Chapter 5
A Primer on Information Theory with Applications to Neuroscience

Felix Effenberger

5.1 Introduction

Neural systems process information. This processing is of fundamental biological importance for all animals and humans alike as its main (if not sole) biological purpose is to ensure the survival of an individual (in the short run) and its species (in the long run) in a given environment by means of perception, cognition, action, and adaption.

Information enters a neural system in form of sensory input representing some aspect of the outside world, perceivable by the sensory modalities present in the system. After processing this information or parts of it, the system may then adjust its state and act according to a perceived change in the environment.

This general model is applicable to very basic acts of cognition as well as to ones requiring higher degrees of cognitive processing. Yet, the underlying principle is the same. Thus measuring, modeling, and (in the long run) understanding information processing in neural systems is of prime importance for the goal of gaining insight to the functioning of neural systems on a theoretical level.

Note that this question is of theoretical and abstract nature so that we take an abstract view on information in what follows. We use Shannon's theory of information [97] as a tool that provides us with a rigid mathematical theory and quantitative measures of information. Using information theory, we will have a conceptual look at information in neural systems. In this context, information theory can provide both explorative and normative views on the processing of

F. Effenberger (✉)
Max-Planck-Institute for Mathematics in the Sciences,
Inselstr. 22, 04103 Leipzig, Germany
e-mail: felix.effenberger@mis.mpg.de

G. Rakocevic et al. (eds.), *Computational Medicine in Data Mining and Modeling*,
DOI 10.1007/978-1-4614-8785-2_5,

information in a neural system as we will see in Sect. 5.6. In some cases, it is even possible to gain insights on the nature of the "neural code," i.e., the way neurons transmit information via their spiking activity.

Information theory was originally used to analyze and optimize man-made communication systems, for which the functioning principles are known. Nonetheless, it was soon realized that the theory could also be used in a broader setting, namely, to gain insight into the functioning of systems for which the underlying principles are far from fully understood, such as neural systems. This was the beginning of the success story of information-theoretic methods in many fields of science such as economics, psychology, biology, chemistry, and physics.

The idea of using information theory to quantitatively assess information processing in neural systems has been around since the 1950s; see the works of Attneave [6], Barlow [9], and Eckhorn and Pöpel [32, 33]. Yet, as information-theoretic analyses are data intensive, these methods were rather heavily restricted by (a) the limited resources of computer memory and computational power available and (b) the limited accuracy and amount of measured data that could be obtained from neural systems (on the single cell as well as at the systems level) at that time. However, given the constant rise in available computing power and the evolution and invention of data acquisition techniques that can be used to obtain data from neural systems (such as magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), or calcium imaging), information-theoretic analyses of all kinds of biological and neural systems became more and more feasible and could be carried out with greater accuracy and for larger and larger (sub)systems.

Over the last decades such analyses became possible using an average workstation computer, a situation that could only be dreamed of in the 1970s. Additionally, the emergence of new noninvasive data collection methods such as fMRI and MEG that outperform more traditional methods like electroencephalography (EEG) in terms of spatial resolution (fMRI, MEG) or noise levels (MEG) made it possible to even obtain and analyze system-scale data of the human brain in vivo.

The goal of this chapter is to give a short introduction to the fundamentals of information theory and its application to data analysis problems in the neurosciences. And although information-theoretic analyses of neural systems were not often used in order to gain insight on or characterize neural dysfunction so far, this could prove to be a helpful tool in the future.

The chapter is organized as follows. We first talk a bit about the process of modeling in Sect. 5.2 that is fundamental for all what follows as it connects reality with theory. As information theory is fundamentally based on probability theory, following this we give an introduction to the mathematical notions of probabilities, probability distributions, and random variables in Sect. 5.3. If you are familiar with probability theory, you may well skim or skip this section. Section 5.4 deals with the main ideas of information theory. We first take a view on what we mean by information and introduce the core concept of information theory, namely, *entropy*. Starting from the concept of entropy, we will then continue to look at more complex notions such as *conditional entropy* and *mutual information* in Sect. 4.3. We will

then consider a variant of *conditional mutual information* called *transfer entropy* in Sect. 4.5. We conclude the theoretical part by discussing methods used for the estimation of information-theoretic quantities from sampled data in Sect. 5.5. What follows will deal with the application of the theoretical measures to neural data. We then give a short overview of applications of the discussed theoretical methods in the neurosciences in Sect. 5.6, and last (but not least), Sect. 5.7 constrains a list of software packages that can be used to estimate information-theoretic quantities for some given data set.

5.2 Modeling

In order to analyze the dynamics and gain a theoretical understanding of a given complex system, one usually defines a model first, i.e., a simplified theoretical version of the system to be investigated. The rest of the analysis is then based on this model and can only capture aspects of the system that are also contained in the model. Thus, care has to be taken when creating the model as the following analysis crucially depends on the quality of the model.

When building a model based on measured data, there is an important thing we have to pay attention to, namely, that any data obtained by measurement of physical quantities is only accurate up to a certain degree and corrupted by noise. This naturally also holds for neural data (e.g., electrophysiological single- or multi-cell measurements, EEG, fMRI, or MEG data). Therefore, when observing the state of some system by measuring it, one can only deduce the true state of the system up to a certain error determined by the noise in the measurement (which may depend both on the measurement method and the system itself). In order to model this uncertainty in a mathematical way, one uses probabilistic models for the states of the measured quantities of a system. This makes probability theory a key ingredient to many mathematical models in the natural sciences.

5.3 Probabilities and Random Variables

The roots of the mathematical theory of probability lie in the works of Cardano, Fermat, Pascal, Bernoulli, and de Moivre in the sixteenth and seventeenth centuries, in which the authors attempted to analyze games of chance. Pascal and Bernoulli were the first to treat the subject as a branch of mathematics; see [106] for a historical overview. Mathematically speaking, probability theory is concerned with the analysis of random phenomena. Over the last centuries, it has become a well-established mathematical subject. For a more in-depth treatment of the subject see [47, 52, 98].

5.3.1 A First Approach to Probabilities via Relative Frequencies

Let us consider an experiment that can produce a certain fixed number of outcomes (say a coin toss, where the possible outcomes are heads or tails or the throw of a die where the die will show one of the numbers 1 to 6). The set of all possible outcomes is called the *sample space* of the experiment.

One possible result of an experimen t is called *outcome,* and a set of outcomes is called an *event* (for the mathematically adept: an event is a subset of the power set of all outcomes). Take, for example, the throw of a regular, six-sided die as an experiment. The set of results in this case would be the set of natural numbers $\{1, \ldots,6\}$, and examples of events are $\{1,3,5\}$ or $\{2,4,6\}$ corresponding to the events "an odd number was thrown" and "an even number was thrown," respectively.

The classical definition of the probability of an event is due to Laplace: "The probability of an event to occur is the number of cases favorable for the event divided by the number of total outcomes possible" [106].

We thus assign each possible outcome a *probability*, a real number between 0 and 1 that is thought of as to describe how "likely" it is that the given event will occur, where 0 means "the event does not ever occur" and 1 means "the event always occurs." The sum of all the assigned numbers is restricted to be 1 as we assume that one of our considered events always occurs. For the coin toss, the possible outcomes heads and tails thus each have probability $\frac{1}{2}$ (considering that the number of favorable outcomes is one and the number of possible outcomes is two), and for the throw of a die this number is $\frac{1}{6}$ for each digit. This assumes that we have a so-called *fair* coin or die, i.e., one that does not favor any particular outcome over the others.

The probability of a given event to occur is then just the sum of the probabilities of the outcomes the event is composed of, e.g., when considering the throw of a die, the probability of the event "an odd number is thrown" is $\frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$.

Such types of experiments in which all possible outcomes have the same probability (they are called *equiprobable*) are called *Laplacian experiments*. The simplest case of an experiment not having equiprobable outcomes is the so-called Bernoulli experiment. Here, two possible outcomes "success" and "failure" with probabilities $p \in [0,1]$ and $1 - p$ are considered. Let us now consider probabilities in the general setting.

5.3.2 An Axiomatic Description of Probabilities

The foundations of modern probability theory were laid by Kolmogorov [54] in the 1930s. He was the first to give an axiomatic description of probability theory based on measure theory, putting the field on a mathematically sound basis. We will state

his axiomatic description of probabilities in the following. This rather technical approach might seem a little complicated and cumbersome at first, and we will try to give well-understandable explanations of the concepts and notions used as they are of general importance.

Kolmogorov's definition is based on what is known as measure theory, a field of mathematics that is concerned with measuring the (geometric) size of subsets of a given space. Measure theory gives an axiomatic description of a *measure* (as a function μ assigning a nonnegative number to each subset) that fulfills the usual properties of a geometric measure of length (in one-dimensional space), area (in two-dimensional space), volume (in three-dimensional space), and so on. For example, if we take the measure of two disjoint (i.e., non-overlapping) sets, we expect the measure of their union to be the sum of the measures of the two sets and so on.

One prior remark on the definition: When looking at sample spaces (remember, these are the sets of possible outcomes of a random experiment), we have to make a fundamental distinction between *discrete sample spaces* (i.e., ones in which the outcomes can be separated and counted, like in a pile of sand, where we think of each little sand particle representing one possible outcome) and *continuous sample spaces* (where the outcomes form a continuum and cannot be separated and counted, think of this sample space as some kind of dough in which the outcomes cannot be separated). Although in most cases the continuous setting can be treated as a straightforward generalization of the discrete case and we just have to replace sums by integrals in the formulas, some technical subtleties exist, that makes a distinction between the two cases necessary. This is why we separate the two cases in all of what follows.

Definition 3.1 Measure Space and Probability Space. *A measure space is a triple* (Ω,F,μ). *Here*

- The *base space* Ω denotes an arbitrary nonempty set.
- F denotes the set of *measurable sets* in Ω which has to be a so-called σ-algebra over Ω, i.e., it has to fulfill:
  - $\varnothing \in \mathrm{F}$
  - F is closed under complements: if $E \in \mathrm{F}$, then $(\Omega \backslash E) \in \mathrm{F}$.
  - F is closed under countable unions: if $E_i \in \mathrm{F}$ for $i = 1, 2, \ldots$, then $(\cup_i E_i) \in \mathrm{F}$.
- μ is the so-called measure: It is a function $\mu : \mathrm{F} \to \mathrm{R} \cup \{\infty\}$ with the following properties
  - $\mu(\varnothing) = 0$ and $\mu \geq 0$ (non-negativity).
  - μ is countably additive: if $E_i \in \mathrm{F}$, $i = 1, 2, \ldots$ is a collection of pairwise disjoint (i.e., non-overlapping) sets, then $\mu(\cup_i E_i) = \sum_i \mu(E_i)$.

Why this complicated definition of measurable sets, measures, etc.? Well, this is mathematically the probably (no pun intended) most simple way to formalize

the notion of a "measure" (in terms of geometric volume) as we know it over the real numbers.

When defining a measure, we first have to fix the whole space in which we want to measure. This is the base space Ω. Ω can be any arbitrary set: the sample space of a random experiment, e.g., $\Omega = \{\text{heads, tails}\}$ when we look at a coin toss or $\Omega = \{1, \ldots,6\}$ when we look at the throw of a die (these are two examples of discrete sets), the set of real numbers R, the real plane R^2 (these are two examples of continuous sets), or whatever you choose it to be. When modeling the spiking activity of a neuron, the two states could be "neuron spiked" or "neuron did not spike."

In a second step we choose a collection of subsets of Ω that we name F, the collection of subsets of Ω that we want to be measurable. Note that the measurable subsets of Ω are not given a priori, but that we determine those by choosing F. So, you may ask, why this complicated setup with F, why not make every possible subset of Ω measurable, i.e., make F the power set of Ω (the power set is the set of all subsets of Ω)? This is totally reasonable and can easily been done when the number of elements of Ω is finite. But as with many things in mathematics, things get complicated when we deal with the continuum: In many natural settings, e.g., when Ω is a continuous set, this is just not possible or desirable for technical reasons. That is why we choose only a subset of the power set (you might refer to its elements as the "privileged" subsets) and make only the contained subsets measurable. We want to choose this subset in a way that the usual constructions that we know from geometric measures still work in the usual way, though. This motivates the properties that we impose on F: We expect to be able to measure the complements of measurable sets, as well as the union and intersection of a finite number of measurable sets to again be measurable. These properties are motivated by the corresponding properties of geometric measures (i.e., the union, intersection and complement of intervals of certain lengths has a length and so on). So to sum up, the set F is a subset of the power set of Ω, and sets that are not in F are not measurable.

In a last step, we choose a function μ that assigns a measure (think of it as a generalized geometric volume) to each measurable set (i.e., each element of F), where the measure has to fulfill some basic properties that we know from geometric measures: The measure is nonnegative, the empty set (that is contained in every set) should have measure 0, and the measure is additive.

All together, this makes the triple (Ω,F,μ) a space in which we can measure events and use constructions that we know from basic geometry. Our definition makes sure that the measure μ behaves in the way we expect it to (mathematicians call this a natural construction). Take some time to think about it: Definition 3.1 above generalizes the notion of the geometric measure in terms of the length $l(I) = b - a$ of intervals $I = [a,b]$ over the real numbers.

In fact, when choosing the set $\Omega = \mathrm{R}$, we can construct the so-called Borel σ-algebra B that contains all closed intervals $I = [a,b]$, $a < b$, and a measure μ_B that assigns each interval $I = [a,b] \in \mathrm{B}$ its length $\mu_\mathrm{B}(I) = b - a$. The measure μ_B is called *Borel measure*. It is the standard measure of length that we know from geometry and makes $(\mathrm{R},\mathrm{B},\mu_\mathrm{B})$ a measure space. This construction can easily be

extended to arbitrary dimensions (using closed sets) resulting in the measure space $(\mathrm{R}^n, \mathrm{B}^n, \mu_{\mathrm{B}^n})$ that fulfills the properties of a n-dimensional geometric measure of volume.

Let us look at some examples of measure spaces now:

1. Let $\Omega = \{0,1\}$, $\mathrm{F} = \{\emptyset,\{0\},\{1\},\Omega\}$, and P with $P(0) = P(1) = 0.5$. This makes (Ω,F,P) a measure space for our coin toss experiment. Note that in this simple case, F equals the full power set of Ω.
2. Let $\Omega = \{a,b,c,d\}$ and let $\mathrm{F} = \{\emptyset,\{a,b\},\{c,d\},\Omega\}$ with $P(\{a,b\}) = p$ and $P(\{c,d\}) = 1 - p$, where p denotes an arbitrary number between 0 and 1. This makes (Ω,F,P) a measure space.

Having understood the general case of a measure space, defining a probability space and a probability distribution is easy.

Definition 3.2 Probability Space, Probability Distribution. *A probability space is a measure space* (Ω,F,μ) *for which the measure* μ *is normed,* i.e., $\mu : \Omega \rightarrow [0,1]$ *with* $\mu(\Omega) = 1$. *The measure* μ *is called probability distribution and is often also denoted by P (for probability).* Ω *is called the sample space, elements of* Ω *are called outcomes and* F *is the set of events.*

Note that again, we make the distinction between discrete and continuous sample spaces here. In the course of history, a probability distribution on a discrete sample space came to be called *probability mass function* (or *pmf*), and a probability distribution defined on a continuous sample space came to be called *probability density function* (or *pdf*).

Let us look at a few examples, where the probability spaces in the following are given by the triple (Ω,F,P):

1. Let $\Omega = \{\text{heads,tails}\}$ and let $\mathrm{F} = \{\emptyset,\{\text{head}\},\{\text{tails}\},\Omega\}$. This is a probability space for our coin toss experiment, where $\emptyset$ relates to the event "neither heads nor tails" and Ω to the event "either heads or tails." Note that in this simple case, F equals the full power set of Ω.
2. Let $\Omega = \{1, \ldots,6\}$ and let F be the full power set of Ω (i.e., the set of all subsets of Ω, there are $6^2 = 36$, can you enumerate them all?). This is a probability for our experiment of dice throws, where we can distinguish all possible events.

5.3.3 Theory and Reality

It is important to stress that probabilities themselves are a mathematical and purely theoretical construct to help in understanding and analyzing random experiments, and per se they do not have to do anything with reality. They can be understood as an "underlying law" that generates the outcomes of a random experiment and *can never* be directly observed; see Fig. 5.1. But with some restrictions they can be estimated for a certain given experiment by looking at the outcomes of many repetitions of that experiment.

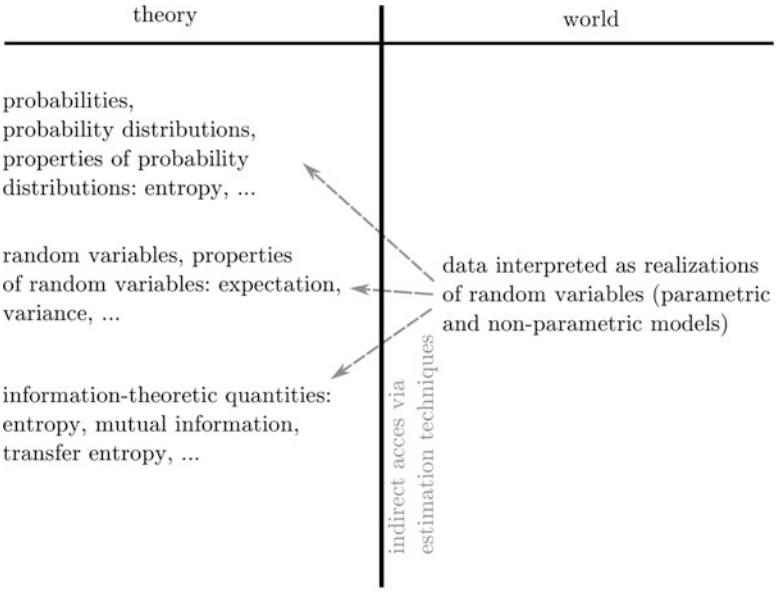

Fig. 5.1 Theoretical quantities and measurable quantities: The only things observable and accessible usually are data (measured or generated), all theoretical quantities are not directly accessible. They have to be estimated using statistical methods

Let us consider the following example. Assume that our experiment is the roll of a six-sided die. When repeating the experiment for ten times (also called *trials*), we will obtain frequencies for each of the numbers as given in Fig. 5.1. Repeating the experiment for 100 times, we will get frequencies that look similar to the ones given in Fig. 5.1. If we look at the relative frequencies (i.e., the frequency divided by the total number of trials), we see that these converge to the theoretically predicted value of $\frac{1}{6}$ as our number of trials grows larger.

This fundamental finding is also called the "Borel's law of large numbers."

Theorem 3.3 Borel's Law of Large Numbers. *Let Ω be a sample space of some experiment and let P be a probability mass function on Ω. Furthermore, let $N_n(E)$ be the number of occurrences of the event $E \subset \Omega$ when the experiment is repeated n times. Then the following holds:*

$$\frac{N_n(E)}{n} \to P(E) \quad \text{as} \quad n \to \infty.$$

Borel's law of large numbers states that if an experiment is repeated many times (where the trials have to be independent and done under identical conditions), then the relative frequency of the outcomes converge to their probability as assigned by the probability mass function. The theorem thus establishes the notion of probability as the long run relative frequency of an event occurrence and thereby connects

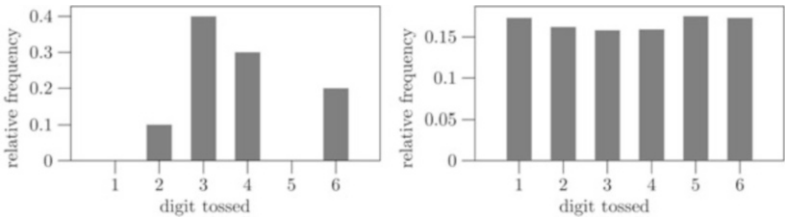

Fig. 5.2 Relative frequencies of tossed digits using a fair die: after 10 tosses (*left*) and after 1,000 tosses (*right*)

the theoretical side to the experimental side. Keep in mind though that we can never directly measure probabilities, and although relative frequencies will converge to the probability values, they will usually not be exactly equal (Fig. 5.2).

5.3.4 Independence of Events and Conditional Probabilities

A fundamental notion in probability theory is the idea of independence of events. Intuitively, we call two events independent if the occurrence of one does not affect the probability of occurrence of the other. Consider, for example, the events that it rains and the event that the current day of the week is Monday. These two are clearly independent, unless we lived in a world where there would be a correlation between the two, i.e., where the probability of rain would be different on Mondays compared to the other days of the week which is clearly not the case.

Similarly, we establish the notion of independence of two events in the sense of probability theory as follows.

Definition 3.4 Independent Events. *Let A and B be two events of some probability space* (Ω,Σ,*P*). *Then A and B are called independent if and only if*

$$P(A \cap B) = P(A)P(B). \tag{5.1}$$

The term $P(A \cap B)$ is referred to *joint probability* of A and B; see Fig. 5.3.

Another important concept is the notion of conditional probability, i.e., the probability of one event A occurring, given the fact that another event B occurred.

Definition 3.5 Conditional Probability. *Given two events A and B of some probability space* (Ω,F,*P*) *with* $P(B) > 0$ *we call*

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

the *conditional probability of A given B*.

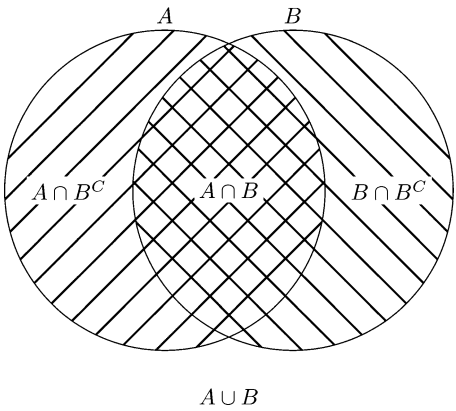

Fig. 5.3 Two events A and B, their union $A \cup B$, their intersection $A \cap B$ (i.e., common occurrence in terms of probability) and their exclusive occurrences $A \cap B^C$ (A and not B occurs), $B \cap A^C$ (B occurs and not A), where $\cdot^C$ denotes the complement in $A \cup B$

Note that for independent events A and B, we have $P(A \cap B) = P(A)P(B)$ and thus $P(A|B) = P(A)$ and $P(B|A) = P(B)$. We can thus write

$$\begin{aligned} P(A \cap B) &= P(A)P\big(B\big), \\ \Leftrightarrow P(A) &= \frac{P\big(A \cap B\big)}{P(B)} = P(A|B), \\ \Leftrightarrow P(B) &= \frac{P(A \cap B)}{P(A)} = P(B|A), \end{aligned}$$

and this means that the occurrence of A does not affect the conditional probability of B given A (and vice versa). This exactly reflects the intuitive definition of independence that we gave in the first paragraph of this section. Note that we could have also used the conditional probabilities to define independence in the first place. Nonetheless the definition of Eq. 5.1 is preferred, as it is shorter, symmetrical in A and B and more general as the conditional probabilities above are not defined in the case where $P(A) = 0$ or $P(B) = 0$.

5.3.5 *Random Variables*

In many cases the sample spaces of random experiments are a lot more complicated than the ones of the toy examples we looked at so far. Think, for example, of measurements of membrane potentials of certain neurons that we want to model mathematically, or the state of some complicated system, e.g., a network of neurons receiving some stimulus.

Thus mathematicians came up with a way to tame the sample spaces by looking at the events indirectly, namely, by first mapping the events to some better understood space, like the set of real numbers (or some higher dimensional real vector space), and then look at outcomes of the random experiment in the simplified space

rather than in the complicated original space. Looking at spaces of numbers has many advantages: order relations exist (smaller, equal, larger), we can form averages, and much more. This leads to the concept of random variables.

A (real) *random variable* is a function that maps each outcome of a random experiment to some (real) number. Thus, a random variable can be thought of as a variable whose value is subject to variations due to chance. But keep in mind that a random variable is a mapping and not a variable in the usual sense.

Mathematically, a random variable is defined using what is called a *measurable function*. A measurable function is nothing more than a map from one measurable space to another for which the pre-image of each measurable set is again measurable (with respect to the two different measures in the two measure spaces involved). So a measurable map is nothing more than a "nice" map respecting the structures of the spaces involved (take as an example for such maps the continuous functions over R).

Definition 3.6 Random Variable. *Let* (Ω,Σ,P) *be a probability space and* (Ω',Σ') *a measure space. A* (Σ,Σ')*-measurable function* $X : \Omega \to \Omega'$ *is called* Ω'*-valued random variable (or just* Ω'*-random variable) on* Ω*.*

Commonly, a distinction between *continuous random variables* and *discrete random variables* is made, the former taking values on some continuum (in most cases R) and the latter on a discrete set (in most cases Z).

A type of random variable that plays an important role in modeling is the so-called Bernoulli random variable that only takes two distinct values 0 with probability p and 1 with probability $1 - p$ (i.e., it has a Bernoulli distribution as its underlying probability distribution). Spiking behavior of a neuron is often modeled that way, where 1 stands for "neuron spiked" and 0 for "neuron did not spike" (in some interval of time).

A real- or integer-valued random variable X thus assigns a number $X(E)$ to every event $E \in \Sigma$. A value $X(E)$ corresponds to the occurrence of the event E and is called a *realization of* X. Thus, random variables allow for the change of space in which outcomes of probabilistic processes are considered. Instead of considering an outcome directly in some complicated space, we first project it to a simpler space using our mapping (the random variable X) and interpret its outcome in that simpler space.

In terms of measure theory, a random variable $X : (\Omega,\Sigma,P) \to (\Omega',\Sigma')$ (again, considered as a measurable mapping here) induces a probability measure P_X on the measure space (Ω',Σ') via

$$P_X\left(S'\right) := P\left(X^{-1}\left(S'\right)\right),$$

where again $X^{-1}(S')$ denotes the pre-image of $S' \in \Sigma'$. This also justifies the restriction of X to be measurable: If it were not, such a construction would not be possible, but this is a technical detail. As a result, this makes (Ω',Σ',P_X) a probability space and we can think of the measure P_X as the "projection" of the measure P from Ω onto Ω' (via the measurable mapping X).

The measures P and P_X are probability densities for the probability distributions over Ω and Ω': They measure the likelihood of occurrence for each event (P) or value (P_X).

As a simple example of a random variable, consider again the example of the coin toss. Here, we have $\Omega = \{\text{heads,tails}\}$, $\mathrm{F} = \{\emptyset,\{\text{heads}\},\{\text{tails}\},\Omega\}$, and P that assigns to both heads and tails the probability $\frac{1}{2}$ forming the probability space. Consider as a random variable $X : \Omega \to \Omega'$ with $\Omega' = \{0,1\}$ that maps Ω to S such that $X(\text{heads}) = 0$ and $X(\text{tails}) = 1$. If we choose $\mathrm{F}' = \{\emptyset,\{0\},\{1\},\{0,1\}\}$ as a σ-algebra for Ω', this makes $M = (\Omega',\mathrm{F}')$ a measurable space and X induces a measure $P' = P_X$ on M with $P'(\{0\}) = P'(\{1\}) = \frac{1}{2}$. That makes (Ω',F',P') a measure space, and since P' is normed, it is a probability space.

5.3.5.1 Cumulative Distribution Function

Using random variables that take on values of whole or the real numbers, the natural total ordering of elements in these spaces enables us to define the so-called cumulative distribution function (or *cdf*) for a random variable.

Definition 3.7 Cumulative Distribution Function. *Let X be a* R*-valued or* Z*-valued random variable on some probability space* (Ω,Σ,P)*. Then the function*

$$F(x) := P(X \leq x)$$

is called the *cumulative distribution function* of X.

The expression $P(X \leq x)$ evaluates to

$$P(X \leq x) = \int_{\tau \leq x} P(X = \tau)\ \mathrm{d}\tau,$$

in the continuous case and to

$$P(X \leq x) = \sum_{k \leq x} P(X = k)$$

in the discrete case.

In that sense, the measure P_X can be understood as the derivative of the cumulative distribution function F

$$P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1),$$

and we also write $F(x) = \int_{\tau \leq x} P_X(\tau)\ \mathrm{d}\tau$ in the continuous case.

5.3.5.2 Independence of Random Variables

The definition of independent events directly transfers to random variables: Two random variables X, Y are called independent if the conditional probability distribution of $X(Y)$ given an observed value of $Y(X)$ does not differ from the probability distribution of $X(Y)$ alone.

Definition 3.8 Independent Random Variables. *Let X, Y be two random variables. Then X and Y are called independent, if the following holds for any observed values x of X and y of Y:*

$$P(X|Y = y) = P(X) \quad \text{and} \quad P(Y|X = x) = P(Y).$$

This notion can be generalized to the case of three or more random variables naturally.

5.3.5.3 Expectation and Variance

Two very important concepts of random variables are the so-called *expectation value* (or just *expectation*) and the *variance*. The expectation of a random variable X is the mean value of the random variable, where the weighting of the values corresponds to the probability density distribution. It thus tells us what value of X we should expect "on average."

Definition 3.9 Expectation Value. *Let X be a* R- *or* Z*-valued random variable. Then its expectation value (sometimes also denoted by μ) is given by*

$$E[X] := \int_{\mathrm{R}} x P_X(x) \ \mathrm{d}x = \int_{\mathrm{R}} x \ \mathrm{d}P_X,$$

for a real-valued random variable X and by

$$E[X] := \sum_{x \in \mathrm{Z}} x P_X(x)$$

if X is Z-valued.

Note that if confusion can be made as to which probability distribution the expectation value is taken, we will include the probability distribution to which the expectation value is taken in the index. Consider, for example, two random variables X and Y defined on the same base space but with different underlying probability distributions. In this case, we denote by $E_X[Y]$ the expectation value of Y taken with respect to the probability distribution of X.

Let us now look at an example. If we consider the throw of a fair die with $P(i) = \frac{1}{6}$ for each digit $i = 1, \ldots, 6$ and take X as the random variable that just assigns each digit its integer value $X(i) = i$, we get $E[X] = \frac{1}{6}(1 + \ldots + 6) = 3.5$.

Another important concept is the so-called *variance* of a random variable. The variance is a measure for how far the values of the random variable are spread around its expected value. It is defined as follows.

Definition 3.10 Variance. *Let X be a* R- *or* Z*-valued random variable. Then its variance is given as*

$$\mathrm{var}[X] := E\Big[\big(E[X] - X\big)^2\Big] = E\big[X^2\big] - \big(E[X]\big)^2$$

sometimes also denoted as σ^2.

The variance is thus the expected squared distance of the values of the random variable to its expected value. Another commonly used measure is the so-called standard deviation $\sigma(X) = \sqrt{\mathrm{var}(X)}$, a measure for the average deviation of realizations of X from the mean value.

Often one also talks about the expectation value as "first-order moment" of the random variable, the variance as a "second-order moment." Higher-order moments can be constructed by iteration, but will not be of interest to us in the following.

Note again that the concepts of expectation and variance live on the theoretical side of the world, i.e., we cannot measure these quantities directly. The only thing that we can do is try to estimate them from a set of measurements (i.e., realizations of the involved random variables); see Fig. 5.1. The statistical discipline of estimation theory deals with question regarding the estimation of theoretical quantities from real data. We will talk about estimation in more detail in Sect. 5.5 and just give two examples here.

For estimating the expected value we can use what is called the *sample mean*.

Definition 3.11 Sample Mean. *Let X be a* R- *or* Z*-valued random variable with n realizations $x_1, \ldots, x_n$. Then the sample mean $\hat{\mu}$ of the realizations is given as*

$$\hat{\mu}\,(x_1, \ldots, x_n) := \frac{1}{n}\sum_{i=1}^{n} x_i$$

As we will see below, this sample mean provides a good estimation of the expected value if the number n of samples is large enough. Similarly, we can estimate the variance as follows.

Definition 3.12 Sample Variance. *Let X be a* R- *or* Z*-valued random variable with n realizations $x_1, \ldots, x_n$. Then the population variance $\hat{\sigma}$ of the realizations is given as*

$$\hat{\sigma}^2(x_1, \ldots, x_n) := \frac{1}{n}\sum_{i=1}^{n} \big(x_i - \hat{\mu}\,(x_1, \ldots, x_n)\big)^2,$$

where $\hat{\mu}$ denotes the sample mean.

Before going on let us calculate some examples of expectations and variances of random variables. Take the coin toss example from above. Here, the expected value of X is $E[X] = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1 = \frac{1}{2}$, the variance $var(X) = E\left[\left(E[X] - X\right)^2\right] = \frac{1}{2} \cdot \left(0 - \frac{1}{2}\right)^2 + \frac{1}{2} \cdot \left(1 - \frac{1}{2}\right)^2 = \frac{1}{4}$. For the example of the dice roll (where the random variable X takes the value of the number thrown) we get $E[X] = \frac{1+2+3+4+5+6}{6} = \frac{7}{2} = 3.5$ and $var(X) = E\left[X^2\right] - \left(E[X]\right)^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12} \approx 2.92$.

5.3.6 *Laws of Large Numbers*

The laws of large numbers (there exist two versions as we will see below) state that the sample average of a set of realizations of a random variable "almost certainly" converges the random variable's expected value when the number of realizations grows to infinity.

Theorem 3.13 Law of Large Numbers. *Let $X_1, X_2, \ldots$ be an infinite sequence of independent, identically distributed random variables with expected values $E(X_1) = E(X_2) = \ldots = \mu$. Let $\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$ be the sample average.*

a. **Weak law of large numbers**. The sample average converges in probability towards the expected value, i.e., for any $\varepsilon > 0$

$$\lim_{n \to \infty} P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) = 0.$$

This is sometimes also expressed as

$$\overline{X}_n \xrightarrow{P} \mu \quad when \;\; n \to \infty.$$

b. **Strong law of large numbers**. The sample average converges almost surely towards the expected value, i.e.,

$$P\left(\lim_{n \to \infty} \overline{X}_n = \mu\right) = 1.$$

This is sometimes also expressed as

$$\overline{X}_n \xrightarrow{a.s.} \mu \quad when \;\; n \to \infty.$$

The weak version of the law states that the sample average $\overline{X}_n$ is likely to be close to μ for some large value of n. But this does not exclude the possibility of $\left|\overline{X}_n - \mu\right| > \varepsilon$ occurring an infinite number of times.

The strong law says that this "almost surely" will not be the case: With probability 1, the inequality $\left|\overline{X}_n - \mu\right| < \varepsilon$ holds for all $\varepsilon > 0$ and all large enough n.

5.3.7 Some Parametrized Probability Distributions

Certain probability distributions often occur naturally when looking at typical random experiments. In the course of history, these were thus put (mathematicians like doing such things) into families or classes, and the members of one class are distinguished by a set of parameters (a parameter is just a number than can be chosen freely in some specified range). To specify a certain probability distribution we simply have to specify in which class it lies and which parameter values it exhibits, which is more convenient than specifying the probability distribution explicitly every time. This also allows proving (and reusing) results for whole classes of probability distributions and facilitates communication with other scientists.

Note that we will only give a concise version of the most important distributions relevant in neuroscientific applications here and point the reader to [47, 52, 98] for a more in-depth treatment of the subject.

The *normal distribution* $N(\mu,\sigma^2)$ is a family of continuous probability distributions parametrized by two real-valued parameters $\mu \in \mathrm{R}$ and $\sigma^2 \in \mathrm{R}^+$, called *mean* and *variance*. Its probability density function is given as

$$\begin{aligned} f(x;\mu,\sigma) : \mathrm{R} &\to \mathrm{R}_0^+ \\ x &\mapsto \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \end{aligned}$$

The family is closed under linear combinations, i.e., linear combinations of normally distributed random variables are again normally distributed. It is the most important and often used probability distribution in probability theory and statistics as many other probability distributions can be approximated by a normal distribution when the sample size is large enough (this fact is called the *central limit theorem*). See Fig. 5.4 for examples of the pdf and cdf for normally distributed random variables.

The *Bernoulli probability distribution* $Ber(p)$ describes the two possible outcomes of a Bernoulli experiment with the probability of success and failure being p and $1 - p$, respectively. It is thus a discrete probability distribution on two elements and it is parametrized by one parameter $p \in [0,1] \subset \mathrm{R}$. Its probability mass function is given by the two values $P(\text{success}) = p$ and $P(\text{failure}) = 1 - p$.

The *binomial probability distribution* $B(n,p)$ is a discrete probability distribution parametrized by two parameters $n \in \mathbf{N}$ and $p \in [0,1] \subset \mathrm{R}$. Its probability mass function is

$$f(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}, \tag{5.2}$$

and it can be thought of as a model for the probability of k successful outcomes in a trial with n independent Bernoulli experiments, each having success probability p, see Fig. 5.5.

The *Poisson distribution* $Poiss(\lambda)$ is a family of discrete probability distributions parametrized by one real parameter $\lambda \in \mathrm{R}^+$. Its probability mass function is given by

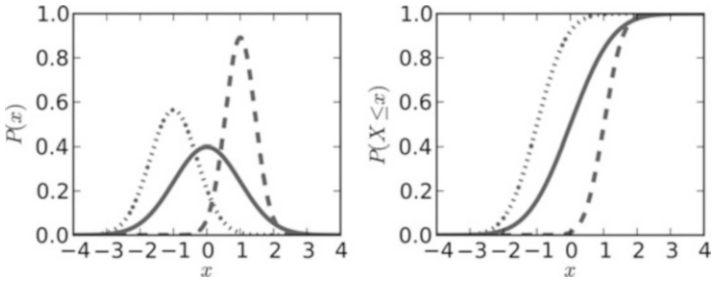

Fig. 5.4 Normal distribution: probability density function (*left*) and cumulative density function (*right*) for selected parameter values of μ and σ. *Solid line*: $\mu = 0$, $\sigma = 1$; *dashed line*: $\mu = 1$, $\sigma^2 = 0.2$; *dotted line*: $\mu = -1$, $\sigma^2 = 0.5$

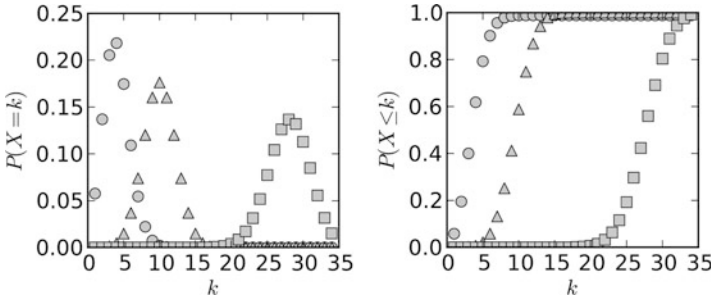

Fig. 5.5 Binomial distribution: probability mass function (*left*) and cumulative density function (*right*) for selected parameter values of p and n. *Circle*: $p = 0.2$, $n = 20$; *triangle*: $p = 0.5$, $n = 20$; *square*: $p = 0.7$, $n = 40$

$$\begin{array}{rcl} f(k;\lambda) : \mathbf{N} & \rightarrow & \mathrm{R}_0^+ \\ k & \mapsto & \dfrac{\lambda^k e^{-\lambda}}{k!}. \end{array}$$

The Poisson distribution plays an important role in the modeling of neuroscience data. This is the case because the firing statistics of cortical neurons (and also other kinds of neurons) can often be well fit by a Poisson process, where λ is considered the mean firing rate of a given neuron; see [24, 74, 101].

This fact comes at no surprise if we invest some thought. The Poisson distribution can be seen as a special case of the binomial distribution. A theorem known as Poisson limit theorem (sometimes also called "law of rare events") now tells us that in the limit $p \rightarrow 0$ and $n \rightarrow \infty$ the binomial distribution converges to the Poisson distribution with $\lambda = np$. Consider, for example, the spiking activity of our neuron that we could model via a Binomial distribution. We discretize time and consider time bins of say 2 ms and assume a mean firing rate of the neuron denoted by λ (measured in Hertz). Clearly, in most time bins the neuron does not spike (corresponding to a small value of p), and the number of bins is large (corresponding

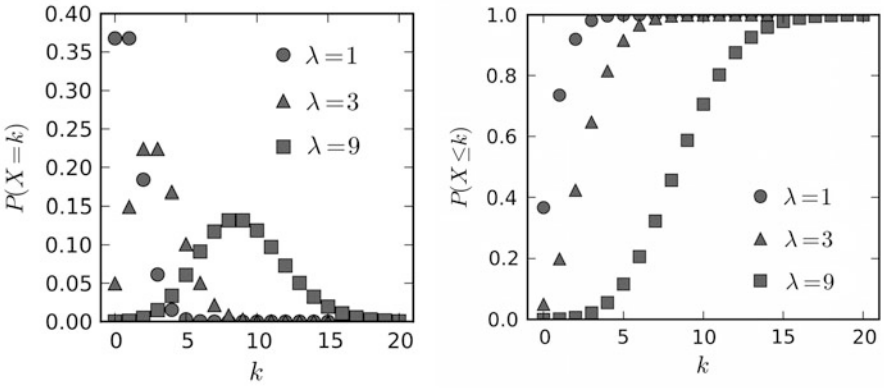

Fig. 5.6 Poisson distribution: probability mass function (*left*) and cumulative density function (*right*) for selected parameter values of λ

to a large n). The Poisson limit theorem tells us that in this case the probability distribution concerning spike emission is well matched by a Poisson distribution.

See Fig. 5.6 for examples of the pmf and cdf for Poisson-distributed random variables for a selection of parameters λ.

The so-called exponential distribution $Exp(\lambda)$ is a continuous probability distribution parametrized by one real parameter $\lambda \in \mathrm{R}^+$. Its probability density function is given by

$$\begin{aligned} f(x;\lambda) : \mathrm{R} &\to \mathrm{R}_0^+ \\ x &\mapsto \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}. \end{aligned}$$

The exponential distribution with parameter λ can be interpreted as the probability distribution describing the time between two events in a Poisson process with parameter λ; see the next section.

See Fig. 5.7 for examples of the pdf and cdf for exponentially distributed random variables for a selection of parameters λ.

We want to conclude our view on families on probability distributions at this point and point the interested reader to [47, 52, 98] regarding further examples and details of families of probability distributions.

5.3.8 Stochastic Processes

A *stochastic process* (sometimes also called *random process*) is a collection of random variables indexed by a totally ordered set, which is usually taken as time. Stochastic processes are commonly used to model the evolution of some random variable over time. We will only look at discrete-time processes in the following,

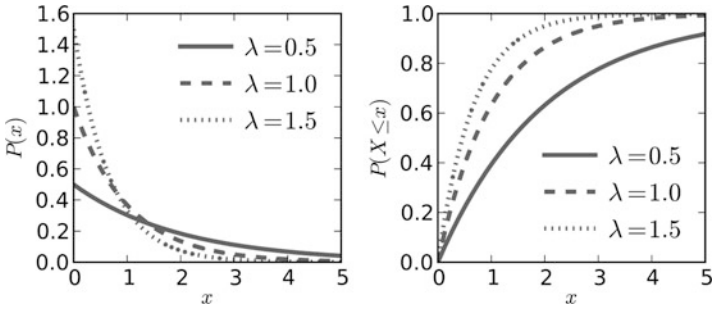

Fig. 5.7 Exponential distribution: probability density function (*left*) and cumulative density function (*right*) for selected parameter values of λ

i.e., stochastic processes that are indexed by a discrete set. The extension to the continuous case is straightforward; see [18] for an introduction to the subject.

Mathematically, a stochastic process is defined as follows.

Definition 3.14. *Let* (Ω,**F**,*P*) *be a probability space and let* (*S*,**S**) *be a measure space. Let furthermore* $X_t : \mathbf{F} \to \mathbf{S}$ *be a set of random variables, where* $t \in T$. *Then an S-valued stochastic process* P *is given by*

$$\mathbf{P} := \{X_t : t \in T\},$$

where T is some totally ordered set, commonly interpreted as time. The space S is referred to as the *sample space of the process* P.

If the distribution underlying the random variables X_t does not vary over time, the process is called *homogeneous*, in the case where the probability distributions P_{X_t} depend on the time t, it is called *inhomogeneous*.

A special kind and well-studied type of stochastic process is the so-called Markov process. A discrete Markov process of order $k \in \mathbf{N}$ is a inhomogeneous stochastic process subject to the restriction that for any time $t = 0, 1, \ldots$, the probability distribution underlying X_t only depends on the preceding k probability distributions of $X_{t-1}, \ldots, X_{t-k}$, i.e., that for any t and any set of realizations x_i of X_i ($0 \leq i \leq t$), we have

$$P(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_{t-k} = x_{t-k}) = P(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_0 = x_0).$$

Another process often considered in neuroscientific applications is the *Poisson process*. It is a discrete-time stochastic process P for which the random variables are Poisson distributed with some parameter $\lambda(t)$ (in the inhomogeneous case, for the homogeneous case, we have $\lambda(t) = \lambda = \text{constant}$). As can be shown, the time delay between each pair of consecutive events of a Poisson process is exponentially distributed. See Fig. 5.8 for examples of the number of instantaneous (occurring during one time slice) and the number of cumulated events (over all preceding time slices) of Poisson processes for a selection of parameters λ.

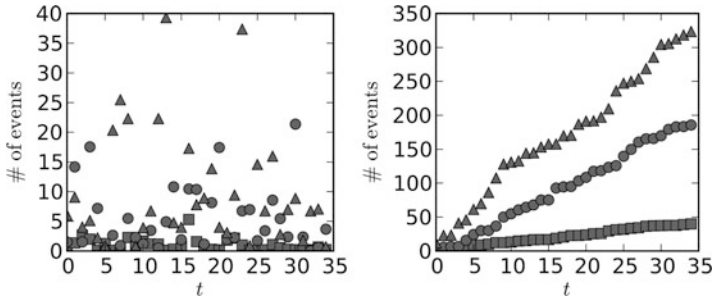

Fig. 5.8 Examples of the number of events in one time window of size $\Delta t = 1$ (*left*) and the number of accumulated events since $t = 0$ (*right*) for Poisson processes with certain rates $\lambda = 1$ (*circle*), $\lambda = 5$ (*triangle*) and $\lambda = 10$ (*square*)

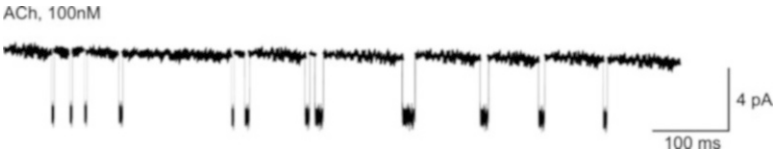

Fig. 5.9 Random opening and closing of ion channels (Modified from [23], Fig. 12)

Poisson processes have proven to be a good model for many natural as well as man-made processes such as radioactive decay, telephone calls and queues, and also for modeling neural data. An influential paper in the neurosciences was [23], showing the random nature of the closing and opening of single ion channels in certain neurons. Using as a model a Poisson process with the right parameter provides a good fit to the measured data here Fig. 5.9.

Another prominent example of neuroscientific models employing a Poisson process is the commonly used model for the sparse and highly irregular firing patterns of cortical neurons in vivo [24, 74, 101]. The firing patterns of such cells are usually modeled using inhomogeneous Poisson processes (with $\lambda(t)$ modeling the average firing rate of a cell).

5.4 Information Theory

Information theory was introduced by Shannon [97] as a mathematically rigid theory to describe the process of transmission of information over some channel of communication. His goal was to quantitatively measure the "information content" of a "message" sent over some "channel"; see Fig. 5.10. In what follows we will not go into detail regarding all aspects of Shannon's theory, but we will mainly focus on his idea of measuring "information content" of a message. For a more in-depth treatment of the subject, the interested reader is pointed to the excellent book [26].

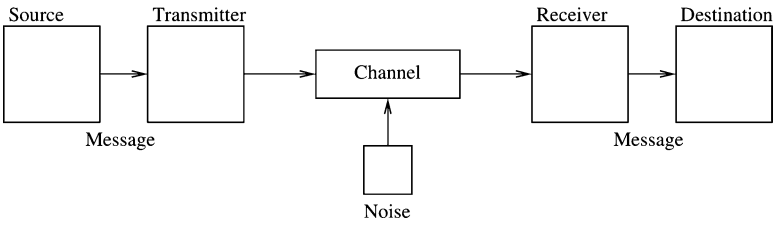

Fig. 5.10 The setting of Shannon's information theory: information is transferred from a source to a destination via a message that is first encoded and then subsequently sent over a noisy channel to be decoded by the receiver

The central elements of Shannon's theory are depicted in Fig. 5.10. In the standard setting considered in information theory, an *information source* produces *messages* that are subsequently encoded using *symbols* from an *alphabet* and sent over a noisy *channel* to be received by a *receiver* that decodes the message and attempts to reconstruct the original message.

A communication channel (or just channel) in Shannon's model transmits the encoded message from the sender to the receiver. Due to noise present in the channel, the receiver does not receive the original message dispatched by the sender but rather some noisy version of it.

The whole theory is set in the field of probability theory (hence our introduction to the concepts in the last section) and in this context, the messages emitted by the source are modeled as a random variable X with some underlying probability distribution P_X. For each message x (a realization of X), the receiver sees a corrupted version y of x and this fact is modeled by interpreting the received messages as realizations of a random variable Y with some probability distribution P_Y (that depends both on P_X and the channel properties). The transmission characteristics of the channel itself are characterized by the stochastic correspondence of the signals transmitted by the sender to the ones received by the receiver, i.e., by modeling the channel as a conditional probability distribution $P_{Y|X}$.

Being based upon probability theory, keep in mind that all the information-theoretic quantities that we will look at in the following such as "entropy" or "mutual information" are just properties of the random variables involved, i.e., properties of the probability distributions underlying these random variables.

Information-theoretic analyses have proven to be a valuable tool in many areas of science such as physics, biology, chemistry, finance, and linguistics and generally in the study of complex systems [62, 88]. We will have a look at applications in the neurosciences in Sect. 5.6.

Note that a vast number of works was published in the field of information theory and its applications since its first presentation in the 1950s. We will focus on the core concepts in the following and point the reader to [26] for a more in-depth treatment of the subject.

In the following we will start by looking at a notion of information and using this proceed to define *entropy* (sometimes also called *Shannon entropy*), a core concept

in information theory. As all further information-theoretic concepts are based on the idea of entropy, it is of vital importance to understand this concept well. We will then look at mutual information, the information shared by two or more random variables. Furthermore, we will look at a measure of distance for probability distributions called Kullback–Leibler divergence and give an interpretation of mutual information in terms of Kullback–Leibler divergence. After a quick look at the multivariate case of mutual information between more than two variables and the relation between mutual information and channel capacity, we will then proceed to an information-theoretic measure called transfer entropy. Transfer entropy is based on mutual information but in contrast to mutual information is of directed nature.

5.4.1 A Notion of Information

Before defining entropy, let us try to give an axiomatic definition of the concept of "information", see [115]. The entropy of a random variable will then be nothing more than the expected (i.e., average) amount of information contained in a realization of that random variable.

We want to consider a probabilistic model in what follows, i.e., we have a set of events, each occurring with a given probability. The goal is to assess how informative the occurrence of a given event is. What would we intuitively expect from a measure of information h that maps the set of the events to the set of nonnegative real number, i.e., when we restrict h to be a non-negative real number?

First of all, it should certainly be additive for independent events and sub-additive for non-independent events. This is easily justified: If you read two newspaper articles about totally unrelated subjects, the total amount of information you obtain consists of both the information in the first and the second article. When you read articles about related subjects on the other hand, they often have some common information.

Furthermore, events that occur regularly and unsurprisingly are not considered informative and the more seldom or surprising an event occurs, the more informative it is. Think about an article about your favorite sports team winning a match that usually wins all matches. You will consider this not very informative. But when the local newspaper reports about an earthquake with its epicenter being in the part of town where you live, this will certainly be informative to you (unless you were at home during the time the earthquake happened), assuming that earthquakes do not occur on a regular basis where you live.

We thus have the following axioms for the information content h of an event, where we look at the information content of events contained in some probability space (Ω,Σ,P):

1. h is nonnegative: $h : \Sigma \to \mathbf{R}^+$.
2. h is sub-additive: For any two messages $\omega_1, \omega_2 \in \Sigma$, we have $h(\omega_1 \cap \omega_2) \leq h(\omega_1) + h(\omega_2)$, where equality holds if and only if ω_1 and ω_2 are independent.

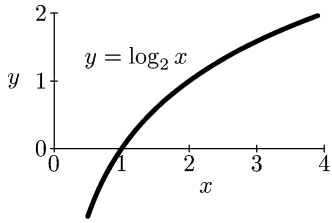

Fig. 5.11 The logarithm to the basis of 2

3. h is continuous and monotonic with respect to the probability measure P.
4. Events with probability 1 are not informative: $h(\omega) = 0$ for $\omega \in \Sigma$ with $P(\omega) = 1$.

Now calculus tells us (this is not hard to show – you paid attention in the mathematics class at school, did you not?) that these four requirements leave only one possible function that fulfills all these requirements: the logarithm (Fig. 5.11). This leads us to the following natural definition.

Definition 4.1 Information. *Let* (Ω,Σ,P) *be a probability space. Then the information* h *of an event* $\sigma \in \Sigma$ *is defined as*

$$h(\sigma) := h(P(\sigma)) = -\log_b(P(\sigma)),$$

where b denotes the basis of the logarithm.

For the basis of the logarithm, usually $b = 2$ or $b = e$ is chosen, fixing the unit of h as "bit" or "nat," respectively. We resort to using $b = 2$ for the rest of this chapter and write *log* for the logarithm to the basis of two. The natural logarithm will be denoted by *ln*.

Note that the information content in our definition only depends on the probability of the occurrence of the event and not the event itself. It is thus a property of the probability distribution P.

Let us give some examples in order to illustrate this idea of information content.

Consider a toss of a fair coin, where the possible outcomes are heads (H) or tails (T), each occurring with probability $\frac{1}{2}$. What is the information contained in a coin toss? As the information solely depends on the probability, we have $h(H) = h(T)$, which comes at no surprise. Furthermore we have $h(H) = h(T) = -\log\frac{1}{2} = -(\log(1) - \log_2(2)) = \log 2 = 1$ bit, when we apply the fundamental logarithmic identity $\log(a \cdot b) = \log(a) + \log(b)$. Thus one toss of a fair coin gives us one bit of information. This fact also lets us explain the unit attached to h. If measured in bit (i.e., with $b = 2$), this is the amount of bits needed to store that information. For the toss of a coin we need one bit, assigning each outcome to either 0 or 1.

Repeating the same game for the roll of a fair die where each digit has probability $\frac{1}{6}$, we again have the same amount of information for each digit $E \in \{1, \ldots,6\}$, namely, $h(E) = \log(6) \approx 2.58$ bit. This means that in this case we need three bits to store the information associated to each outcome, namely, the number shown.

Looking at the two examples above, we can give another (hopefully intuitive) characterization of the term information content: It is the minimal number of yes-no questions that we have to ask until we know which event occurred, assuming that we have a knowledge of the underlying probability distribution. Consider the example of the coin toss above. We have to ask exactly one question and we know the outcome ("Was it heads?" "Was it tails?").

Things get more interesting when we look at the case of the die throw. Here, several question asking strategies are possible and you can freely choose your favorite – we will give one example below.

Say a digit d was thrown. The first question could be, "Was the digit less or equal to 3?" (other strategies, "Was the digit greater than 3?" "Was the digit even?" "Was the digit odd?"). We then go on depending on the answer and cut off at least half of the remaining probability mass in each step, leaving us with a single possibility after at most 3 steps. From the information content, we know that on average we have to ask 2.58 times on average.

The two examples above were both cases with uniform probability distributions but in principle the same applies to arbitrary probability distributions.

5.4.2 Entropy as Expected Information Content

The term entropy is at the heart of Shannon's information theory [97]. Using the notion of the information as discussed in Sect. 4.1, we can readily define the entropy of a discrete random variable as its expected information.

Definition 4.2 Entropy. *Let X be a random variable on some probability space (Ω,Σ,P) with values in the integer or the real numbers. Then its entropy*[1] *(sometimes also called Shannon entropy or self-information) $H(X)$ is defined as the expected amount of information of X,*

$$H(X) := E[h(X)]. \tag{5.3}$$

If X is a random variable that takes integer values (i.e., a discrete random variable), Eq. 5.3 evaluates to

$$H(X) = \sum_{x\in\mathrm{Z}} P(X=x)h(P(X=x)) = -\sum_{x\in\mathrm{Z}} P(X=x)\log(P(X=x)),$$

in the case of a real-valued, continuous random variable, we get

[1] Shannon chose the letter H for denoting entropy after Boltzmann's H-theorem in classical statistical mechanics.

$$H(X) = \int_{\mathrm{R}} P(X = x) h(P(X = x)) \ \mathrm{d}x$$

and the resulting quantities is called *differential entropy* [26].

As the information content is a function solely dependent on the probability of the events one also speaks of the entropy of a probability distribution.

Looking at the definition in Eq. 5.3, we see that entropy is a measure for the average amount of information that we expect to obtain when looking at realizations of a given random variable X. An equivalent characterization would be to interpret it as the average information one is missing when one would not know the value of the random variable (i.e., its realization), and a third one would be to interpret it as the average reduction of uncertainty about the possible values of a random variable having observed one or more realizations.

Akin to the information content h, entropy H is a dimensionless number and usually measured in bits (i.e., the expected number of binary digits needed to store the information), taking a logarithm to the base of 2.

Shannon entropy has many applications as we will see in the following and constitutes the core of all things labeled "information theory." Let us thus look a bit closer at this quantity.

Lemma 4.3. *Let X be some discrete random variable. Then its entropy $H(X)$ satisfies the two inequalities*

$$0 \leq H(X) \leq \log(n).$$

Note that the first inequality is a direct consequence of the properties of the information content, and the second follows from Gibbs' inequality [26].

With regard to entropy, probability distributions having maximal entropy are often of interest in applications as they can be seen as the least restricted ones (i.e., having the least a priori assumptions), given the model parameters. The *principle of maximum entropy* states that when choosing among a set of probability distributions with certain fixed properties, the preference should be given to distributions that have the maximal entropy among all considered distributions. This choice is justified as the one making the fewest assumptions on the shape of the distribution apart from the prescribed properties.

For discrete probability distributions, the uniform distribution is the one with the highest entropy among all other distributions on the same base set. This can be well seen in the example in Fig. 5.12: The entropy of a Bernoulli distribution takes its maximum at $p = 1/2$, the parameter value for which it corresponds to the uniform probability distribution on the two elements 0 and 1, each occurring with probability 1/2.

For continuous, real-valued random variables with a given finite mean μ and variance σ^2, the normal distribution with mean μ and variance σ^2 has highest entropy. Demanding non-negativity and a non-vanishing probability on the positive real numbers (i.e., an infinite support) with positive given mean μ yields the exponential distribution with parameter $\lambda = 1/\mu$ as a maximum entropy distribution.

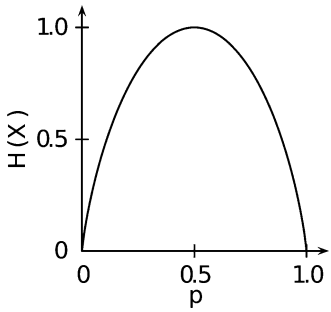

Fig. 5.12 Entropy $H(X)$ of a Bernoulli random variable X as a function of success probability $p = P(X = 1)$. The maximum is attained at $p = 1/2$

Examples

Before continuing, let us now compute some more entropies in order to get a feeling for this quantity.

For a uniform probability distribution P on n events $\Omega = \{\omega_1, \ldots, \omega_n\}$ each event has probability $P(\omega_i) = 1/n$ and we obtain

$$H(P) = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} = \log n,$$

as the maximal entropy for all discrete probability distributions on the set Ω.

Let us now compute the entropy of a Bernoulli random variable X, i.e., a binary random variable X taking values 0 and 1 with probability p and $1 - p$, respectively. For the entropy of X we get

$$H(X) = -(p \ \log \ p + (1-p) \ \log(1-p)).$$

See Fig. 5.12 for a plot of the entropy seen as a function of the success probability p. As expected, the maximum is attained at $p = 1/2$, corresponding to the case of the uniform distribution.

Computing the differential entropy of a normal distribution $N(\mu,\sigma^2)$ with mean μ and variance σ^2 yields

$$H\big(\big(N\big(\mu, \sigma^2\big)\big) = \frac{1}{2} \log\big(2\pi e \sigma^2\big),$$

and we see that the entropy does not depend on the mean value of the distribution but just its variance. This is not surprising, as the shape of the probability distribution is only changed by σ^2 and not μ.

For an example of how to compute the entropy of spike trains, see Sect. 5.6.

5.4.2.1 Joint Entropy

Generalizing the notion of entropy to two or more variables, we can define the so-called joint entropy to quantify the expected uncertainty (or expected information) in a joint distribution of random variables.

Definition 4.4 Joint Entropy. *Let X and Y be discrete random variables on some probability spaces. Then the joint entropy of X and Y is given by*

$$H(X,Y) = -E_{X,Y}[\log\ P(x,y)] = -\sum_{x,y} P(x,y)\ \log P(x,y), \tag{5.4}$$

where $P_{X,Y}$ denotes the joint probability distribution of X and Y and the sum runs over all possible values x and y of X and Y, respectively.

This definition allows a straightforward extension to the case of more than two random variables.

The conditional entropy $H(X|Y)$ of two random variables X and Y quantifies the expected uncertainty (respectively expected information) remaining in a random variable X under the condition that a second variable Y was observed or equivalently as the reduction of the expected uncertainty in X upon the knowledge of Y.

Definition 4.5 Conditional Entropy. *Let X and Y be discrete random variables on some probability spaces. Then the conditional entropy of X given Y is given by*

$$H(X|Y) = -E_{X,Y}[\log\ P(x|y)] = -\sum_{x,y} P(x,y)\ \log P(x|y),$$

where $P_{X,Y}$ denotes the joint probability distribution of X and Y.

5.4.3 Mutual Information

In this section we will introduce the notion of mutual information, an entropy-based measure for the information shared between two (or more) random variables. Mutual information can be thought of as a measure for the mutual dependence of random variables, i.e., as a measure for how far they are from being independent.

We will give two different approaches to this concept in the following: a direct one based on the point-wise mutual information i and one using the idea of conditional entropy. Note that in essence, these are just different approaches to defining the same object. We give the two approaches in the following, hoping that they help in understanding the concept better. In Sect. 4.4 we will see yet another characterization in terms of the Kullback–Leibler divergence.

5.4.3.1 Point-Wise Mutual Information

In terms of information content, the case of considering two events that are independent is straightforward: One of the axioms tells us that the information content of the two events occurring together is the sum of the information contents of the single events. But what about the case where the events are non-independent? In this case we certainly have to consider the conditional probabilities of the two

events occurring: If one event often occurs given that the other one occurs (think of the two events "It is snowing" and "It is winter"), the information overlap is higher than when the occurrence of one given the other is rare (think of "It is snowing" and "It is summer").

Using the notion of information from Sect. 4.1, let us express this in a mathematical way by defining the *mutual information* (i.e., shared information content) of two events. We call this the *point-wise mutual information* or *pmi*.

Definition 4.6 Point-Wise Mutual Information. *Let x and y be two events of a probability space* (Ω,Σ,P). *Then their point-wise mutual information (pmi) is given as*

$$\begin{aligned} i(x;y) : &= -\log \frac{P(x,y)}{P(x)P(y)} \\ &= -\log \frac{P(x|y)}{P(x)} \\ &= -\log \frac{P(y|x)}{P(y)}. \end{aligned} \tag{5.5}$$

Note that we used joint probability distribution of x and y is for the definition of $i(x;y)$ to avoid the ambiguities introduced by the conditional distributions. Yet, the latter are probably the easier way to gain a first understanding of this quantity.

Let us note that this measure of shared information is symmetric ($i(x;y) = i(y;x)$) and that it can take any real value, particularly also negative values. Such negative values of point-wise mutual information are commonly referred to as *misinformation* [64]. Point-wise mutual information is zero if the two events x and y are independent and it is bounded above by the information content of x and y. More generally, the following inequality holds:

$$-\infty \leq i(x;y) \leq \min\left\{ \underbrace{-\log\ P(x)}_{=h(x)}, \underbrace{-\log\ P(y)}_{=h(y)} \right\}.$$

Defining the information content of the co-occurrence of x and y as

$$i(x,y) := -\log\ P(x,y),$$

another way of writing the point-wise mutual information is

$$\begin{aligned} i(x;y) &= i(x) + i\big(y\big) - i(x,y\big), \\ &= i(x) - i(x|y), \\ &= i(y) - i\big(y|x\big), \end{aligned} \tag{5.6}$$

where the first identity above is readily obtained from Eq. 5.5 by just expanding the logarithmic term, and in the second and third line the formula for the conditional probability was used.

Table 5.1 Table of joint probabilities $P(\omega_a,\omega_b)$ of two events ω_a and ω_b

ω_a	ω_b	P(x,y)
a_1	b_1	0.2
a_1	b_2	0.5
a_2	b_1	0.25
a_2	b_2	0.05

Before considering mutual information of random variables as a straightforward generalization of the above, let us look at an example.

Say we have two probability spaces (Ω_a,Σ_a,P_a) and (Ω_b,Σ_b,P_b), with $\Omega_a = \{a_1,a_2\}$ and $\Omega_b = \{b_1,b_2\}$. We want to compute the point-wise mutual information of two events $\omega_a \in \Omega_a$ and $\omega_b \in \Omega_b$ subject to the joint probability distributions of ω_a and ω_b as given in Table 5.1. Note that the joint probability distribution can also be written as matrix

$$P(\omega_a, \omega_b) = \begin{pmatrix} 0.2 & 0.5 \\ 0.25 & 0.05 \end{pmatrix},$$

if we label rows by possible outcomes of ω_a and columns by possible outcomes of ω_b. The marginal distributions $P(\omega_a)$ and $P(\omega_b)$ are now obtained as row, respectively, column sums as $P(\omega_a = a_1) = 0.7$, $P(\omega_a = a_2) = 0.3$, $P(\omega_b = b_1) = 0.45$, and $P(\omega_b = b_2) = 0.55$.

We can now calculate the point-wise mutual information of, for example,

$$i(a_2; b_2) = -\log \frac{0.05}{0.3 \cdot 0.55} \approx 1.7 \text{ bits},$$

and

$$i(a_1; b_1) = -\log \frac{0.2}{0.7 \cdot 0.45} \approx -0.65 \text{ bits}.$$

Note again that in contrast to mutual information (that we will discuss in the next section), point-wise mutual information can take negative values called; see [64].

5.4.3.2 Mutual Information as Expected Point-Wise Mutual Information

Using point-wise mutual information, the definition of mutual information of two random variables is straightforward: Mutual information of two random variables is the expected value of the point-wise mutual information of all realizations.

Definition 4.7 Mutual Information. *Let X and Y be two discrete random variables. Then the mutual information $I(X;Y)$ is given as the expected point-wise mutual information:*

$$\begin{aligned} I(X;Y): &= E_{X,Y}[i(x;y)] \\ &= \sum_y \sum_x P(x,y) i(x;y) \\ &= -\sum_y \sum_x P(x,y) log\left(\frac{P(x,y)}{P(x)P(y)}\right), \end{aligned} \tag{5.7}$$

where the sums are taken over all possible values x of X and y of Y.

Remember again that the joint probability $P(x,y)$ is just a two-dimensional matrix where the rows are indexed by X-values and the columns by Y-values and that each row (column) tells us how likely each possible value of $Y(X)$ is, given the value x of X (y of Y) determined by the row (column) index. The rows (columns) sum to the marginal probability distributions $P(x)$ ($P(y)$), that can be written as vectors.

If X and Y are continuous random variables we just replace the sums by integrals and obtain what is known as *differential mutual information*:

$$I(X;Y) := \int_{\mathrm{R}} \int_{\mathrm{R}} P(x,y) \log\left(\frac{P(x,y)}{P(x)P(y)}\right) \, \mathrm{d}x \, \mathrm{d}y. \tag{5.8}$$

Here $P(x,y)$ denotes the joint probability distribution function of X and Y, and $P(x)$ and $P(y)$ the marginal probability distribution functions of X and Y, respectively.

As we can see, mutual information can be interpreted as the information (i.e., entropy) shared by the two variables, hence its name. Like point-wise mutual information, it is a symmetric quantity $I(X;Y) = I(Y;X)$ and in contrast to point-wise mutual information it is nonnegative, $I(X;Y) \geq 0$. Note though that it is not a metric, as in the general case it does not satisfy the triangle inequality. Furthermore, we have $I(X;X) = H(X)$, and this identity is the reason why entropy is sometimes also called *self-information*.

Taking the expected value of Eq. 5.5 and using the notion of conditional entropy, we can define mutual information between two random variables as follows:

$$\begin{aligned} I(X;Y): &= H(X) + H\big(Y\big) - H\big(X,Y\big), \\ &= H(X) - H(X|Y), \\ &= H(Y) - H\big(Y|X\big), \end{aligned} \tag{5.9}$$

where in the last two steps the identity $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$ was used. Note that Eq. 5.9 is the generalization of Eq. 5.6 to the case of random variables. See Fig. 5.13 for an illustration of the relation between the different entropies and mutual information.

A possible interpretation of mutual information of two random variables X and Y is to consider it as a measure for the shared entropy between the two variables.

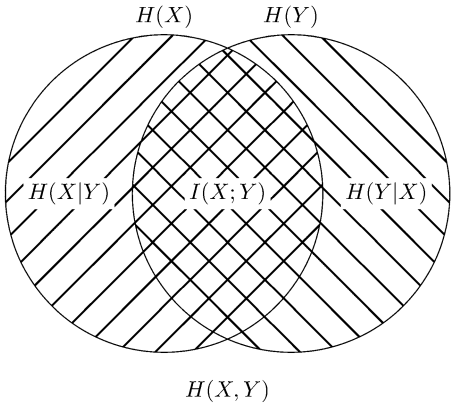

Fig. 5.13 Venn diagram showing the relation between the entropies $H(X)$ and $H(Y)$, the joint entropy $H(X,Y)$, the conditional entropies $H(X|Y)$ and $H(Y|X)$, and mutual information $I(X;Y)$

5.4.3.3 Mutual Information and Channel Capacities

We will look at channels in Shannon's sense of communication in the following and relate mutual information to channel capacity. But rather than looking at the subject in its full generality, we restrict ourselves to discrete, memoryless channels. The interested reader is pointed to [26] for a more thorough treatment of the subject.

Let us take as usual X and Y for the signal transmitted by some sender and received by some receiver, respectively. In terms of information transmission, we can interpret mutual information $I(X;Y)$ as the average amount of information the received signal constrains about the transmitted signal, where the averaging is done over the probability distribution of the source signal P_X. This makes mutual information a function of P_X and $P_{Y|X}$ and as we know, it is a symmetric quantity.

Shannon defines the *capacity* C of some channel as the maximum amount of information that a signal Y received by the receiver can contain about the signal X transmitted through the channel by the source.

In terms of mutual information $I(X;Y)$ we can define the channel capacity as the maximum mutual information $I(X;Y)$ among all realizations of the signal X. Channel capacity is thus not dependent on the distribution of P_X of X but rather a property of the channel itself, i.e., a property of the conditional distribution $P_{Y|X}$ and as such asymmetric and causal [85, 112].

Note that channel capacity is bound from below by 0 and from above by the entropy $H(X)$ of X, with the maximal capacity being attained by a noise-free channel. In the presence of noise the capacity is lower.

We will have a look at channels again when dealing with applications of the theory in Section 6.

5.4.3.4 Normalized Measures of Mutual Information

In many applications one is often interested in making values of mutual information comparable by employing a suitable normalization. Consequently, there exists a

variety of proposed normalized measures of mutual information, most based on the simple idea of normalizing by one of the entropies that appear in the upper bounds of the mutual information. Using the entropy of one variable as a normalization factor, there are two possible choices and both were proposed: The so-called *coefficient of constraint* $C(X|Y)$ [25]

$$C(X|Y) := \frac{I(X;Y)}{H(Y)}$$

and the *uncertainty coefficient* $U(X|Y)$ [105]

$$U(X|Y) := \frac{I(X;Y)}{H(X)}.$$

These two quantities are obviously nonsymmetric but can easily be symmetrized, for example, by setting

$$U(I,J) := \frac{H(I)U(I|J) + H(J)U(J|I)}{H(I) + H(J)}.$$

Another symmetric normalized measure for mutual information, usually referred to as *redundancy measure*, is obtained when normalizing using the sum of the entropy of the variables

$$R = \frac{I(X;Y)}{H(X) + H(Y)}.$$

Note that R takes its minimum of 0 when the two variables are independent and its maximum when one variable is completely redundant knowing the other.

Note that the list of normalized variants of mutual information given here is far from complete. But as said earlier, the principle behind most normalizations is to use one or a combination of the entropies of the involved random variables as a normalizing factor.

5.4.3.5 Multivariate Case

What if we want to calculate the mutual information between not only two random variables but rather three or more? A natural generalization of mutual information to this so-called *multivariate* case is given by the following definition using conditional entropies and is also called *multi-information* or *integration* [107].

The mutual information of three random variables $X_1, \ldots, X_3$ is given by

$$I(X_1;X_2;X_3) := I(X_1;X_2) - I(X_1;X_2|X_3),$$

where the last term is defined as

$$I(X_1; X_2|X_3) := E_{X_3}[I(X_1; X_2)|X_3],$$

the latter being called the *conditional mutual information* of X_1 and X_2 given X_3. The conditional mutual information $I(X_1; X_2|X_3)$ can also be interpreted as the average common information shared by X_1 and X_2 that is not already contained in X_3.

Inductively, the generalization to the case of n random variables $X_1, \ldots, X_n$ is straightforward:

$$I(X_1; \ldots; X_n) := I(X_1; \ldots; X_{n-1}) - I(X_1; \ldots; X_{n-1}|X_n),$$

where the last term is again the conditional mutual information

$$I(X_1; \ldots; X_{n-1}|X_n) := E_{X_n}[I(X_1; \ldots; X_{n-1})|X_n].$$

Beware that while the interpretations of mutual information directly generalize from the bivariate case $I(X;Y)$ to the multivariate case $I(X_1; \ldots;X_n)$, there is an important difference between the bivariate and the multivariate measure. Whereas mutual information $I(X;Y)$ is a nonnegative quantity, multivariate mutual information (MMI for short) behaves a bit differently than the usual mutual information in the aspect that it can also take negative values which makes this information-theoretic quantity sometimes difficult to interpret.

Let us first look at an example of three variables with positive MMI. To make things a bit more hands on, let us look at three binary random variables, one telling us whether it is cloudy, the other whether it is raining, and the third one whether it is sunny. We want to compute $I(rain; no\ sun; cloud)$. In our model, clouds can cause rain and can block the sun, and so we have

$$I(\mathrm{rain}; \mathrm{no\ \ sun}|\mathrm{cloud}) \leq I(\mathrm{rain}; \mathrm{no\ \ sun}),$$

as it is more likely that it is raining and there is no sun visible when it is cloudy than when there are no clouds visible. This results in positive MMI for $I(rain; no\ sun; cloud)$, a typical situation for a common-cause structure in the variables: here, the fact that the sun is not shining can partly be due to the fact that it is raining and partly due to the fact that there are clouds visible.

In a sense the inverse is the situation where we have two causes with a common effect: This situation can lead to negative values for the MMI; see [67]. In this situation, observing a common effect induces a dependency between the causes that did not exist before. This fact is called "explaining away" in the context of Bayesian networks; see [84]. Pearl [84] also gives a car-related example where the three (binary) variables are "engine fails to start" (X), "battery dead" (Y), and "fuel pump broken" (y). Clearly, both Y and Z can cause X and are uncorrelated if we have no knowledge of the value of X. But fixing the common effect X, namely, observing that the engine did not start, induces a dependency between Y and Z that can lead to negative values of the MMI.

Another problem with the n-variate case to keep in mind is the combinatorial explosion of the degrees of freedom regarding their interactions. As a priori every nonempty subset of the variables could interact in an information-theoretic sense, this yields 2^{n-1} degrees of freedom.

5.4.4 A Distance Measure for Probability Distributions: The Kullback–Leibler Divergence

The *Kullback–Leibler divergence* [57] (or *KL divergence* for short) is a kind of "distance measure" on the space of probability distributions: Given two probability distributions on the same base space Ω interpreted as two points in the space of all probability distributions over the base set Ω, it tells us how far they are "apart."

We again use the usual expectation-value construction as used for the entropy before.

Definition 4.8 Kullback–Leibler Divergence. *Let P and Q be two discrete probability distributions over the same base space Ω. Then the Kullback–Leibler divergence of P and Q is given by*

$$D_{\mathrm{KL}}(P||Q) := \sum_{\omega\in\Omega} P(\omega)\log\frac{P(\omega)}{Q(\omega)}. \tag{5.10}$$

The Kullback–Leibler divergence is nonnegative $D_{\mathrm{KL}}(P\|Q) \geq 0$ (and it is zero if P equals Q almost everywhere), but it is not a metric in the mathematical sense as in general it is nonsymmetric $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$, and it does not fulfill the triangle inequality. Note that in their original work, Kullback and Leibler [57] defined the divergence via the sum

$$D_{\mathrm{KL}}(P||Q) + D_{\mathrm{KL}}(Q||P),$$

making it a symmetric measure. $D_{\mathrm{KL}}(P\|Q)$ is additive for independent distributions, namely,

$$D_{\mathrm{KL}}(P||Q) = D_{\mathrm{KL}}(P_1||Q_1) + D_{\mathrm{KL}}(P_2||Q_2),$$

where the two pairs P_1, P_2 and Q_1, Q_2 are independent probability distributions with the joint distributions $P = P_1P_2$ and $Q = Q_1Q_2$, respectively.

Note that the expression in Eq. 5.10 is nothing else than the expected value $E_P[\log P - \log Q]$ with the expectation value taken with respect to P, which in term can be interpreted as "expected distance of P and Q," measured in terms of the information content. Another interpretation can be given in the language of codes: $D_{\mathrm{KL}}(P\|Q)$ is the average number of extra bits needed to code samples from P using a code book based on Q.

Analogous to previous examples, the KL divergence can also be defined for continuous random variables in a straightforward way via

$$D_{\mathrm{KL}}(P||Q) = \int_{\mathrm{R}} p(x)\log\left(\frac{p(x)}{q(x)}\right)\,\mathrm{d}x,$$

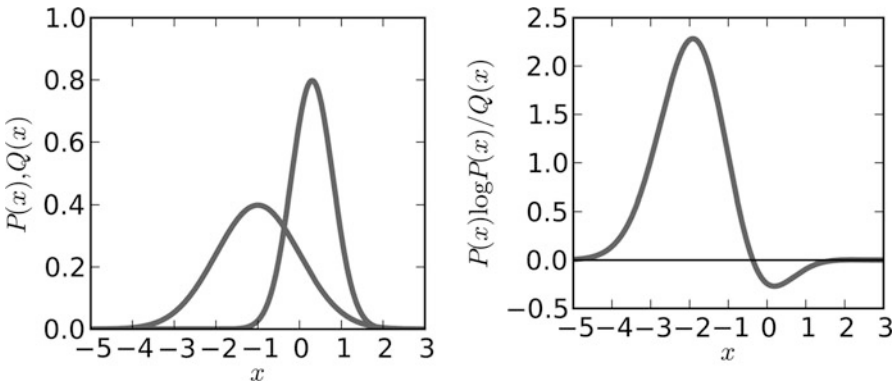

Fig. 5.14 The probability densities of two Gaussian probability distributions (*left*) and the quantity $P(x)logP(x)/Q(x)$ that yields the KL divergence when integrated (*right*)

where P_1, P_2 and Q_1, Q_2 denote the pdf of two continuous probability distributions $P = P_1P_2$ and $Q = Q_1Q_2$.

Expanding the logarithm in Eq. 5.10 we can write the Kullback–Leibler divergence between two probability distributions P and Q in terms of entropies as

$$D_{\mathrm{KL}}(P||Q) = -E_P(\log\ q(x)) + E_P(\log\ p(x)) = H^{\mathrm{cross}}(P,Q) - H(P),$$

where p and q denote the pdf or pmf of the distributions P and Q and $H(P,Q)^{\mathrm{cross}}$ is the so-called *cross-entropy* of P and Q given by

$$\mathrm{H}^{\mathrm{cross}}(P,Q) := -E_P(\log\ Q).$$

This relation lets us easily compute a closed form of the KL divergence for many common families of probability distributions. Let us, for example, look at the value of the KL divergence between two normal distributions $P : N(\mu_1,{\sigma_1}^2)$ and $Q : N(\mu_2,{\sigma_2}^2)$; see Fig. 5.14. This can be calculated as

$$D_{\mathrm{KL}}(P||Q) = \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2} + \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - \log\frac{\sigma_1^2}{\sigma_2^2} - 1\right).$$

Another example: The KL divergence between two exponential distributions $P : Exp(\lambda_1)$ and $Q : Exp(\lambda_2)$ is

$$D_{\mathrm{KL}}(P||Q) = \log(\lambda_1) - \log(\lambda_2) + \frac{\lambda_2}{\lambda_1} - 1.$$

Using the Kullback–Leibler divergence, we can give yet another characterization of mutual information: It is a measure of how far two measured variables are from being independent, this time in terms of the Kullback–Leibler divergence.

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= \underbrace{-\sum_{x} P(x)\log(P(x))}_{=-\sum_{x,y} P(x,y)log(P(x))} + \sum_{x,y} P(x,y)\log(P(x|y)) \\
&= \sum_{x,y} P(x,y)\log\left(\frac{P(x|y)}{P(x)}\right) \\
&= \sum_{x,y} P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right) \\
&= D_{\mathrm{KL}}(P(x,y)||P(x)P(y))
\end{aligned}
\tag{5.11}
$$

Thus, mutual information of two random variables can be seen as the KL divergence of their underlying joint probability distribution from the products of their marginal probability distributions, i.e., as a measure for how far the two variables are from being independent.

5.4.5 Transfer Entropy: Conditional Mutual Information

In the past, mutual information was often used as a measure of information transfer between units (modeled as random variables) in some system. This approach faces the problem that mutual information is a symmetric measure and does not have an inherent directionality. In some applications this symmetry is not desired though, namely, whenever we want to explicitly obtain information about the "direction of flow" of information, for example, to measure causality in an information-theoretic setting; see Sect. 6.5.

In order to make mutual information a directed measure, a variant called *time-lagged mutual information* was proposed, calculating mutual information for two variables including a previous state of the source variable and a next state of the destination variable (where discrete time is assumed).

Yet, as Schreiber [94] points out, while time-lagged mutual information provides a directed measure of information transfer, it does not allow for a time-dynamic aspect as it measures the statically shared information between the two elements. With a suitable conditioning on the part of the variables, the introduction of a time-dynamic aspect is possible though. The resulting quantity is commonly referred to as *transfer entropy* [94]. Its common definition is the following.

Definition 4.9 Transfer Entropy. *Let X and Y be discrete random variables given on a discrete-time scale, and let $k, l \geq 1$ be two natural numbers. Then the transfer entropy from Y to X with k memory steps in X and l memory steps in Y is defined as*

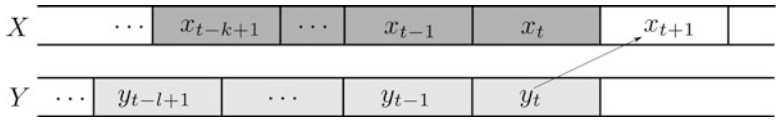

Fig. 5.15 Computing transfer entropy $TE_{Y \to X}$ from source Y to target X at time t as a measure of the average information present in y_t about the future state x_{t+1}. The memory vectors x_n^k and y_n^k are shown in *gray*

$$TE_{Y \to X} : \sum_{x_{n+1}, x_n^k, y_n^l} P\left(x_{n+1}, x_n^k, y_n^l\right) \log \frac{P\left(x_{n+1} | x_n^k, y_n^l\right)}{P\left(x_{n+1} | x_n^k\right)},$$

where we denoted by x_n, y_n the value of X and Y at time n and by y the past k values of X, counted from time n on $x_n^k := (x_n, x_{n-1}, \ldots, x_{n-k+1})$ and analogously $y_n^l := (y_n, y_{n-1}, \ldots, y_{n-l+1})$, Fig. 5.15.

Although this definition might look complicated at first, the idea behind it is quite simple. It is merely the Kullback–Leibler divergence between the two conditional probability distributions $P(x_{n+1} | x_n^k)$ and $P(x_{n+1} | x_n^k, y_n^l)$,

$$TE_{Y \to X} = D_{\mathrm{KL}}\left(P\left(x_{n+1} | x_n^k\right) || P\left(x_{n+1} | x_n^k, y_n^l\right)\right),$$

i.e., a measure of how far the two distributions are from fulfilling the generalized Markov property (see Sect. 3.8)

$$P\left(x_{n+1} | x_n^k\right) = P\left(x_{n+1} | x_n^k, y_n^l\right). \tag{5.12}$$

Note that for small values of transfer entropy, we can say that Y has little influence on X at time t, whereas we can say that information is transferred from Y to X at time t when the value is large. Yet, keep in mind that transfer entropy is just a measure of statistical correlation; see Sect. 6.5.

Another interpretation of transfer entropy is seeing it as a conditional mutual information $I(Y^{(l)}; X' | X^{(k)})$, measuring the average information the source Y constrains about the next state X' of the destination X that was not contained in the destination's past $X^{(k)}$ (see [62]) or alternatively as the average information provided by the source about the state transition in the destination; see [51, 62].

As so often before, the concept can be generalized to the continuous case [51], although the continuous setting introduces some subtleties that have to be addressed.

Concerning the memory parameters k and l of the source and the destination, although arbitrary choices are possible, the values chosen fundamentally influence the nature of the questions asked. In order to get correct measures for systems being far from Markovian (i.e., systems which states are not influenced by more than a certain fixed number of preceding system states), high values of k have to be used, and for non-Markovian systems, the case $k \to \infty$ has to be considered. On the other hand, commonly just one previous state of the source variable is considered in

applications, setting $l = 1$ [62], this being also due to the growing data intensity in k and l and the usually high computational cost of the method.

Note that akin to the case of mutual information, there exist point-wise versions of transfer entropy (also called *local transfer entropy*), as well as extensions to the multivariate case; see [62].

5.5 Estimation of Information-Theoretic Quantities

As we have seen in the preceding sections, one needs to know the full sample spaces and probability distributions of the random variables involved in order to precisely calculate information-theoretic quantities such as the entropy, mutual information, or transfer entropy. But obtaining this data is in most cases impossible in reality, as the spaces are usually high dimensional and sparsely sampled, rendering the direct methods for the calculation of such quantities impossible to carry out. A way around this problem is to come up with estimation techniques that estimate entropies and derived quantities such as mutual information from the data. Over the last decades a large body of research was published concerning the estimation of entropies and related quantities, leading to a whole zoo of estimation techniques, each class having its own advantages and drawbacks. So rather than a full overview, we will give a sketch of some central ideas here and give references to further literature. The reader is also pointed to the review articles [10, 79].

Before looking at estimation techniques for neural (and other) data, let us first give a swift and painless review of some important theoretical concepts regarding statistical estimation.

5.5.1 A Bit of Theory Regarding Estimations

From a statistical point of view, the process of estimation in its most general form can be regarded in the following setting: We have some data (say measurements or data obtained via simulations) that is believed to be generated by some stochastic process with an underlying non-autonomous, i.e., time dependent or autonomous probability distribution. We then want to estimate either the value of some function defined on that probability distribution (e.g., the entropy) or the shape of this probability distribution as a whole (from which we can then obtain an estimate of a derived quantity). This process is called *estimation* and a function mapping the data to an estimated quantities *estimator*. In this section we will first look at estimators and their desired properties and then look at what is called maximum likelihood estimation, the most commonly used method for the estimation of parameters in the field of statistics.

5.5.1.1 Estimators

Let $x = (x_1, \ldots, x_n)$ be a set of realizations of the random variable X that is believed to have a probability distribution that comes from a family of probability distributions P_θ parametrized by a parameter θ and assume that the underlying probability distribution of X is $P_{\theta_{\text{true}}}$.

Let $T : x \mapsto \hat{\theta}_{\text{true}}$ be an estimator for the parameter θ with the true value θ_{true}. For the value of the estimated parameter, we usually write $\hat{\theta}_{\text{true}} := T(x)$. The *bias* of $T(x)$ is the expected difference between $\hat{\theta}_{\text{true}}$ and θ_{true}:

$$bias(T) := E_X\left[\hat{\theta}_{\text{true}} - \theta_{\text{true}}\right],$$

and an estimator with vanishing bias is called *unbiased*.

One usually strives to obtain unbiased estimators that are also *consistent*, i.e., for which the estimated value $\hat{\theta}_{\text{true}}$ converges to the value of the true parameter θ_{true} in probability as the sample x increases in size, i.e., as $n \to \infty$:

$$\lim_{n\to\infty} P(|T(X) - \theta_{\text{true}}| > \varepsilon) = 0.$$

Another important property of an estimator is its variance $var(T)$, and an unbiased estimator having the minimal variance among all unbiased estimators of the same parameter is called *efficient*.

Yet another measure often used when assessing the quality of an estimator T is its mean squared error

$$MSE(T) = (bias(T))^2 + var(T),$$

and as we can see, any unbiased estimator with minimal variance minimizes the mean squared error.

Without further going into detail here, it is noted that there exists a theoretical lower bound to the minimal variance obtainable by an unbiased estimator, the *Cramér-Rao bound*. The Cramér-Rao bound sets the variance of the estimator in relation to the so-called Fisher information (that can be set into relation with mutual information, see [17, 113]). The interested reader is pointed to [2, 58].

5.5.1.2 Estimating Parameters: The Maximum Likelihood Estimator

Maximum likelihood estimation is the most widely used estimation technique in statistics and, as we will see in the next few paragraphs, a straightforward procedure that in essence tells us what the most likely parameter value in an assumed family of probability distributions is, given a set of realizations of a random variable that is believed to have an underlying probability distribution from the family considered.

In statistical applications one often faces the following situation: We have a finite set of realizations $\{x_i\}_i$ of a random variable X. We assume X to have a probability distribution $f(x,\theta_{\text{true}})$ in a certain parametrized class of probability distributions $\{f(x,\theta)\}_\theta$, where the true parameter θ_{true} is unknown. The goal is to get an estimate $\hat{\theta}_{\text{true}}$ of θ_{true} using the realizations $\{x_i\}_i$, i.e., to do statistical inference of the parameter θ. Let us consider the so-called likelihood function

$$L(\theta|x) = P_\theta(X = x) = f(x|\theta)$$

as a function of θ. It is a measure of how likely it is that the parameter of the probability distribution has the value θ, given the observed realization x of X. In maximum likelihood estimation, we look for the parameter that maximizes the likelihood function. This is $\hat{\theta}_{\text{true}}$:

$$\hat{\theta}_{\text{true}} = argmax_\theta L(\theta|x).$$

Choosing a value of $\theta = \hat{\theta}_{\text{true}}$ minimizes the KL divergence between P_θ and $P_{\theta_{\text{true}}}$ for all possible values of θ. The value $\hat{\theta}_{\text{true}}$, often written as $\hat{\theta}_{\text{MLE}}$, is called the *maximum likelihood estimate* (MLE for short) of θ_{true}.

In this setting, one often not uses the likelihood function directly, but works with the *log* of the likelihood function (this is referred to as log-likelihood). Why? The likelihood functions are often very complicated and situated in high dimensions, making it impossible to find a maximum of the function analytically. Thus, numerical methods (such as Newton's method and variants or the simplex method) have to be employed in order to find a solution. These numerical methods work best (and can be shown to converge to a unique solution) if the function they operate on is concave (bowl-shaped, where the closed end is on the top). The log function has the property to make the likelihood function concave in many cases, that being the reason why one considers the log-likelihood function, rather than the likelihood function directly; see also [80].

5.5.2 Regularization

Having looked at some core theoretical concepts regarding the estimation of quantities depending on probability distributions, let us now come back to dealing with real data.

As in real-world data, the involved probability distributions are often continuous and infinite-dimensional, the resulting estimation problem is very difficult (if not impossible) to solve in its original setting. As a remedy, the problem is often *regularized*, i.e., mapped to a discrete, more easily solvable problem. This of course introduces errors and often makes a direct estimation of the information-theoretic quantities impossible, but even in that simplified model we can estimate lower bounds of the quantities that we are interested in.

By using Shannon's *information-processing inequality* [26]

$$I(X;Y) \geq I(S(X);T(Y)),$$

where X and Y are (discrete) random variables and S and T are measurable maps and choosing the mappings S and T as our regularization mappings (you might also regard them as parameters) we can change the coarseness of the regularization. The regularization can be chosen arbitrarily coarse, i.e., choosing S and T as constant functions, but this of course comes with a price. For example, in the latter case of constant S and T, the mutual information $I(S(X); S(Y))$ would be equal to 0, clearly not a very useful estimate. This means that a trade-off between complexity reduction and the quality of the estimation has to be made. In general, there exists no all-purpose recipe for this, each problem requiring an appropriate regularization.

As this discretization technique has become the standard method in many fields, we will solely consider the regularized, discrete case in the following and point the reader to the review article [10] concerning the continuous case.

In the neurosciences, such a regularization technique was also proposed and is known as the "direct method" [19, 104]. Here, spike trains of recorded neurons are discretized into time bins of a given fixed width, and the neuronal spiking activity is interpreted as a series of symbols from an alphabet defined via the observed spiking pattern in the time bins.

5.5.3 Nonparametric Estimation Techniques

Commonly, two different classes of estimation techniques regarding the shape of probability distributions are distinguished. Parametric estimation techniques assume that the probability distribution is contained in some family of probability distributions having some prescribed shape (see Sect. 3.7). Here, one estimates the value of the parameter from the data observed, whereas nonparametric estimation techniques make no assumptions about the shape of the underlying distribution. We will solely look at nonparametric estimation techniques in the following as in many cases one tries to not assume prior information about the shape of the distribution.

Histogram-based estimation is the most popular and most widely used estimation technique. As the name implies, this method uses a histogram obtained from the data to estimate the probability distribution of the underlying random generation mechanism.

For the following, assume that we obtained a finite set of N samples $x = \{x_i\}_i$ of some real random variable X defined over some probability space (Ω,Σ,P). We then divide the domain of X into $m \in \mathrm{N}$ equally sized bins $\{b_i\}_i$ and subsequently count the number of realizations x_i in our data set contained in each bin. Here, the number m of bins can be freely chosen. It controls the coarseness of our discretization,

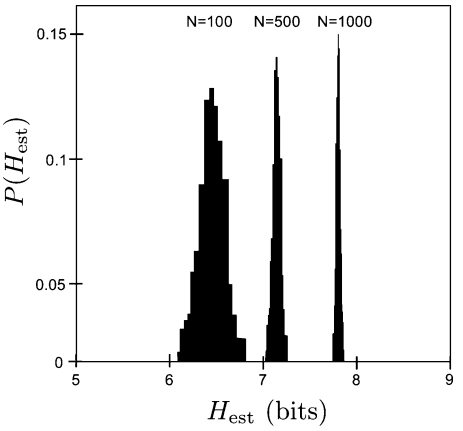

Fig. 5.16 Estimation bias for a non-bias-corrected histogram-based maximum likelihood estimator H_{est} of the entropy of a given distribution with true entropy $H = 8$ bits. Estimated values are shown for three different sample sizes N (Adapted from [79], Fig. 1)

where the limit $m \to \infty$ is the continuous case. This allows us to define relative frequencies of occurrences for X with respect to each bin that we interpret as estimations $\hat{p}_i^m$ (note that we make the dependence on the number of bins m explicit in the notation) of the probability of X taking a value in bin b_i which we denote by $p_i^m = P(X \in b_i)$. The law of large numbers then tells us that our estimated probability values converge to the real probabilities as $N \to \infty$.

Note that although histogram-based estimations are usually called nonparametric as they do not assume a certain shape of the underlying probability distribution, they do have parameters, namely, one parameter for each bin and the estimated probability value $\hat{p}_i^m$. These estimates $\hat{p}_i^m$ can also be interpreted as maximum likelihood estimates of p_i^m.

The following defines an estimator of the entropy based on the histogram. It is often called "plug-in" estimator:

$$\hat{H}_{\text{MLE}}(x) := -\sum_{i=1}^{m} \hat{p}_i^m \log p_i^m. \tag{5.13}$$

The are some problems with this estimator $\hat{H}_{\text{MLE}}(X)$, though. Its convergence to the true value $H(X)$ can be slow and it is negatively biased, i.e., its value is almost always below the true value $H(X)$; see [4, 79, 82, 83]. This shift can be quite significant even for large N; see Fig. 5.16 and [79]. More specifically, one can show that the expected value of the estimated entropy is always smaller than the true value of the entropy

$$E_X\left[\hat{H}_{\text{MLE}}(x)\right] \leq H(X),$$

where the expectation value is taken with respect to the true probability distribution P.

Bias generally is a problem for history-based estimation techniques [14, 82, 95], and although we can correct for the bias, this may not always be a feasible

solution [4]. Nonetheless we will have a look at a bias-corrected version of the estimator given in Eq. 5.13 below.

As a remedy to the bias problem, Miller and Madow [71] calculated the bias of the estimator of Eq. 5.13 and came up with a bias-corrected version of the maximum likelihood estimator for the entropy, referred to as *Miller-Madow* estimator:

$$\hat{H}_{\mathrm{MM}}(x) := \hat{H}_{\mathrm{MLE}}(x) + \frac{\hat{m} - 1}{2N},$$

where $\hat{m}$ is an estimate of the number of bins with nonzero probability. We will not go into the detail of the method here; the interested reader is referred to [71].

Another way of bias-correction $\hat{H}_{\mathrm{MLE}}(X)$ is the so-called "jack-knifed" version of the maximum likelihood estimator by Efron and Stein [34]:

$$\hat{H}_{\mathrm{JK}}(x) := N \cdot \hat{H}_{\mathrm{MLE}}(x) + \frac{N-1}{N} \sum_{j=1}^{N} \hat{H}_{\mathrm{MLE}}(x \backslash \{x_i\}),$$

Yet another bias-corrected variant of the MLE estimator based on polynomial approximation is presented in [79], for which also bounds on the maximal estimation error were derived.

In an effort to overcome the problems faced by histogram-based estimation, many new and more powerful estimation techniques have emerged over the last years, both for entropy and other information-theoretic quantities. As our focus here is to give an introduction to the field, we will not review all of those methods here but rather point the interested reader to the literature where a variety of approaches are discussed. There exist methods based on the idea of adaptive partitioning of sample space [21], ones using entropy production rates and allowing for confidence intervals [99], ones using Bayesian methods [75, 90, 99], and ones based on density estimation using nearest neighbors [55], along with many more. See [46] for an overview concerning several estimation techniques for entropy and mutual information. We note here that in contrast to estimations of entropy, estimators of mutual information are usually positively biased, i.e., tend to overestimate mutual information.

5.6 Information-Theoretic Analyses of Neural Systems

Some time after its discovery by Shannon, neuroscientists started to recognize information theory as a valuable mathematical tool to assess information processing in neural systems. Using information theory, several questions regarding information processing and the neural code can be addressed in a quantitative way, among those:

- How much information single cells or populations carry about a stimulus and how this information is coded.
- What aspects of a stimulus are encoded in the neural system.

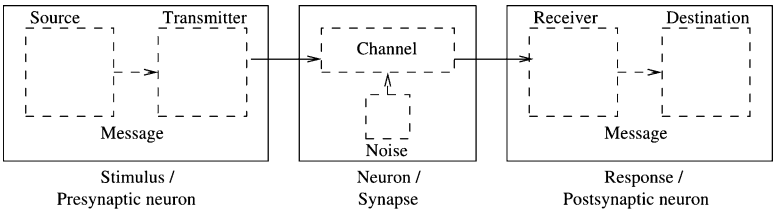

Fig. 5.17 An information-theoretic view on neural systems. Neurons can either act as channels in the information-theoretic sense, relaying information about some stimulus or as senders and receivers with channels being synapses

- How "effective connectivity" [40] in neural systems can be defined via causal relationships between units in the system.

See Fig. 5.17 for an illustration of how Shannon's theory can be used in a neural setting.

Attneave [6] and Barlow [9] were the first to consider information processing in neural systems from an information-theoretic point of view. Subsequently, Eckhorn and Pöpel [32, 33] applied information-theoretic methods to electrophysiologically recorded data of neurons in a cat. But being data intensive in nature, these methods faced some quite strong restrictions during that time, namely, the limited amount of computing power (and computer memory) and the limited amount (and often low quality) of data obtainable via measurements at that time.

But over the last decades, available computing became more and more available, and classical measurement techniques were improved, along with new ones emerging such as fMRI, MEG, and calcium imaging. This made information-theoretic analyses of neural systems more and more feasible, and through the invention of recording techniques such as MEG and fMRI, it is nowadays even possible to perform such analyses on a system scale for the human brain in vivo. Yet, even with the newly available recording techniques today, there are some conceptual difficulties with information-theoretic analyses as it is often a challenge to obtain enough data in order to get good estimates of information-theoretic quantities. Special attention has to be paid to using the data efficiently, and the validity of such analyses has to be assessed to their statistical significance.

In the following we will discuss some conceptual questions relevant when regarding information-theoretic analyses of neural systems. More detailed reviews can be found in [15, 35, 93, 111].

5.6.1 The Question of Coding

Marr described "three levels at which any machine carrying out an information-processing task must be understood" [68] [Chap. 1.2]. They are:

1. Computational theory: What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?

2. Representation and algorithm: How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?
3. Hardware implementation: How can the representation and algorithm be realized physically?

When performing an information-theoretic analysis of a system, one naturally faces the fundamental problem related to the coding of the information: In order to calculate (i.e., estimate) information-theoretic quantities, one has to define a family of probability distributions over the state space of the system, each member of that family describing one system state that is to be considered. As we know, all information-theoretic quantities such as entropy and mutual information (between the system state and the state of some external quantity) are determined by the probability distributions involved. The big question now is how to define the system state in the first point, a question which is especially difficult to answer in the case of neural systems on all scales.

One possible way to construct such a probabilistic model for a sensory neurophysiological experiment involving just one neuron is the following. Typically, the experiment consists of many trials, where per trial $i = 1, \ldots, n$ in some defined time window a stimulus S_i is presented eliciting a neural response $R(S_i)$ consisting of a sequence of action potentials. Presenting the same stimulus S many times allows for the definition of a probability distribution of responses $R(S)$ of the neuron to a stimulus S. This is modeled as a conditional probability distribution $P_{R|S}$. As noted earlier, we usually have no direct access to $P_{R|S}$ but rather have to find an estimate $\hat{P}_{R|S}$ from the available data. Note that in practice, usually the joint probability distribution $P(R,S)$ is estimated and estimates of conditional probability distributions are subsequently obtained from the estimate of the joint distribution.

Let us now assume that the stimuli are drawn from the set of stimuli $S = \{S_1, \ldots, S_k\}$ according to some probability distribution P_S (that can be freely chosen by the experimenter). We can then compute the mutual information between the stimulus ensemble S and its elicited response $R(S)$

$$I(S;R(S)) := H(R(S)|S) - H(S) = H(S|R(S)) - H(R(S))$$

using the probability distributions P_S and $\hat{P}_{R|S}$; see Sect. 4.3.

As usual, by mutual information we assess the expected shared information between the stimulus and its elicited response averaged over all stimuli and responses. In order to break this down to the level of single stimuli, we can either consider the point-wise mutual information or employ one of the proposed decompositions of mutual information such as *stimulus-specific information* or *stimulus-specific surprise*; see [20] for a review.

Having sketched the general setting, let us come back to the question of coding of information by the neurons involved. This is important as we have to adjust our model of the neural responses accordingly, the goal being to capture all relevant features of the neural response in the model.

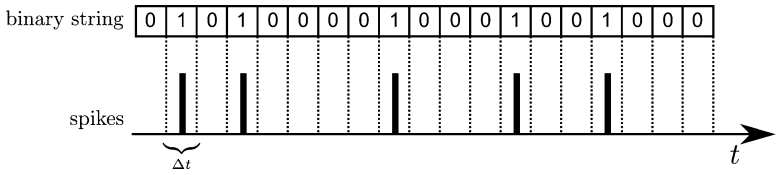

Fig. 5.18 Model of a spike train. The binary string is obtained through a binning of time

Regarding neural coding, there are two main hypotheses of how single neurons might code information: Neurons could use a *rate code*, i.e., encode the information via their mean firing rates, neglecting the timing patterns of spikes, or they could employ a *temporal code*, i.e., a code where the precise timing of single spikes plays an important role. Yet another hypothesis would be that neurons code information in bursts of spikes, i.e., groups of spikes emitted in a small time window, which is a variant of the time code. For questions regarding coding in populations, see the review [89].

Note that the question of neural coding is a highly debated one in the neurosciences as of today (see [42, 96]), and we do not want to favor one view point over the other in the following. As with many things in nature, there does not seem to be a clear black and white picture regarding neuronal coding. Rather it seems that a gradient of different coding schemes is employed depending on which sensory system is considered and at which stage of neuronal processing; see [19, 22, 42, 81, 93].

5.6.2 *Computing Entropies of Spike Trains*

Let us now compute the entropy of spike trains and subsequently single spikes, assuming that the neurons we model employ either a rate or a time code. We are especially interested in the maximal entropy attainable by our model spike trains as these can give us upper bounds for the amount of information such trains and even single spikes can carry in theory. The following examples here are adapted from [108]. Concerning the topics of spike trains and their analysis, the interested reader is also pointed to [92].

First, we define a model for the spike train emitted by a neuron measured for some fixed time interval of length T. We can consider two different models for the spike train, a continuous and a discrete one. In the continuous case, we model each spike by a Dirac delta function and the whole spike train as a combination of such functions. The discrete model is obtained from the continuous one by introducing small time bins of size Δt in a way that one bin can at most contain one spike, say $\Delta t = 2$ ms. We then assign to each bin in which no spike occurred a value of 0 and ones in which a spike occurred a value of 1; see Fig. 5.18.

Let us use this discrete model for the spike train of a neuron, representing a spike train as a binary string S in the following. Fixing the time span to be T and the bin

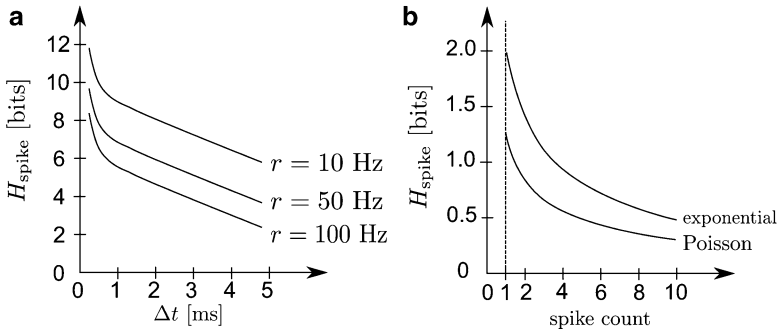

Fig. 5.19 Maximum entropy per spike for spike trains. (**a**) Time code with different rates r as a function of the size Δt of the time bins. (**b**) Rate code using Poisson and exponential spiking statistics (Figure adapted from [108] Fig. D.4)

width to be Δt, each spike train S has length $N = T/\Delta t$. We want to calculate the maximal entropy among all such spike trains n, subject to the condition that the number of spikes in S is a fixed number $r \leq N$ which we call the spike rate of S.

Let us now calculate the entropy in the firing pattern of a neuron of which we assume that spike timing carries important information, i.e., a neuron employing a time code. In order to keep the model simple, let us further assume that the spiking behavior is not restricted in any way, i.e., that all possible binary strings S are equiprobable. Then we can calculate the entropy of this uniform probability distribution P as

$$H(P) = \log \binom{N}{r}, \tag{5.14}$$

where $\binom{N}{r}$ denotes the binomial coefficient $\binom{N}{r} = \frac{N!(N-r)!}{r!}$, the number of all distinct binary strings of length N having exactly r nonzero entries. The entropy in Eq. 5.14 can be approximated by

$$H(P) \approx -\frac{N}{\ln 2}\left(\frac{N}{r}\ln\frac{N}{r} + \left(1-\frac{N}{r}\right)\ln\left(1-\frac{N}{r}\right)\right), \tag{5.15}$$

where ln denotes the natural logarithm to the base e. The expression in Eq. 5.15 is obtained by using the approximation formula

$$\log \binom{n}{k} \approx n\left(\frac{k}{n}\log\left(\frac{k}{n}\right) - \left(1-\frac{k}{n}\right)\log\left(1-\frac{k}{n}\right)\right)$$

which is valid for large n and k and in turn based on Stirling's approximation formula for $\ln n!$.

See Fig. 5.19a for the maximum entropy attainable by the time code as a function of bin size Δt for different firing rates r.

On the other hand, modeling a neuron that reacts to different stimuli with a graded response in its firing rate is usually done using a rate code. Assuming a rate code where the timing of spikes does not play any role yields different results, as we will see in the following; see Fig. 5.19b. In the rate code only the number of spikes N occurring in a given time interval of length T matters, i.e., we consider probability distributions $P_{N,T}$ parametrized by N and T describing how likely the occurrence of N spikes in a time window of length T is. Being well-backed with experimental data [24, 74, 101], a popular choice of $P_{N,T}$ is taking a Poisson distribution with some fixed mean $N = r \cdot T$, where r is thought of as the mean firing rate of the neuron.

The probability $P_{N,T}(n)$ of observing n spikes in an interval of length T now is given by the pmf of the Poisson distribution

$$P_{N,T}(n) = \frac{N^n e^{-N}}{n!}$$

and the entropy of $P_{N,T}$ computes as

$$H(P_{N,T}) = -\sum_n P_{N,T}(n) \log P_{N,T}(n).$$

Again using Stirling's formula this can be written as

$$H(P_{N,T}) \approx \frac{1}{2}(\log N - \log 2\pi). \tag{5.16}$$

Dividing the entropy $H(P_{N,T})$ by the number of spikes that occurred yields the entropy per spike. See Fig. 5.19b for a plot of the entropy per spike as a function of the number of observed spikes.

An interesting question is to ask for the maximal information (i.e., entropy) that spike trains can carry, assuming a rate code. Assuming continuous time and prescribing mean and variance of the firing rate, this leaves the exponential distribution $P_{\exp}$ as the one with the maximal entropy. The entropy of an exponentially distributed spike train with mean rate $r = 1/T(e^{\lambda} - 1)$ is

$$H(P_{\exp}) \approx \log(1+N) + N \log\left(1 + \frac{1}{N}\right),$$

see also Fig. 5.19b.

Note that while it was possible to compute the exact entropies in the preceding as we assumed full knowledge of the underlying probability distributions. This is of course not the case for data obtained by recordings. Here the estimation of entropies faces the bias-related problems of sparsely sampled probability distributions as discussed earlier. Concerning entropy estimation in spike trains, the reader is also pointed to [82].

5.6.3 *Efficient Coding?*

The principle of efficient coding [6, 9, 100] (also called *Infomax principle*) was first proposed by Attneave and Barlow. It views the early sensory pathway as a channel in Shannon's sense and postulates that early sensory systems try to maximize information transmission under the constraint of an efficient code, i.e., that neurons maximize mutual information between a stimulus and their output spike train, using as few spikes as possible. This minimization of spikes for a given stimulus results in a maximal compression of the stimulus data, minimizing redundancies between different neurons on a population level. One key prediction of this optimality principle is that neurons involved in the processing of stimulus data (and ultimately the whole brain) is adapted to natural stimuli, i.e., some form of natural (and structured) sensory input such as sounds or images rather than noise. For some sensory systems it could be shown that there is strong evidence that early stages of processing indeed perform an optimal coding; see, e.g., [77]. While first mainly the visual system was studied and it was shown that the Infomax principle holds here [9], other sensory modalities were also considered in the following years [13, 59–61, 109, 114].

But whereas the Infomax principle could explain certain experimental findings in the early sensory processing stages, the picture becomes less clear the more upstream the information processing in neural networks is considered. Here, other principles were also argued for; see, for example, [43].

On the system level, Friston et al. [36, 38] proposed an information-theoretic measure of free energy in the brain that can be understood as generalization of the concept of efficient coding. Also arguing for optimal information transfer, Norwich [76] gave a theory of perception based on information-theoretic principles. He argues that the information present in some stimulus is relayed to the brain by the sensory system with negligible loss. Many empirical equations of psychophysics can be derived from this model.

5.6.4 *Scales*

There are many scales at which information-theoretic analyses of neural systems can be performed. From the level of a single synapse [30, 65] over the level of single neurons [29, 93] over the population level [27, 35, 50, 87, 89] up to the system level [78, 110]. In the former cases the analyses are usually carried out on electrophysiologically recorded data of single cells, whereas on the system level data is usually obtained by EEG, fMRI, or MEG measurements.

Notice that most of the information-theoretic analyses of neural systems were done for early stages of sensory systems, focusing on the assessment of the amount of mutual information between some stimulus and its neural response. Here different questions can be answered about the nature and efficiency of the

neural code and the information conveyed by neural representations of stimuli; see [12, 15, 91, 93]. This stimulus–response-based approach has already provided a lot of insight into the processing of information in early sensory systems, but things get more and more complicated the more downstream an analysis is performed [22, 93].

On the systems level, the abilities of neural systems to process and store information are due to interactions of neurons, populations of neurons, and sub-networks. As these interactions are highly nonlinear and in contrast to the early sensory systems neural activity is mainly driven by the internal network dynamics (see [5, 110]), stimulus–response-type models often are not very useful here. Here, transfer entropy has proven to be a valuable tool, making analyses of information transfer in the human brain in vivo possible [78, 110]. Transfer entropy can also be used as a measure for causality, as we will discuss in the next section.

5.6.5 Causality in the Neurosciences

The idea of *causality*, namely, the question of what are the causes resulting in the observable state and dynamics of complex systems of physical, biological, or social nature is a deep, philosophical question that has been driving scientists in all fields ever since. In a sense this question lies at the heart of science itself and as such is often notoriously difficult to answer.

In the neurosciences, this principle is related to one of the core questions of neural coding and subsequently neural information processing: What stimuli make neurons spike (or change their membrane potential for non-spiking neurons)? For many years now, neuroscientists have investigated neurophysiological correlates of information presented to a sensory system in form of stimuli.

While considerable progress has been made regarding the answer to this question in the early stages of sensory processing (see the preceding sections), where often a clear correlation between a stimulus and the resulting neuronal activity could be found, things get less and less clear the further downstream this question is addressed. In the latter case, neuronal activity is subject to higher and higher degrees of internal dynamics and a clear stimulus–response relation is often lacking.

Considering early sensory systems, even though merely a correlation between a stimulus and neural activity can be measured, it is justified to speak of causality here, as it is possible to actively influence the stimulus and observe the change in neural activity. Note that the idea of intervention is crucial here; see [7, 85].

Looking at more downstream systems or at the cognitive level, an active intervention albeit possible (but often not as directly as for sensory systems) may not have the same easy to detect effects on system dynamics. Here, often just statistical correlations can be observed, and in most cases, it is very hard if not impossible to show that the principle of causality in its purest form holds. Yet, one

can still make some statements regarding what one might call "statistical causality" in this case, as we will see.

In an attempt to give a statistical characterization of the notion of causality, the mathematician Wiener [112] came up with the following probabilistic framing of this concept that came to be known as *Wiener causality*: Consider two stochastic processes $X = (X_t)_{t \in \mathrm{N}}$ and $Y = (Y_t)_{t \in \mathrm{N}}$. Then Y is said to Wiener-cause X if the knowledge of past values of Y diminishes uncertainty about the future values of X. Note that Wiener causality is a measure of predictive information transfer and not one of causality, and thus the naming is a bit unfortunate; see [63].

The economist Granger employed Wiener's principle of causality and developed the notion of what is nowadays called *Wiener-Granger causality* [16, 44]. Subsequently, the linear Wiener-Granger causality and its generalizaticons were often employed as measure of statistical causality in the neurosciences; see [16, 46]. Another model for causality in the neurosciences is *dynamic causal modeling* [37, 41, 102].

In contrast to dynamic causal modeling, causality measures based on information-theoretic concepts are usually purely data-driven and thus inherently model-free [46, 110]. This fact can be of advantage in some cases but we do not want to make a judgment here, calling one method better per se, as each has its advantages and drawbacks [39].

The directional and time-dynamic nature of transfer entropy allows using it as a measure of Wiener causality, as was proposed in the field of neurosciences recently [110]. As such, transfer entropy can be seen as a nonlinear extension of the concept of Wiener-Granger causality; see [66] for a comparison of transfer entropy to other measures.

Note again that transfer entropy still essentially is a measure of conditional correlation rather than one of direct effect (i.e., causality) and that correlation is not causation. Thus it is a philosophical question to which extent transfer entropy can be used to infer some form of causality, a question that we will not further pursue here, rather pointing the reader to [7, 46, 66, 85].

In any case the statistical significance of the inferred causality (remember that transfer entropy just measures conditional correlation) has to be verified. For trial-based data sets as often found in the neurosciences, this testing is usually done against the null hypothesis H_0 of average transfer entropy obtained by random shuffling of the data.

5.6.6 Information-Theoretic Aspects of Neural Dysfunction

Given the fact that information-theoretic analyses can provide insights about the functioning of neural systems, the next logical step is to ask how these might help in better understanding neural dysfunction and neural diseases.

The field one might call "computational neuroscience of disease" is an emerging field of research within the neurosciences; see the special issue of

Neural Networks [28]. The discipline faces some hard questions as in many cases dysfunction is observed on the cognitive (i.e., systems) level but has causes on many scales of neural function (subcellular, cellular, population, system).

Over the last years, different theoretical models regarding neural dysfunction and disease were proposed, among them computational models applicable to the field of psychiatry [48, 72], models for brain lesions [1], models of epilepsy [3], models for deep brain stimulation [70, 86], and models for aspects of Parkinson's [45, 73] and Alzheimer's [11, 56] disease, of abnormal auditory processing [31, 56], and for congenital prosopagnosia (a deficit in face identification) [103].

Some of these models employ information-theoretic ideas in order to assess differences between the healthy and dysfunctional states [8, 103]. For example, information-theoretic analyses of cognitive and systems-level processes in the prefrontal cortex were carried out recently [8, 53], and differences in information processing could be assessed between the healthy and dysfunctional system by means of information theory [8].

Yet, computational neuroscience of disease is a very young field of research, and it remains to be elucidated if and in what way analyses of neural systems employing information-theoretic principles could be of help in medicine on a broader scale.

5.7 Software

There exist several open source software packages that can be used to estimate information-theoretic quantities of neural data. The list below is by no means complete, but should give a good overview of things; see also [49]:

- Entropy: Entropy and mutual information estimation
  - URL: http://cran.r-project.org/web/packages/entropy.
  - Authors: Jean Hausser and Korbinian Strimmer.
  - Type: R package.
  - From the website: This package implements various estimators of entropy, such as the shrinkage estimator by Hausser and Strimmer, the maximum likelihood and the Millow-Madow estimator, various Bayesian estimators, and the Chao-Shen estimator. It also offers an R interface to the NSB estimator. Furthermore, it provides functions for estimating mutual information.
- Information-dynamics tool kit
  - URL: http://code.google.com/p/information-dynamics-toolkit.
  - Author: Joseph Lizier.
  - Type: standalone Java software.
  - From the website: Provides a Java implementation of information-theoretic measures of distributed computation in complex systems: i.e., information

storage, transfer, and modification. Includes implementations for both discrete and continuous-valued variables for entropy, entropy rate, mutual information, conditional mutual information, transfer entropy, conditional/complete transfer entropy, active information storage, excess entropy/predictive information, and separable information.

- ITE (information-theoretical estimators)
  - URL: https://bitbucket.org/szzoli/ite/.
  - Author: Zoltan Szabo.
  - Type: Matlab/Octave plug-in.
  - From the website: ITE is capable of estimating many different variants of entropy, mutual information, and divergence measures. Thanks to its highly modular design, ITE supports additionally the combinations of the estimation techniques, the easy construction and embedding of novel information-theoretical estimators, and their immediate application in information-theoretical optimization problems. ITE can estimate Shannon and Rényi entropy; generalized variance, kernel canonical correlation analysis, kernel generalized variance, Hilbert-Schmidt independence criterion, mutual information (Shannon, L2, Rényi, Tsallis), copula-based kernel dependency, and multivariate version of Hoeffding's Phi; complex variants of entropy and mutual information; and divergence (L2, Rényi, Tsallis), maximum mean discrepancy, and J-distance. ITE offers solution methods for Independent Subspace Analysis (ISA) and its extensions to different linear-, controlled-, post nonlinear-, complex-valued, partially observed systems, as well as to systems with nonparametric source dynamics.
- PyEntropy
  - URL: http://code.google.com/p/pyentropy.
  - Authors: Robin Ince, Rasmus Petersen, Daniel Swan, and Stefano Panzeri.
  - Type: Python module.
  - From the website: PyEntropy is a Python module for estimating entropy and information-theoretic quantities using a range of bias-correction methods.
- Spike train analysis tool kit
  - URL: http://neuroanalysis.org/toolkit.
  - Authors: Michael Repucci, David Goldberg, Jonathan Victor, and Daniel Gardner.
  - Type: Matlab/Octave plug-in.
  - From the website: Information-theoretic methods are now widely used for the analysis of spike train data. However, developing robust implementations of these methods can be tedious and time-consuming. In order to facilitate further adoption of these methods, we have developed the Spike Train Analysis Toolkit, a software package which implements several information-theoretic spike train analysis techniques.

- TRENTOOL
  - URL: http://trentool.de.
  - Authors: Michael Lindner, Raul Vicente, Michael Wibral, Nicu Pampu, and Patricia Wollstadt.
  - Type: Matlab plug-in.
  - From the website: TRENTOOL uses the data format of the open source MATLAB toolbox Fieldtrip that is popular for electrophysiology data (EEG/MEG/LFP). Parameters for delay embedding are automatically obtained from the data. TE values are estimated by the Kraskov-Stögbauer-Grassberger estimator and subjected to a statistical test against suitable surrogate data. Experimental effects can then be tested on a second level. Results can be plotted using Fieldtrip layout formats.

Acknowledgements The author would like to thank *Nihat Ay*, *Yuri Campbell*, *Aleena Garner*, *Jörg Lehnert*, *Timm Lochmann*, *Wiktor Młynarski,* and *Carolin Stier* for their useful comments on the manuscript.

References

1. J. Alstott, M. Breakspear, P. Hagmann, L. Cammoun, and O. Sporns. Modeling the impact of lesions in the human brain. *PLoS computational biology*, 5(6):e1000408, June 2009.
2. S-I. Amari, H. Nagaoka, and D. Harada. *Methods of information geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 2000.
3. I. S. And and K. Staley, editors. *Computational Neuroscience in Epilepsy*. Academic Press, 2011.
4. A. Antós and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures and Algorithms*, 19(3–4):163–193, 2001.
5. M. M. Arnold, J. Szczepanski, N. Montejo, J. M. Amigó, E. Wajnryb, and M. V. Sanchez-Vives. Information content in cortical spike trains during brain state transitions. *J Sleep Res*, 22(1):13–21, 2013.
6. F. Attneave. Some informational aspects of visual perception. *Psychol Rev*, 61(3):183–193, 1954.
7. N. Ay and D. Polani. Information Flows in Causal Networks. *Advances in Complex Systems*, 11(01):17–41, 2008.
8. F. Barcelo and R. T. Knight. An information-theoretical approach to contextual processing in the human brain: evidence from prefrontal lesions. *Cerebral cortex*, 17 Suppl 1:51–60, 2007.
9. H. B. Barlow. *Sensory Communication*, chapter Possible principles underlying the transformation of sensory messages, pages 217–234. MIT Press, 1961.
10. J. Beirlant and E. J. Dudewicz. Nonparametric entropy estimation: An overview. *Intern J Math Stat Sci*, 6(1):1–14, 1997.
11. B. S. Bhattacharya, D. Coyle, and L. P. Maguire. A thalamo-cortico-thalamic neural mass model to study alpha rhythms in Alzheimer's disease. *Neural Networks*, 24(6):631–645, 2011.
12. W. Bialek, F. Rieke, R. de Ruyter van Steveninck, and D. Warland. Reading a neural code. *Science*, 252(5014):1854–1857, 1991.
13. W. Bialek, R. Scalettar, and A. Zee. Optimal performance of a feed-forward network at statistical discrimination tasks. *Journal of Statistical Physics*, 57(1–2):141–156, 1989.

14. C. R. Blyth. Note on Estimating Information Author. *The Annals of Mathematical Statistics*, 30(1):71–79, 1959.
15. A. Borst and F. E. Theunissen. Information theory and neural coding. *Nat Neurosci*, 2(11):947–957, 1999.
16. S. L. Bressler and A. K. Seth. Wiener-Granger causality: a well established methodology. *Neuroimage*, 58(2):323–329, 2011.
17. N. Brunel and J. P. Nadal. Mutual information, Fisher information, and population coding. *Neural Comput*, 10(7):1731–1757, 1998.
18. Z Brzeniak and T. J Zastawniak. *Basic Stochastic Processes: A Course Through Exercises*. Springer, 1999.
19. G. T. Buracas, A. M. Zador, M. R. DeWeese, and T. D. Albright. Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. *Neuron*, 20(5):959–969, 1998.
20. D. A. Butts. How much information is associated with a particular stimulus? *Network*, 14(2):177–187, 2003.
21. C. Cellucci, A. Albano, and P. Rapp. Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms. *Physical Rev E*, 71(6): 066208, 2005.
22. G. Chechik, M. J. Anderson, O. Bar-Yosef, E. D. Young, N. Tishby, and I. Nelken. Reduction of information redundancy in the ascending auditory pathway. *Neuron*, 51(3):359–368, 2006.
23. D. Colquhoun and B. Sakmann. Fast events in single-channel currents activated by acetylcholine and its analogues at the frog muscle end-plate. *The Journal of Physiology*, 369:501–557, 1985.
24. A. Compte, C. Constantinidis, J. Tegner, S. Raghavachari, M. V. Chafee, P. S. Goldman-Rakic, and X-J. Wang. Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task. *J Neurophysiol*, 90(5):3441–3454, 2003.
25. C. H. Coombs, R. M. Dawes, and A. Tversky. *Mathematical psychology: an elementary introduction*. Prentice-Hall, 1970.
26. T. M. Cover and J. A. Thomas. *Elements of Information Theory*, volume 2012. John Wiley & Sons, 1991.
27. M. Crumiller, B. Knight, Y. Yu, and E. Kaplan. Estimating the amount of information conveyed by a population of neurons. *Frontiers in Neurosci*, 5(July):90, 2011.
28. V. Cutsuridis, T. Heida, W. Duch, and K. Doya. Neurocomputational models of brain disorders. *Neural Networks*, 24(6):513–514, 2011.
29. R. de Ruyter van Steveninck and W. Bialek. Real-time performance of a movement-sensitive neuron in the blowfly visual system: coding and information transfer in short spike sequences. *Proc. R. Soc. Lond. B*, 234(1277):379–414, 1988.
30. R. de Ruyter van Steveninck and S. B. Laughlin. The rate of information transfer at graded-potential synapses. *Nature*, 379:642–645, 1996.
31. X. Du and B. H. Jansen. A neural network model of normal and abnormal auditory information processing. *Neural Networks*, 24(6):568–574, 2011.
32. R. Eckhorn and B. Pöpel. Rigorous and extended application of information theory to the afferent visual system of the cat. I. Basic concepts. *Kybernetik*, 16(4):191–200, 1974.
33. R. Eckhorn and B. Pöpel. Rigorous and extended application of information theory to the afferent visual system of the cat. II. Experimental results. *Biol Cybern*, 17(1):71–77, 1975.
34. B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3):586–596, 1981.
35. A. Fairhall, E. Shea-Brown, and A. Barreiro. Information theoretic approaches to understanding circuit function. *Curr Opin Neurobiol*, 22(4):653–659, 2012.
36. K. Friston. The free-energy principle: a unified brain theory? *Nat Rev Neurosci*, 11(2):127–138, 2010.
37. K. Friston. Dynamic causal modeling and Granger causality Comments on: the identification of interacting networks in the brain using fMRI: model selection, causality and deconvolution. *Neuroimage*, 58(2):303–310, 2011.

38. K. Friston, J. Kilner, and L. Harrison. A free energy principle for the brain. *J Physiol Paris*, 100(1–3):70–87, 2006.
39. K. Friston, R. Moran, and A. K. Seth. Analysing connectivity with Granger causality and dynamic causal modelling. *Current opinion in neurobiology*, pages 1–7, December 2012.
40. K. J Friston. Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, 2(1–2):56–78, October 1994.
41. K. J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.
42. W. Gerstner, A. K. Kreiter, H. Markram, and A. V. Herz. Neural codes: firing rates and beyond. *Proc Natl Acad Sci U S A*, 94(24):12740–12741, 1997.
43. A. Globerson, E. Stark, D. C. Anthony, R. Nicola, B. G. Davis, E. Vaadia, and N. Tishby. The minimum information principle and its application to neural code analysis. *Proc Natl Acad Sci U S A*, 106(9):3490–3495, 2009.
44. C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37:424–438, 1969.
45. M. Haeri, Y. Sarbaz, and S. Gharibzadeh. Modeling the Parkinson's tremor and its treatments. *J Theor Biol*, 236(3):311–322, 2005.
46. K. Hlavackovaschindler, M. Palus, M. Vejmelka, and J. Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46, 2007.
47. P. G. Hoel, S. C. Port, and C. J. Stone. *Introduction to probability theory*. Houghton Mifflin Co., Boston, Mass., 1971.
48. Q. J. M. Huys, M. Moutoussis, and J. Williams. Are computational models of any use to psychiatry? *Neural Networks*, 24(6):544–551, 2011.
49. R. A. A. Ince, A. Mazzoni, R. S. Petersen, and S. Panzeri. Open source tools for the information theoretic analysis of neural data. *Frontiers in Neurosci*, 4(1):62–70, 2010.
50. R. A. A. Ince, R. Senatore, E. Arabzadeh, F. Montani, M. E. Diamond, and S. Panzeri. Information-theoretic methods for studying population codes. *Neural Networks*, 23(6):713–727, 2010.
51. A. Kaiser and T. Schreiber. Information transfer in continuous processes. *Physica D*, 166(March):43–62, 2002.
52. A. Klenke. *Probability Theory*. Universitext. Springer London, London, 2008.
53. E. Koechlin and C. Summerfield. An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11(6):229–235, 2007.
54. A. Kolmogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer-Verlag, Berlin, 1973.
55. A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Rev E*, 69(6):066138, 2004.
56. S. Krishnamurti, L. Drake, and J. King. Neural network modeling of central auditory dysfunction in Alzheimer's disease. *Neural Networks*, 24(6):646–651, 2011.
57. S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
58. E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 1998.
59. R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.
60. R. Linsker. Perceptual neural organization: some approaches based on network models and information theory. *Annu Rev Neurosci*, 13:257–281, 1990.
61. R. Linsker. Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Comput*, 702(1):691–702, 1992.
62. J. T. Lizier. *The Local Information Dynamics of Distributed Computation in Complex Systems*. Number October. Springer, springer edition, 2013.
63. J. T. Lizier and M. Prokopenko. Differentiating information transfer and causal effect. *The European Physical Journal B*, 73(4):605–615, January 2010.
64. J.T. Lizier, M. Prokopenko, and A.Y. Zomaya. The information dynamics of phase transitions in random Boolean networks. In *Proc Eleventh Intern Conf on the Simulation and Synthesis of Living Systems (ALife XI)*, pages 374–381. MIT Press, 2008.

65. M. London, A. Schreibman, M. Häusser, M. E. Larkum, and I. Segev. The information efficacy of a synapse. *Nat Neurosci*, 5(4):332–340, 2002.
66. M. Lungarella, K. Ishiguro, Y. Kuniyoshi, and N. Otsu. Methods for Quantifying the Causal Structure of Bivariate Time Series. *International Journal of Bifurcation and Chaos*, 17(03):903–921, 2007.
67. D. J. C. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.
68. David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press, 1982.
69. R. Marschinski and H. Kantz. Analysing the information flow between financial time series. *The European Physical Journal B*, 30(2):275–281, 2002.
70. C. C. McIntyre, S. Miocinovic, and C. R. Butson. Computational analysis of deep brain stimulation. *Expert Rev Med Devices*, 4(5):615–622, 2007.
71. G. A. Miller. *Information Theory in Psychology: Problems and Methods*, chapter Note on the bias of information estimates, pages 95–100. Free Press, 1955.
72. P. R. Montague, R. J. Dolan, K. J. Friston, and P. Dayan. Computational psychiatry. *Trends in Cognitive Sciences*, 16(1):72–80, 2012.
73. A. A. Moustafa and M. A. Gluck. Computational cognitive models of prefrontal-striatal-hippocampal interactions in Parkinson's disease and schizophrenia. *Neural Networks*, 24(6):575–591, 2011.
74. M. P. Nawrot, C. Boucsein, V. Rodriguez Molina, A. Riehle, A. Aertsen, and S. Rotter. Measurement of variability dynamics in cortical spike trains. *J Neurosci Methods*, 169(2):374–390, 2008.
75. I. Nemenman, W. Bialek, and R. R. de Ruyter van Steveninck. Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Rev E*, 69(5):056111, 2004.
76. K. H. Norwich. *Information, Sensation, and Perception*. Academic Press, 1993.
77. B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by VI? *Vision Res*, 37(23):3311–3325, 1997.
78. D. Ostwald and A. P. Bagshaw. Information theoretic approaches to functional neuroimaging. *Magn Reson Imaging*, 29(10):1417–1428, 2011.
79. L. Paninski. Estimation of entropy and mutual information. *Neural Comput*, 15(6):1191–1254, 2003.
80. L. Paninski. Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4):243–262, November 2004.
81. S. Panzeri, R. S. Petersen, S. R. Schultz, M. Lebedev, and M. E. Diamond. The role of spike timing in the coding of stimulus location in rat somatosensory cortex. *Neuron*, 29(3):769–777, 2001.
82. S. Panzeri, R. Senatore, M. A. Montemurro, and R. S. Petersen. Correcting for the sampling bias problem in spike train information measures. *Journal of neurophysiology*, 98(3):1064–72, 2007.
83. S. Panzeri and A. Treves. Analytical estimates of limited sampling biases in different information measures. *Network*, 7:87–107, 1995.
84. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Kaufmann, M, 1988.
85. J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.
86. M. Pirini, L. Rocchi, M. Sensi, and L. Chiari. A computational modelling approach to investigate different targets in deep brain stimulation for Parkinson's disease. *J Comput Neurosci*, 26(1):91–107, 2009.
87. A. Pouget, P. Dayan, and R. Zemel. Information processing with population codes. *Nat Rev Neurosci*, 1(2):125–132, 2000.
88. M. Prokopenko, F. Boschetti, and A. J. Ryan. An information-theoretic primer on complexity, self-organization, and emergence. *Complexity*, 15(1):11–28, 2009.

89. R. Q. Quiroga and S. Panzeri. Extracting information from neuronal populations: information theory and decoding approaches. *Nat Rev Neurosci*, 10(3):173–195, 2009.
90. K. R. Rad and L. Paninski. Information Rates and Optimal Decoding in Large Neural Populations. In *NIPS 2011: Granada, Spain*, pages 1–9, 2011.
91. F. Rieke, D. Warland, and W. Bialek. Coding efficiency and information rates in sensory neurons. *EPL (Europhysics Letters)*, 22(2):151–156, 1993.
92. F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek. *Spikes: Exploring the Neural Code (Computational Neuroscience)*. A Bradford Book, 1999.
93. E. T. Rolls and A. Treves. The neuronal encoding of information in the brain. *Prog Neurobiol*, 95(3):448–490, 2011.
94. T. Schreiber. Measuring Information Transfer. *Phys Rev Lett*, 85(2):461–464, 2000.
95. T. Schürmann. Bias analysis in entropy estimation. *Journal of Physics A: Mathematical and General*, 37(27):L295–L301, 2004.
96. T. J. Sejnowski. Time for a new neural code? *Nature*, 376(July):21–22, 1995.
97. C. E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(July, October 1948):379–423, 623–656, 1948.
98. A. N. Shiryayev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1984.
99. J. Shlens, M. B. Kennel, H. D. I. Abarbanel, and E. J. Chichilnisky. Estimating information rates with confidence intervals in neural spike trains. *Neural Comput*, 19(7):1683–1719, 2007.
100. E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annu Rev Neurosci*, 24:1193–1216, 2001.
101. W. R. Softky and C. Koch. The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *The Journal of Neuroscience*, 13(1):334–350, 1993.
102. K. E. Stephan, L. M. Harrison, S. J. Kiebel, O. David, W. D. Penny, and K. J. Friston. Dynamic causal models of neural system dynamics:current state and future extensions. *Journal of biosciences*, 32(1):129–144, 2007.
103. R. Stollhoff, I. Kennerknecht, T. Elze, and J. Jost. A computational model of dysfunctional facial encoding in congenital prosopagnosia. *Neural Networks*, 24(6):652–664, 2011.
104. S. Strong, R. Koberle, R. de Ruyter van Steveninck, and W. Bialek. Entropy and Information in Neural Spike Trains. *Phys Rev Lett*, 80(1):197–200, 1998.
105. H. Theil. *Henri Theil's Contributions to Economics and Econometrics: Econometric Theory and Methodology*. Springer, 1992.
106. I. Todhunter. *A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace*. Elibron Classics, 1865.
107. G. Tononi, O. Sporns, and G. M. Edelman. A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proc Natl Acad Sci U S A*, 91(11):5033–5037, 1994.
108. T. Trappenberg. *Fundamentals of Computational Neuroscience*. Oxford University Press, 2010.
109. J. H. van Hateren. A theory of maximizing sensory information. *Biol Cybern*, 29:23–29, 1992.
110. R. Vicente, M. Wibral, M. Lindner, and G. Pipa. Transfer entropy–a model-free measure of effective connectivity for the neurosciences. *J Comput Neurosci*, 30(1):45–67, 2011.
111. J. D. Victor. Approaches to information-theoretic analysis of neural activity. *Biological theory*, 1(3):302–316, 2006.
112. N Wiener. The theory of prediction. In E. Beckenbach, editor, *Modern mathematics for engineers*. McGraw-Hill, New-York, 1956.
113. S. Yarrow, E. Challis, and P. Seriès. Fisher and shannon information in finite neural populations. *Neural Comput*, 1780:1740–1780, 2012.
114. L. Zhaoping. Theoretical understanding of the early visual processes by data compression and data selection. *Network*, 17(4):301–334, 2006.
115. I. Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273, 2008.