risk Assessment*, 25 (1), 21–33.

Tyralis, H., and Koutsoyiannis, D., 2014. A Bayesian statistical model for deriving the predictive distribution of hydroclimatic variables. *Climate Dynamics*, 42 (11-12), 2867–2883, doi: 10.1007/s00382-013-1804-y.

Tyralis, H., Koutsoyiannis, D., and Kozanis, S., 2013. An algorithm to construct Monte Carlo confidence intervals for an arbitrary function of probability distribution parameters. *Computational Statistics*, 28 (4), 1501–1527, doi: 10.1007/s00180-012-0364-7.

Uffink, J., 1995. Can the maximum entropy principle be explained as a consistency requirement?. *Studies In History and Philosophy of Modern Physics*, 26 (3), 223-261.

U.S. Standard Atmosphere, 1976. U.S. Government Printing Office, Washington, DC, USA, https://www.ngdc.noaa.gov/stp/space-weather/online-publications/miscellaneous/us-standard-atmosphere-1976/us-standard-atmosphere_st76-1562_noaa.pdf.

van der Sluijs, J., 2005. Uncertainty as a monster in the science–policy interface: Four coping strategies. *Water Sci. Technol*., 52, 87–92.

van Wijngaarden, W.A., and Happer, W., 2023. Atmosphere and greenhouse gas primer. *arXiv Prepr*. arXiv:2303.00808, doi: 10.48550/arXiv.2303.00808.

Visconti, E.Q., 1817. *Planches de l'Iconographie Grecque*. De l'Imprimerie de P. Didot l'Ainé, Paris, 58 plates (engravings), https://archive.org/details/gri_33125010850713/ and https://arachne.dainst.org/entity/1884649.

Volpi, E., Fiori, A., Grimaldi, S., Lombardo, F., and Koutsoyiannis, D., 2019. Save hydrological observations! Return period estimation without data decimation. *Journal of Hydrology*, doi: 10.1016/j.jhydrol.2019.02.017.

von Mises, R., 1936. La distribution de la plus grande de n valeurs. *Rev. Math. Union Interbalcanique*, 1, 141-160.

von Neumann, J., 1956. Probabilistic logics and the synthesis of reliable organisms from unreliable components. *Automata Studies* , 43-98, doi: 10.1515/9781400882618-003.

Wagner, W., and Pruss, A. 2002. The IAPWS formulation 1995 for the thermodynamic properties of ordinary water substance for general and scientific use. *J. Phys. Chem. Data*, 31, 387-535.

Wallis, J. R., and O'Connell, P.E., 1972. Small sample estimation of ρ_1. *Water Resour. Res.*, 8(3), 707-712.

Walker, G., 1931. On periodicity in series of related terms. *Proceedings of the Royal Society of London, Ser. A*, 131, 518–532.

Wannier, G.H., 1987. *Statistical Physics*. Dover, New York, 532 pp.

Whittle, P., 1951. *Hypothesis Testing in Times Series Analysis*. PhD thesis, Almqvist & Wiksells, Uppsala, Sweden.

Whittle, P., 1952. Tests of fit in time series. *Biometrika*, 39(3/4), 309-318.

Whittle, P., 1953. The analysis of multiple stationary time series. *Journal of the Royal Statistical Society B*, 15 (1), 125-139.

Wiener, N., 1948a. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT Press, Cambridge, Mass., USA, 212 pp.

Wiener, N., 1948b. Time, communication, and the nervous system. *Ann. N. Y. Acad. Sci.*, *50*, 197–220.

Wiuff, R., 2023. Was Franz Baur's infamous long-range weather forecast for the winter of 1941/42 on the Eastern Front really wrong?. *Bulletin of the American Meteorological Society*, 104(1), E107-E125.

WMO (World Meteorological Organization), 1957. Meteorology—a Three-Dimensional Science. Second Session for the Commission for Aerology. *WMO Bull. 6,* 134–138.

Wold, H.O., 1938. *A Study in the Analysis of Stationary Time-Series*. PhD thesis, Almquist and Wicksell, Uppsala, Sweden.

Wold, H.O., 1948. On prediction in stationary time series. *The Annals of Mathematical Statistics*, 19(4), 558-567.

Wornell, G.W., 1993. Wavelet-based representations for the 1/f family of fractal processes. *Proc. IEEE*, 81, 1428–1450.

Wright, P.G., 1970. Entropy and disorder. *Contemp. Phys.*, 11, 581–588, doi: 10.1080/00107517008202196.

Yule, G.U. 1927. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London A*, 226, 267–298.

Zou, L., Hoffmann, L., Müller, R. and Spang, R., 2023. Variability and trends of the tropical tropopause derived from a 1980–2021 multi-reanalysis assessment. *Frontiers in Earth Science*, 11, 1177502, doi: 10.3389/feart.2023.1177502.