Chapter 2
Improbability and Novelty of Descriptions

In this chapter we define the information of an event $A \in \Sigma$, or in our terminology the *novelty* of a *proposition* A as $-\log_2 p(A)$. We further define the important new concept of a *description* and extend the definition of *novelty* from events to *descriptions*. Finally we introduce the notions of *completeness* and *directedness of descriptions* and thereby the distinction between *surprise* and *information*, which are opposite special cases of *novelty*. This deviation from classical information theory is further elaborated in the fourth part of this book. The interested expert may go directly from Chap. 3 to Part IV.

2.1 Introductory Examples

You meet somebody (Mr. Miller) and ask him about his children. During the conversation he tells you two facts: "I have two children" and "This is my son" (pointing to a person next to him). Given these two facts, what is the probability that he has a daughter?

The task seems to be the evaluation of a conditional probability concerning the children of Mr. Miller. If we take his two statements at face value, this probability is $p[\,\text{he has a daughter} \mid \text{he has a son and he has two children}\,]$.

However, the situation is not so simple. For example, Mr. Miller's statement "I have two children" is also true, if he has three or more children. Now, if he has two or more children, at least one of them a son, and for each child the probability of being a daughter is about $\frac{1}{2}$, then the probability asked for is certainly at least $\frac{1}{2}$, maybe larger, depending on the number of children he actually has. But there is another, more plausible way of interpreting the two statements: You consider the situation of this conversation and what else Mr. Miller could have said. For example, if you asked him "Do you have two children?" and he simply answered "yes," then it may be that he actually has more than two children. However, if he could equally well answer your question with "I have three children" instead of "I have two children" you would expect that he chooses the first statement if he actually has

G. Palm, *Novelty, Information and Surprise*, DOI 10.1007/978-3-642-29075-6_2,

three children, since this describes his situation more accurately. This is what we normally expect in a conversation.

For example, it may have started by Mr. Miller pointing out "This is my son." Then you may have asked "Do you have more children?" and he may have answered "Yes, I have two." In this case, one statement would actually mean [he has two children and no more]. Let us turn to the other statement "this is my son." We would assume that Mr. Miller happened to be accompanied by one of his children. Indeed, if he would be accompanied by two of them, we would expect him to mention both of them in a usual conversation. So the other statement means [Mr. Miller was accompanied by one of his two children and this was his son]. Now we can work out the desired probability if we assume that it is equally probable that Mr. Miller is accompanied by one or the other of his two children, if he is accompanied by just one of them (which seems quite reasonable).

Mathematically, the situation can be described by three random variables X_1 and $X_2 \in \{m, f\}$ for child one and two, and $C \in \{0, 1, 2\}$ for his companion: $C = 0$ means he is not accompanied by just one child, $C = 1$, he is accompanied by child one, $C = 2$ for child two. We assume that

$$p(X_1 = m) = p(X_1 = f) = p(X_2 = m) = p(X_2 = f) = \tfrac{1}{2}$$

and $p[C = 1] = p[C = 2] = p$ and work out the answer to the problem.

The face value of the statement that he has a son (assuming he has exactly two children) is $A_m = [X_1 = m \text{ or } X_2 = m]$. With $A_f = [X_1 = f \text{ or } X_2 = f]$, the face-value probability that we asked for is

$$p(A_f|A_m) = \frac{p(A_f \cap A_m)}{p(A_m)} = \frac{\frac{1}{2}}{\frac{3}{4}} = \frac{2}{3}.$$

The other interpretation of the same statements is what is described more formally in this chapter. Mr. Miller describes the situation differently, depending on the values of X_1, X_2, and C:

- If $X_1 = m$ and $C = 1$ he says A_m,
- If $X_2 = m$ and $C = 2$ he says A_m,
- If $X_1 = f$ and $C = 1$ he says A_f,
- If $X_2 = f$ and $C = 2$ he says A_f,
- If $C = 0$ he says nothing (i.e., Ω) about the sex of his children.

Since he has said A_m, we have to ask for the set of all conditions under which he says A_m. This is called $\widetilde{A}_m$ in our theory:

$$\widetilde{A}_m = [X_1 = m, C = 1] \cup [X_2 = m, C = 2].$$

The desired probability is

$$\begin{aligned} p(A_f|\widetilde{A}_m) &= \frac{p(A_f \cap \widetilde{A}_m)}{p(\widetilde{A}_m)} \\ &= \frac{p[X_1 = m, X_2 = f, C = 1] + p[X_1 = f, X_2 = m, C = 2]}{p[X_1 = m, C = 1] + p[X_2 = m, C = 2]} \\ &= \frac{\frac{p}{4} + \frac{p}{4}}{\frac{p}{2} + \frac{p}{2}} = \frac{1}{2}. \end{aligned}$$

Another example for this distinction between A and $\widetilde{A}$ is the famous Monty Hall problem (Gardner 1959, 1969, see also Selvin 1975; Seymann 1991; Bapeswara-Rao and Rao 1992; Gillman 1992; Granberg and Brown 1995).

A quizmaster M gives the successful candidate C the opportunity to win a sportscar S. There are three doors and the sportscar is behind one of them (behind each of the other two doors is a goat). The candidate points at a door and if the sportscar is behind it it's his. Now the quizmaster opens one of the other doors and shows him a goat behind it (he knows where the sportscar is). Then he asks the candidate wether he wants to change his previous decision. Should the candidate change?

Again it is a problem of conditional probabilities. It can be described by three variables $S, C, M \in \{1, 2, 3\}$, describing the position of the sportscar, the initial choice of the candidate and the door opened by the quizmaster. There is a restriction on M, namely $S \neq M \neq C$. By opening one door (door i), the quizmaster effectively makes a statement $A_i = [$ the sportscar is not behind door $i]$ $(i = 1, 2, 3)$.

Here the face-value probability is $p[S = C|A_i]$. For reasons of symmetry, we may assume that all these probabilities are the same for $i = 1, 2, 3$. Thus we may assume $C = 1$ and $i = 2$. Then $p[S = 1|A_2, C = 1] = p[S = 1|S = 1 \text{ or } S = 3] = \frac{1}{2}$.

When, however, we ask for the conditions under which the quizmaster says A_i (i.e., for $\widetilde{A}_i$), the answer is $[M = i]$. Thus the desired probability is

$$p[S = C|M = i] = \frac{p[S = C, M = i]}{p[M = i]}.$$

Again for reasons of symmetry we may assume $C = 1$ and $i = 2$. Thus

$$\begin{aligned} p[S = 1|M = 2, C = 1] &= \frac{p[S = 1, C = 1, M = 2]}{p[M = 2, C = 1]} \\ &= \frac{p[S = 1, C = 1, M = 2]}{p[S = 1, C = 1, M = 2] + p[S = 3, C = 1, M = 2]} \\ &= \frac{\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2}}{\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{3}} = \frac{1}{3}. \end{aligned}$$

2.2 Definition and Properties

In this chapter we will define the *novelty* (on this level we might as well call it information or surprise) of a proposition, which should be a measure of its "unexpectedness."

The first idea is that a proposition certainly is the more surprising, the more improbable, i.e., the less probable it is.

A real function $f:\mathbb{R} \to \mathbb{R}$ is called *isotone* (or increasing), if $x \leq y$ implies $f(x) \leq f(y)$, and it is called *antitone* (or decreasing), if $x \leq y$ implies $f(x) \geq f(y)$.

Given a probability p on Ω and an antitone real function f, the function $f \circ p$ (defined by $(f \circ p)(A) = f(p(A))$ for $A \in \Sigma$) may be called an *improbability*.

Definition 2.1. For a proposition $A \in \Sigma$ we define the *novelty* $\mathcal{N}$ of A as

$$\mathcal{N}(A) := -\log_2 p(A).^{1}$$

We note that $\mathcal{N}$ is an improbability, since $x \mapsto -\log_2 x$ is antitone. But why did we choose $f = -\log_2$—why the base two? The basic idea is that $\mathcal{N}(A)$ should measure the number of *yes–no question*s needed to guess A. This will become much clearer in Chap. 4; here we just want to give a hint to make this choice of $f = -\log_2$ plausible.

Obviously, with one yes–no question we can decide between two possibilities, with 2 questions between 4 possibilities, with 3 questions between 8 possibilities, and so on, since we can use the first question to divide the 8 possibilities into 2 groups of 4 possibilities each, and decide which group it is, and then use the two remaining questions for the remaining 4 possibilities. In this way, with each additional question the number of possibilities that we can decide between is doubled. This means that with n questions we can decide between 2^n possibilities. If we want to find out the number of questions from the number of possibilities, we have to use the inverse relationship, i.e., for k possibilities we need $\log_2 k$ questions.

The most important property that is gained by the choice of a logarithmic function is the *additivity of novelty*: $\mathcal{N}(A \cap B) = \mathcal{N}(A) + \mathcal{N}(B)$ for independent propositions A and B. To explain this, we have to expand a little on the notion of *independence*.

Given two propositions A and B, we may try to find a statistical relation between the two. For example, we might ask whether it has an influence on the probability of A to occur, when we already know that B is true. The question is, whether the probability $p(A)$ is the same as the so-called *conditional probability* of A given B, which is defined as:

[1] This definition is the classical basic definition of information or entropy, which goes back to Boltzmann (1887) (see also Brush 1966).

$$p_B(A) = p(A|B) := \frac{p(A \cap B)}{p(B)}.$$

If $p(A|B) = p(A)$ then there is no such influence, and we call A and B *independent*.

Of course, we could also reverse the roles of A and B, and say that A and B are independent if $p(B|A) = p(B)$. It turns out that this condition is essentially equivalent to the other one, because

$$p(B|A) = \frac{p(A \cap B)}{p(A)} = \frac{p(A \cap B)}{p(B)} \cdot \frac{p(B)}{p(A)} = p(A|B) \cdot \frac{p(B)}{p(A)} \text{ equals } p(B)$$

if $p(A|B) = p(A)$.

Of course, all of this only makes sense if $p(A)$ and $p(B)$ are not zero. It is also clear that the two equivalent conditions are essentially the same as $p(A \cap B) = p(A) \cdot p(B)$, because $p(A|B) = p(A)$ also implies

$$p(A) \cdot p(B) = p(A|B) \cdot p(B) = p(A \cap B).$$

These considerations are summarized in the following definition:

Definition 2.2. Two propositions A and B are called *independent*, if

$$p(A \cap B) = p(A) \cdot p(B).$$

Proposition 2.1. *If we define the conditional novelty of A given B as*

$$\mathcal{N}(A|B) = -\log_2 p(A|B),$$

then we have

i) $\mathcal{N}(A \cap B) = \mathcal{N}(B) + \mathcal{N}(A|B)$
ii) $\mathcal{N}(A \cap B) = \mathcal{N}(A) + \mathcal{N}(B)$ *if A and B are independent.*

Proof. Obvious. □

2.3 Descriptions

Let us come back to an observation already made in Sect. 1.1. In general, the set Ω of all possible events may be very large. In a typical physical model the events $\omega \in \Omega$ would be represented as real vectors $\omega = (\omega_1, \ldots, \omega_n) \in \mathbb{R}^n$ and thus Ω would be larger than countable. On the other hand, Σ may well be countable, and so a description of an element $\omega \in \Omega$ by propositions $A \in \Sigma$ would be essentially inexact. Moreover, different persons may use different propositions to describe the

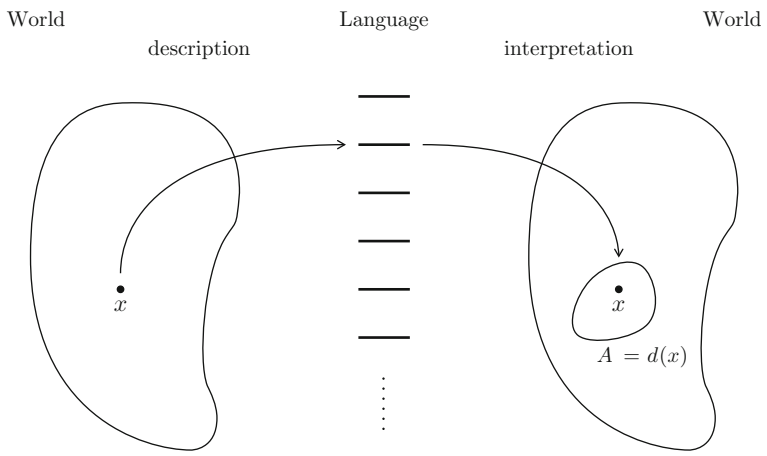

Fig. 2.1 Model for descriptions of events

same event ω. For example, when we walk on the street, we may see an event ω, let's say we see a car passing by. But this is not an exact description, and if we say that we see a blue Mercedes driven by an old man passing by very slowly, this still is not an exact description and somebody else might rather describe the same event as "Mr. Miller is driving with his wife's car into town."

What goes on here is that

1. Given the event, somebody describes it by a statement in a certain language.
2. Then this statement is interpreted again in our model Ω of the possible events as a proposition A, i.e., as the set A of all events y which are also described by the same statement (Fig. 2.1).

The point of view taken by a particular observer in describing the events $\omega \in \Omega$ by propositions $A \in \Sigma$, constitutes a particular description of these events. This process of *description* can be defined mathematically as a mapping.

Definition 2.3. Given (Ω, Σ, p), a mapping $d: \Omega \to \Sigma$ that assigns to each $\omega \in \Omega$ a proposition $d(\omega) = A \in \Sigma$ such that $\omega \in A$, is called a *description*.

In addition we require[2] that for every $A \in R(d)$

$$p[d = A] = p(\{\omega \in \Omega : d(\omega) = A\}) \neq 0.$$

This means that we don't want to bother with propositions that may happen only with probability 0. This additional requirement is quite helpful for technical reasons, but it is quite restrictive and it rules out some interesting examples (see Example 2.3). Note that the requirement that $\omega \in d(\omega)$ means that the description has to be true. So every event ω is described by a *true* proposition about it.

[2]This requirement obviously implies that the propositions $[d = A]$ are in Σ for every $A \in R(d)$. It also implies that $R(d)$ is finite or countable.

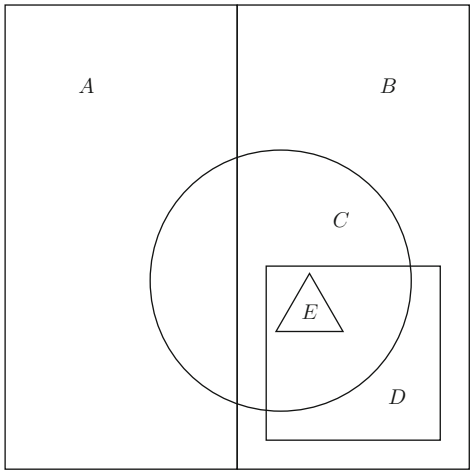

Fig. 2.2 Example of propositions about positions on a table

Example 2.1. Consider the throwing of a dice, i.e., the space of possible events $\Omega = \{1, 2, 3, 4, 5, 6\}$. Consider the following descriptions:

1) "Even vs. odd": For $A = \{2, 4, 6\}$ and $B = \{1, 3, 5\} = A^c$ we define the description e by $e\colon 1 \mapsto B, 2 \mapsto A, 3 \mapsto B, 4 \mapsto A, 5 \mapsto B, 6 \mapsto A$.
2) "Small vs. large":

$$d : 1 \mapsto \{1, 2, 3, 4\} = A,\ 2 \mapsto A,\ 3 \mapsto A,\\ 4 \mapsto \{3, 4, 5, 6\} = B,\ 5 \mapsto B,\ 6 \mapsto B.$$

3) "Pairs":

$$c : 1 \mapsto \{1, 2\},\ 2 \mapsto \{2, 3\},\ 3 \mapsto \{3, 4\},\\ 4 \mapsto \{3, 4\},\ 5 \mapsto \{4, 5\},\ 6 \mapsto \{5, 6\}.$$

□

Example 2.2. Try to define some descriptions that make use of the propositions indicated in Fig. 2.2 about locations on a square table including some logical operations on them.[3] □

Example 2.3. Without the requirement added to Definition 2.3 a description d may have an uncountable range $R(d)$. Here are two examples for this.

Take $\Omega = \mathbb{R}$ and a (continuous) probability p on $(\mathbb{R}, \mathcal{B})$ and $\delta > 0$. For $\omega \in \Omega$ we define $d(\omega) = (\omega - \delta, \omega + \delta)$ and $c(\omega) = [\omega, \infty) = \{x \in \mathbb{R}\colon x \geq \omega\}$.

Both c and d are interesting descriptions, but for every A in $R(c)$ or $R(d)$ we observe $p[c = A] = 0$ and $p[d = A] = 0$. □

Definition 2.4. A finite or countable collection $\{A_1, \ldots, A_n\}$ or $\{A_i\colon i \in \mathbb{N}\}$ of (measurable) subsets of Ω is called a (measurable) *partition*, if

[3] For example, one can describe points $x \in A \setminus C$ by $d(x) = A$, points x in $A \cap C$ by $d(x) = A \cap C$, points x in $B \setminus C$ by $d(x) = B$, and points x in $B \cap C$ by $d(x) = C$.

i) $\bigcup_i A_i = \Omega$ and
ii) $p(A_i \cap A_j) = 0$ for $i \neq j$.

For a partition we always have $\sum_i p(A_i) = 1$ and *essentially*[4] every $\omega \in \Omega$ is in exactly one of the sets A_i.

Usually part (ii) of the definition says $A_i \cap A_j = \emptyset$ for $i \neq j$. Here we again disregard sets of probability 0 as "essentially empty." In the following we will sometimes identify sets (A and B) that differ only by an essentially empty set, i.e., we may write $A = B$ (essentially), meaning that

$$p[A \neq B] = p(A \backslash B) + p(B \backslash A) = 0.$$

Example 2.4. Given a partition $\alpha = \{A_1, \ldots, A_n\}$ we may define $d_\alpha(\omega) := A_i$ for $\omega \in A_i$. Obviously, d_α is a description. □

2.4 Properties of Descriptions

Definition 2.5. Given a description $d: \Omega \to \Sigma$, we consider the *novelty mapping* $\mathcal{N}: \Sigma \to \mathbb{R}$ defined by $\mathcal{N}(A) = -\log_2 p(A)$ and call

$$N_d(\omega) = (\mathcal{N} \circ d)(\omega) = \mathcal{N}(d(\omega))$$

the *novelty provided by ω for the description d*. Note that $N_d: \Omega \mapsto \mathbb{R}$ is a random variable.[5] We further define the *average novelty* for the description d as

$$\mathcal{N}(d) := E(N_d).$$

For this definition it is required that $\mathcal{N} \circ d$ is measurable. Let us illustrate this in the case where d has finite range (see Definition 1.2). Let us say d takes the values $A_1, \ldots, A_n$, then $\mathcal{N} \circ d$ also has a finite range, and the value of $\mathcal{N} \circ d$ is $-\log_2 p(A_i)$, when the value of d is A_i.

This occurs on the set $\widetilde{A}_i := \{\omega \in \Omega: d(\omega) = A_i\} = d^{-1}(A_i)$. Now it is clear that $\mathcal{N} \circ d$ is a step function, namely

$$\mathcal{N} \circ d = \sum_{i=1}^{n} -\log_2 p(A_i) 1\!\!1_{\widetilde{A}_i}.$$

Clearly in this case we have to require that the sets $\widetilde{A}_i = d^{-1}(A_i)$ are in Σ for $i = 1, \ldots, n$, and then we can calculate

[4] From Definition (ii) it is obvious that the probability that ω is in two sets, e.g., A_i and A_j, is 0. So, *disregarding propositions with probability* 0, every $\omega \in \Omega$ is in exactly one of the sets A_i. We will usually disregard propositions with probability 0 and this is meant by the word "essentially".

[5] Due to the additional requirement in Definition 2.3 the function N_d is measurable. However, it may happen that $E(N_d)$ is infinite. For an example see Proposition 2.17.

$$\mathcal{N}(d) = E(\mathcal{N} \circ d) = -\sum_{i=1}^{n} p\left(\widetilde{A}_i\right) \log_2 p(A_i).$$

But these sets $\widetilde{A}_i$ do have still another significance. Let us ask what we can infer about the event ω that has taken place from its description $d(\omega) = A_i$ given by an observer. The obvious answer is of course that he tells us that ω is in A_i. But if we know the attitude of the observer well enough, we can infer even more. If we know his procedure of description, i.e., the mapping d he uses, we can infer that ω is such that $d(\omega) = A_i$, i.e., we can infer that $\omega \in [d = A_i] = \{y : d(y) = A_i\} = d^{-1}(A_i) =: \widetilde{A}_i$, and $\widetilde{A}_i$ is a more exact information about ω than A_i, because $\widetilde{A}_i \subseteq A_i$. (Indeed every ω in $\widetilde{A}_i$ satisfies $d(\omega) = A_i$).

Let us give an example of this kind of inference. If an observer says: "the car went slowly" this may be interpreted literally as "speed of the car below 30 km/h," say, but when we know that the observer would say that a car goes "very slowly," when its speed is below 10 km/h, we can even infer that the speed of the car was between 10 km/h and 30 km/h. Of course, an accurate observer might have said that the car goes slowly but not very slowly, but it is quite common experience that this additional information is neither mentioned nor used (nor of any interest).

If this kind of additional inference is already explicitly contained in a description d, then we call it *complete*. Given an arbitrary description d it is quite easy to construct another description $\tilde{d}$ that gives exactly this additional implicit information; this description $\tilde{d}$ we call the *completion* of d.

Definition 2.6. For a description d and for $A \in R(d)$ we define

i) $\widetilde{A} := [d = A] = \{\omega \in \Omega : d(\omega) = A\}$
ii) The description $\tilde{d}$ by $\tilde{d}(\omega) = \{\omega' \in \Omega : d(\omega) = d(\omega')\}$.

$\tilde{d}$ is called the *completion* of d. A description d is called *complete* if $d = \tilde{d}$.

Proposition 2.2. *The following properties of a description d are equivalent:*

i) *d is complete,*
ii) *If $d(\omega) \neq d(\omega')$ then $d(\omega) \cap d(\omega') = \emptyset$,*
iii) *The range $R(d)$ of d is a partition*[6] *of Ω.*

Proof. This should be obvious from the definitions. If not, the reader should try to understand the definition of $\tilde{d}$ and why the range of $\tilde{d}$ must be a partition. □

If we ask how *surprising* the outcome of a fixed description d will be, we again encounter the sets $\widetilde{A}_i$. The idea here is to consider the surprise of one particular outcome in comparison to all the other (usual) outcomes of the description d.

[6] Strictly speaking, here a partition should be defined by $A_i \cap A_j = \emptyset$ instead of $p(A_i \cap A_j) = 0$ (compare Definition 2.4). If we disregard 0-probability-propositions we should interpret $d = \tilde{d}$ in Definition 2.6 as $p[d = \tilde{d}] = 1$ and we should use the weaker formulation $p(d(\omega) \cap d(\omega')) = 0$ in part (ii) of this definition.

As a first step we can order the sets A_i according to their probability such that $p(A_1) \leq p(A_2) \leq \ldots \leq p(A_n)$. Then we can say that A_1 is more surprising than A_2 and so on. To quantify the amount of surprise we really get from A_i we determine the probability p_i that d gives us at least as much surprise as A_i does. This is $p_i = p[d(x) = A_1 \vee d(x) = A_2 \vee \ldots \vee d(x) = A_i]$. Since d can take only one value for every $\omega \in \Omega$, this is the sum of probabilities

$$p_i = \sum_{j=1}^{i} p[d = A_j] = \sum_{j=1}^{i} p(\widetilde{A}_j).$$

So for our description d, the surprise of A_i is $-\log p_i$, whereas its novelty is simply $-\log p(A_i)$, and its information is $-\log p(\widetilde{A}_i)$. Given a description d we can construct in the above fashion another description $\vec{d}$ which gives the surprise of d.

Definition 2.7. For a description d and for $A \in R(d)$, we define

$$\begin{aligned}\vec{A} &:= \bigcup \{\widetilde{B} \colon B \in R(d), p(B) \leq p(A)\} \\ &= [p \circ d \leq p(A)] \\ &= \{\omega \in \Omega \colon p(d(\omega)) \leq p(A)\}\end{aligned}$$

and the description $\vec{d}$ by

$$\vec{d}(\omega) := \{\omega' \colon p(d(\omega')) \leq p(d(\omega))\}.$$

A description d is called *directed*, if $d = \vec{d}$.

Definition 2.8. A description d is called *symmetric*, if for every $x, y \in \Omega$, $x \in d(y)$ implies $y \in d(x)$.

We can now reintroduce the set-theoretical operations that are defined on propositions, on the level of descriptions; in particular, we have the natural ordering and the union and intersection of descriptions:

Definition 2.9. We say that a description c is *finer* than a description d, or that d is *coarser* than c and write $c \subseteq d$, if $c(\omega) \subseteq d(\omega)$ for every $\omega \in \Omega$.

With this definition we see that the completion $\tilde{d}$ of any description d is always finer than d. This is because for any $\omega \in \Omega$, $x \in \tilde{d}(\omega)$ means $d(x) = d(\omega)$ and this implies $x \in d(x) = d(\omega)$. Obviously we also have $\tilde{d} \subseteq \vec{d}$, since $d(\omega') = d(\omega)$ implies $p(d(\omega')) \leq p(d(\omega))$.

Usually $d \subseteq \vec{d}$, but $\vec{d} \subseteq d$ is also possible.

Example 2.5. Let $\Omega = \{1, \dots, 6\}$.

$$d(\omega) := \{\omega\}, \qquad b(\omega) = \{1, \dots, \omega\}, \qquad and$$
$$c(1) = \{1, \dots, 4\}, \qquad c(2) = \{1, \dots, 5\}, \qquad c(i) = \Omega \text{ for } i > 2.$$

Then

$$\vec{d}(\omega) = \Omega, \qquad \vec{b}(\omega) = b(\omega), \qquad and$$
$$\vec{c}(1) = \{1\}, \qquad \vec{c}(2) = \{1, 2\}, \qquad \vec{c}(i) = \Omega \text{ for } i > 2.$$

Thus we have $\vec{b} = b$, $\vec{c} \subset c$, and $\vec{d} \supset d$. □

Definition 2.10. For any two descriptions c and d we define

i) $c \cap d$ by $(c \cap d)(\omega) := c(\omega) \cap d(\omega)$,
ii) $c \cup d$ by $(c \cup d)(\omega) := c(\omega) \cup d(\omega)$,
iii) d^c by $d^c(\omega) = (d(\omega))^c \cup \{\omega\}$ for every $\omega \in \Omega$.

Note that the complement or negation has to be slightly adjusted in order to keep the property that $\omega \in d^c(\omega)$. Still the complement has the nice property that $d \cap d^c$ is the finest possible description, namely $d \cap d^c(\omega) = \{\omega\}$, and $d \cup d^c$ is the coarsest possible description, namely $d \cup d^c(\omega) = \Omega$.

The point is that descriptions have to be true, i.e., $\omega \in d(\omega)$, and therefore the flat negation of a description cannot be a description. But given $d(\omega)$ one can try to describe the same event ω in the opposite or most different way, which is $d^c(\omega)$. It is like saying a glass of wine is half-empty instead of half-filled.

Proposition 2.3. *$c \subseteq d$ implies $\mathcal{N} \circ c \geq \mathcal{N} \circ d$, and this implies $\mathcal{N}(c) \geq \mathcal{N}(d)$.*

Proof. Obvious. □

This property of the novelty is called *monotonicity*. It is one of the essential properties that are needed for developing information theory.

Proposition 2.4. *Let c and d be two descriptions. Then $\tilde{c} \cap \tilde{d} \subseteq \widetilde{c \cap d}$.*

Proof. Let $\omega' \in \tilde{c}(\omega) \cap \tilde{d}(\omega)$. Then $c(\omega') = c(\omega)$ and $d(\omega') = d(\omega)$. Therefore $c \cap d(\omega') = c \cap d(\omega)$, i.e., $\omega' \in \widetilde{c \cap d}(\omega)$. □

Unfortunately $\tilde{c} \cap \tilde{d} \neq \widetilde{c \cap d}$ in general, as the following example shows.

Example 2.6. Let $\Omega = \{1, \dots, 6\}$ and define

$$c(1) = c(2) = \{1, 2\}, \qquad c(3) = c(4) = c(5) = c(6) = \{3, 4, 5, 6\}$$

and

$$d(1) = \{1,2,3,4,5\}, \qquad d(3) = \{2,3,4,5,6\},$$
$$d(2) = \{1,2,3,4\}, \qquad d(4) = \{1,3,4,5,6\},$$
$$d(5) = d(6) = \Omega.$$

We observe that $\tilde{d} \subset \tilde{c} = c \subset d$. Here $\tilde{d}(\omega) = \{\omega\}$ for $\omega = 1,2,3,4$ and $\tilde{d}(5) = \tilde{d}(6) = \{5,6\}$. Thus $\tilde{c} \cap \tilde{d} = \tilde{d}$ and $c \cap d = c$ implying $\widetilde{c \cap d} = \tilde{c} = c$. □

Also $c \subseteq d$ does not imply $\tilde{c} \subseteq \tilde{d}$, in general. The above is even an example where $c \subset d$ and $\tilde{c} \supset \tilde{d}$. These problems can be circumvented, if we consider so-called *consequential* or *tight* descriptions. The reason for the second name will become clear in Chap. 9.3.

Definition 2.11. A description d is called *consequential* or *tight*, if for every $x, y \in \Omega$

$$x \in d(y) \text{ implies } d(x) \subseteq d(y).$$

Proposition 2.5. *Let c, d be descriptions.*

i) *$\tilde{d}$ is tight and $\vec{d}$ is tight.*
ii) *If c and d are tight, then $c \cap d$ is tight.*

Proof. (i) Let $x \in \tilde{d}(y)$. Then $d(x) = d(y)$ and therefore $\tilde{d}(x) = \tilde{d}(y)$.
Let $x \in \vec{d}(y)$. Then $p(d(x)) \leq p(d(y))$.
Let $\omega \in \vec{d}(x)$. Then $p(d(\omega)) \leq p(d(x)) \leq p(d(y))$, i.e., $\omega \in \vec{d}(y)$.
Thus $\vec{d}(x) \subseteq \vec{d}(y)$.
(ii) Let $x \in c \cap d(y)$. Then $c(x) \subseteq c(y)$ and $d(x) \subseteq d(y)$ since c and d are tight.
Thus $c \cap d(x) = c(x) \cap d(x) \subseteq c(y) \cap d(y) = c \cap d(y)$. □

Proposition 2.6. *If c and d are tight descriptions, then*

i) *$c \subseteq d$ implies $\tilde{c} \subseteq \tilde{d}$,*
ii) *$\widetilde{c \cap d} = \tilde{c} \cap \tilde{d}$.*

Proof. (i) We show that $\tilde{c} \subseteq \tilde{d}$ which implies $\mathcal{N}(\tilde{c}) \geq \mathcal{N}(\tilde{d})$. Take any $x \in \Omega$. For $y \in \tilde{c}(x)$ we have to show $y \in \tilde{d}(x)$. Let $y \in \tilde{c}(x)$, i.e., $c(y) = c(x)$. This implies

$$d(y) \supseteq c(x) \ni x \quad \text{and} \quad d(x) \supseteq c(y) \ni y.$$

By tightness of d we get $d(x) \subseteq d(y)$ and $d(y) \subseteq d(x)$, i.e., $d(x) = d(y)$. This means $y \in \tilde{d}(x)$.
(ii) From (i) we get $\widetilde{c \cap d} \subseteq \tilde{c}$ and $\widetilde{c \cap d} \subseteq \tilde{d}$, thus $\widetilde{cd} \subseteq \tilde{c} \cap \tilde{d}$. The reverse inclusion is Proposition 2.4. □

Proposition 2.7. *For a tight description d the following are equivalent:*

i) *d is symmetric*
ii) *d is complete.*

Proof. *(ii)* $\Rightarrow$ *(i)*: $\tilde{d}$ is symmetric by definition.
(i) $\Rightarrow$ *(ii)*: If $x \in d(y)$, then also $y \in d(x)$ and tightness implies $d(x) = d(y)$, i.e., $x \in \tilde{d}(y)$. Thus $d(y) = \tilde{d}(y)$. □

Proposition 2.8. *For a tight description d the following are equivalent:*

i) *d is directed.*
ii) *$d(\omega) = \{\omega' : d(\omega') \subseteq d(\omega)\} = \{\omega' : p(d(\omega')) \leq p(d(\omega))\} = \vec{d}(\omega)$ for $\omega \in \Omega$.*
iii) *$d(\omega) \subseteq d(\omega')$ or $d(\omega') \subseteq d(\omega)$ for any $\omega, \omega' \in \Omega$.*

Proof. Let d be tight.

(i) $\Rightarrow$ *(ii)*: $\omega' \in d(\omega)$ implies $d(\omega') \subseteq d(\omega)$ which in turn implies $p(d(\omega')) \leq p(d(\omega))$. Therefore $d(\omega) \subseteq \{\omega' : d(\omega') \subseteq d(\omega)\} \subseteq \{\omega' : p(d(\omega')) \leq p(d(\omega))\} = \vec{d}(\omega)$ By (i) all these sets are equal.

(ii) $\Rightarrow$ *(iii)*: Assume $p(d(\omega)) \geq p(d(\omega'))$. Then $\vec{d}(\omega) \supseteq \vec{d}(\omega')$, or vice versa.

(iii) $\Rightarrow$ *(i)*: $p(d(\omega')) \leq p(d(\omega))$ implies $d(\omega') \subseteq d(\omega)$ by (iii) and therefore $\omega' \in d(\omega)$. Thus $\vec{d}(\omega) \subseteq d(\omega)$. Conversely $\omega' \in d(\omega)$ implies $d(\omega') \subseteq d(\omega)$ which in turn implies $p(d(\omega')) \leq p(d(\omega))$ □

Definition 2.12. For a description d we define the description $d^\cap$ by

$$d^\cap(\omega) := \bigcap \{A \in R(d) : \omega \in A\}.$$

$d^\cap$ is called the *tightening* of d.

Proposition 2.9. *For any description d the following holds:*

i) *$d^\cap \subseteq d$ and therefore $\mathcal{N}(d^\cap) \geq \mathcal{N}(d)$,*
ii) *$d^\cap$ is tight*
iii) *$d^\cap = d$, if and only if d is tight*

Proof. (i) Is obvious
(ii) !
(iii) $d^\cap$ is tight: $\omega' \in d^\cap(\omega)$ implies that $d^\cap(\omega') = \bigcap \{A \in R(d) : \omega' \in A\} \subseteq d^\cap(\omega)$ If d is tight and $\omega \in A \in R(d)$, then $d(\omega) \subseteq A$. Thus $d(\omega) \subseteq d^\cap(\omega)$. Together with (i) we obtain $d = d^\cap$ □

The interpretation of $d^\cap$ becomes obvious, when we consider $R(d)$ as the set of propositions that a person (or machine) is willing (or able) to make about an "event" ω. If ω happens, we may collect all he, she or it can say (correctly) about ω. This is $d^\cap(\omega)$.

In Example 2.6 the description c is complete and therefore tight, while d is not tight. In Proposition 2.5 we have seen that also directed descriptions are tight. However, there are many tight descriptions d which are different both from $\vec{d}$ and $\tilde{d}$. The following is an example for this (cf. Exercise 1)).

Example 2.7. Let $\Omega = \{1, 2, \ldots, 6\}$ and define

$$d(1) = d(2) = \{1, 2, 3, 4\},$$
$$d(3) = d(4) = \{3, 4\} \quad \text{and}$$
$$d(5) = d(6) = \{3, 4, 5, 6\}.$$

□

2.5 Information and Surprise of Descriptions

The information of a description is defined as the average novelty provided by its completion.

Definition 2.13. Let d be a description on (Ω, Σ, p). We denote by $\mathcal{I}(d)$ the number $\mathcal{N}(\tilde{d}) = E(\mathcal{N} \circ \tilde{d})$ and call it the *information* provided by d. For $\omega \in \Omega$ we also define the random variable

$$I_d(\omega) := \mathcal{N}(\tilde{d}(\omega)).$$

If d is complete, $\mathcal{I}(d)$ coincides with the usual concept of Shannon information on (Ω, Σ, p) provided by d. (see Shannon and Weaver 1949 and Proposition 2.14).

It is now easy to prove

Proposition 2.10. *For any description d we have*

$$0 \leq N_d \leq I_d$$

and

$$0 \leq \mathcal{N}(d) \leq \mathcal{I}(d) = \mathcal{N}(\tilde{d}) = E(I_d).$$

Proof. see Exercise 4) on page 33.
We first show that $\tilde{d} \subseteq d$. This implies $\mathcal{N}(\tilde{d}) \geq \mathcal{N}(d)$ by Proposition 2.3.

The first elementary observation on $\mathcal{I}(d)$ is that it is always positive, since it is the sum of positive terms $-p(\widetilde{A}) \log p(\widetilde{A})$.[7] Figure 2.3 shows a plot of the function $h(p) = -p \log_2 p$ for $p \in [0, 1]$.

Definition 2.14. Let d be a description.

i) We define the *surprise (of an outcome ω)* of d by

$$S_d(\omega) := -\log_2 p(\vec{d}(\omega)) = \mathcal{N}(\vec{d}(\omega)).$$

ii) We define the *surprise* of a description d by

$$\mathcal{S}(d) := E(S_d).$$

[7] If $p(\widetilde{A}) = 0$ for some $\widetilde{A}$ in the range of $\tilde{d}$, we set $p(\widetilde{A}) \log p(\widetilde{A}) = 0$ since $\lim_{x \to 0^+} x \log x = 0$.

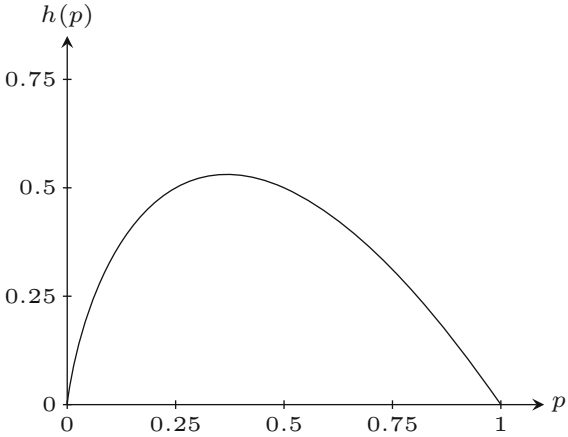

Fig. 2.3 Plot of function $h(p)$

Proposition 2.11. *For every description d we have*

$$0 \leq \mathcal{S}(d) \leq \frac{1}{\ln 2} \text{ and } \mathcal{S}(d) \leq \mathcal{I}(d).$$

Proof. The second inequality follows from $\tilde{d} \subseteq \vec{d}$.

For the first it is sufficient[8] to consider descriptions d with finite $R(d)$. We may further assume that d is directed, i.e., that $R(d) = \{A_1, \ldots, A_n\}$ with $A_1 \subseteq A_2 \subseteq \ldots \subseteq A_n$. Then $\mathcal{S}(d) = -\sum_{i=1}^{n} p(A_i \setminus A_{i-1}) \log p(A_i)$ $(A_0 = \emptyset)$ Let $p_i := p(A_i)$, then

$$\mathcal{S}(d) = -\sum_{i=1}^{n} (p_i - p_{i-1}) \log p_i \leq \sum_{i=1}^{n} -\int_{p_{i-1}}^{p_i} \log_2(x) dx$$

$$\leq -\int_{0}^{1} \log_2(x)\, dx = \frac{1}{\ln 2} \qquad \square$$

Unfortunately, both, information and surprise, do not have the property of monotonicity in general. However, information is monotonic on tight descriptions.

[8] Here we rely on the approximation argument as in Sect. 1.3 for the calculation of an expectation value. If all the finite sum approximations satisfy the same inequality ($\leq \frac{1}{\ln 2}$), then this inequality is also satisfied by the limit.

Proposition 2.12. *If c and d are tight descriptions then $c \subseteq d$ implies $\mathcal{I}(c) \geq \mathcal{I}(d)$.*

Proof. Follows from Proposition 2.6 i). □

From Proposition 2.5 it is obvious that both complete and directed descriptions are tight. So we have monotonicity of information also on complete and on directed descriptions.

Novelty and surprise are both smaller than information, but how do they relate to each other? Usually novelty will be much larger than surprise. For example, a complete description d with $p(d(\omega)) = \frac{1}{n}$ for every $\omega \in \Omega$ has $\mathcal{N}(d) = \mathcal{I}(d) = \log_2 n$, but $\mathcal{S}(d) = 0$. The following example shows, however, that $\mathcal{N} < \mathcal{S}$ is also possible.

Example 2.8. Let $(\Omega, \Sigma, p) = E_{16}$ and[9]

$$\begin{array}{lll} d\colon 1 \to \{1,\dots,11\} & 6 \to \Omega & 14 \to \{7,\dots,16\} \\ \quad 2 \to \{1,\dots,12\} & 7 \to \Omega & 15 \to \{8,\dots,16\} \\ \quad \vdots & \vdots & 16 \to \{9,\dots,16\} \\ \quad 5 \to \{1,\dots,15\} & 13 \to \Omega & \end{array}$$

Ordering $d(\omega)$ by the increasing values of $p(d(\omega))$ for $\omega \in \Omega$, we obtain $d(16), d(15), d(14), d(1), d(2), \dots, d(5), d(6), \dots, d(13)$, where from $d(6)$ on, we have $p(d(i)) = 1$. Thus

$$\begin{array}{lll} \vec{d}\colon 16 \to \{16\} & 1 \to \{1,14,15,16\} & 6 \to \Omega \\ \quad 15 \to \{15,16\} & 2 \to \{1,2,14,15,16\} & 7 \to \Omega \\ \quad 14 \to \{14,15,16\} & \vdots & \vdots \\ & 5 \to \{1,2,3,4,5,14,15,16\} & 13 \to \Omega \end{array}$$

Thus, $N_d(\omega) < S_d(\omega)$ for $\omega \notin \{6,\dots,13\}$ and $N_d(\omega) = S_d(\omega) = 0$ for $\omega \in \{6,\dots,13\}$. Thus $\mathcal{N}(d) < \mathcal{S}(d)$ in this example. □

It is also quite easy to characterize the extreme cases where two of the three quantities $\mathcal{N}, \mathcal{I}$, and $\mathcal{S}$ coincide. This is done in the next proposition.

Proposition 2.13. *Let d be a description, then*

i) $\mathcal{N}(d) = \mathcal{I}(d)$ *implies* $d = \tilde{d}$ *essentially*
ii) $\mathcal{S}(d) = \mathcal{I}(d)$ *implies* $d \equiv \Omega$ *essentially*
iii) *If d is tight then* $\mathcal{N}(d) = \mathcal{S}(d)$ *implies* $d = \vec{d}$ *essentially*

Proof. (i) We have $\mathcal{N}(d) = E(N_d)$, $\mathcal{I}(d) = E(N_{\tilde{d}})$, and $N_d \leq N_{\tilde{d}}$. If for some $\omega \in \Omega$, $N_d(\omega) < N_{\tilde{d}}(\omega)$, the same is true for all $\omega' \in \tilde{d}(\omega)$, and therefore

[9] See Definition 1.3.

$\mathcal{N}(d) < \mathcal{I}(d)$. Thus $N_d(\omega) = N_{\tilde{d}}(\omega)$ and therefore $p(d(\omega)) = p(\tilde{d}(\omega))$ for every $\omega \in \Omega$. Since $\tilde{d}(\omega) \subseteq d(\omega)$, this implies that $\tilde{d}(\omega)$ and $d(\omega)$ are essentially equal.

(ii) Since $\mathcal{S}(d) = E(N_{\vec{d}})$, $\mathcal{I}(d) = E(N_{\tilde{d}})$, and $\tilde{d} \subseteq \vec{d}$ we can again infer that $N_{\vec{d}} = N_{\tilde{d}}$ and that $\vec{d}(\omega) = \tilde{d}(\omega)$ for every $\omega \in \Omega$. For some ω we have $\vec{d}(\omega) = \Omega$, thus $\tilde{d}(\omega) = \Omega$, i.e., $d \equiv \Omega$.

(iii) $\omega' \in d(\omega) \Rightarrow d(\omega') \subseteq d(\omega) \Rightarrow \mathcal{N}_d(\omega') \geq \mathcal{N}_d(\omega) \Rightarrow \omega' \in \vec{d}(\omega)$ □

For further reference we provide a simple comparison of the three basic formulae to calculate $\mathcal{N}$, $\mathcal{I}$, and $\mathcal{S}$.

Proposition 2.14. *For any description d we have*

i) $\mathcal{N}(d) = -\sum_{A \in R(d)} p(\widetilde{A}) \log p(A)$,

ii) $\mathcal{I}(d) = -\sum_{A \in R(d)} p(\widetilde{A}) \log p(\widetilde{A})$,

iii) $\mathcal{S}(d) = -\sum_{A \in R(d)} p(\widetilde{A}) \log p(\vec{A})$.

Proof. Here we definitely need the additional requirement of Definition 2.3. Then we just have to compute the expectation of a step function with at most countably many values. Let us show (iii) for example:

$$\mathcal{S}(d) = E(S_d) = \sum_{x \in R(S_d)} p[S_d = x] \cdot x = -\sum_{A \in R(d)} p[d = A] \cdot \log p(\vec{A}). \qquad \square$$

Example 2.9. Let

$$d(\omega) = \begin{cases} A \text{ for } \omega \in A \text{ and } p = p(A), \\ \Omega \text{ otherwise.} \end{cases}$$

Then

$$\mathcal{I}(d) = -p \log_2 p - (1-p) \log_2 (1-p) =: I(p).$$

This function is plotted in Fig. 2.4. □

In everyday life the common use of the word *information* is closely related with the common use of the word *uncertainty*. In fact, we expect the information provided by the description of a phenomenon to be equivalent to the amount of uncertainty the description eliminates. The concept of uncertainty has been treated in the contexts of thermodynamics and statistical physics (Brush 1966) and has been expressed in terms of *entropy* (a term introduced by Clausius (1865) and defined by a formula similar to Proposition 2.14.(ii) by Boltzmann 1887). It is one of our aims to elucidate the relation between entropy and information in depth (see Chap. 14). For the moment we limit ourselves to observing that the information of a partition is also called its entropy by many authors.

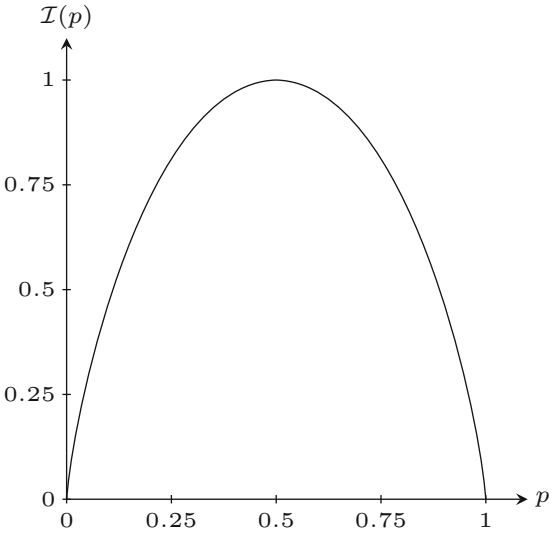

Fig. 2.4 Plot of function $\mathcal{I}(p)$

In a nutshell, the words information, novelty, and surprise introduced here can be distinguished or characterized as follows:

- Information you get whether you can use it or not, whether it is interesting or not,
- Novelty measures how much of this is new and interesting for you,
- Surprise is provided by an event that is comparatively improbable; if everything is equally improbable, nothing is surprising.

The following proposition characterizes descriptions that provide "no information."

Proposition 2.15. *The information of a description d is zero if and only if all sets in the range of its completion $\tilde{d}$ have probability zero except for one set $\widetilde{A}$ with $p(\widetilde{A}) = 1$. In accordance with our additional requirement in Definition 2.3, this is equivalent to $d \equiv \Omega$.*

Proof. Clearly the condition is sufficient: if $p(\widetilde{A}) = 1$ for one set in the range of $\tilde{d}$ and $p(\widetilde{B}) = 0$ for the rest then $\mathcal{I}(d) = 0$. Now, if $\mathcal{I}(d) = 0$ we must have $p(\widetilde{A}) \log p(\widetilde{A}) = 0$ for all $\widetilde{A} \in \tilde{d}(\Omega)$. Therefore $p(\widetilde{A})$ is 0 or 1. Since $\sum_{\widetilde{A} \in \tilde{d}(\Omega)} p(\widetilde{A}) = 1$, exactly one of the sets $\widetilde{A} \in \tilde{d}(\Omega)$ must satisfy $p(\widetilde{A}) = 1$. □

This proposition corresponds to the natural expectation that a description provides no "new" information if we are certain about the outcome of the situation.

The other extreme case that provides maximal information and corresponds to maximal uncertainty is attained by descriptions on which the probability measure is uniformly distributed, as stated by the following proposition.

Proposition 2.16. *For a fixed n consider all descriptions whose range has n elements, that is all d on (Ω, Σ, p) with $d(\Omega) = \{A_1, \ldots, A_n\}$. $\mathcal{I}(d)$ attains a maximum of $\log_2 n$ for $d = \tilde{d}$ and $p(A_i) = \frac{1}{n}$ for each i $(= 1, \ldots, n)$. The same is true for $\mathcal{N}(d)$.*

Proof. see Exercise 8) on page 33. □

Proposition 2.17. *A description with infinite range can have infinite novelty.*

Proof. We give an example for such a description. Let $\Omega = \mathbb{R}_+$, $p(x) = (x+e)^{-1}$ $(\ln(x+e))^{-2}$, where

$$\int_0^\infty p(x)dx = \int_e^\infty x^{-1}(\ln x)^{-2}dx = \Big[-\frac{1}{\ln x}\Big]_e^\infty = 1$$

Let $\alpha = \{[i-1, i) = A_i : i \in \mathbb{N}\}$ and $d(x) = A_i$ for $x \in A_i \in \alpha$. Then $\mathcal{I}(d) = -\sum_{i=1}^{\infty} p(A_i) \log p(A_i)$ where $p(A_i) = p(x_i)$ for some $x_i \in A_i$, so

$$p(A_i) < p(i-1) = (i-1+e)^{-1}(\ln(i-1+e))^{-2} \leq \frac{1}{e} \text{ for } i \in \mathbb{N}$$

and

$$p(A_i) > p(i) = (i+e)^{-1}(\ln(i+e))^{-2}.$$

Thus

$$\begin{aligned} -p(A_i)\log_2 p(A_i) &> -p(i)\log_2 p(i) \\ &> (i+e)^{-1}(\ln(i+e))^{-2}\log_2(i+e) \\ &= (i+e)^{-1}(\ln(i+e))^{-1}(\ln 2)^{-1} \\ &> (\ln 2)^{-1}\int_{i+e}^{i+e+1} \frac{1}{x \ln x}dx. \end{aligned}$$

Therefore

$$\mathcal{I}(d) > (\ln 2)^{-1}\int_{1+e}^{\infty}(x \ln x)^{-1}\,dx = (\ln 2)^{-1}\big[\ln \ln x\big]_{1+e}^{\infty} = \infty.$$ □

Finally we consider two potential properties of information, novelty and surprise that are essential for the further development of a useful information theory.

The first property, already shown in Proposition 2.3 to hold for novelty, is *monotonicity*, i.e., the requirement that finer descriptions should have larger novelty, information, and surprise. Unfortunately, it holds neither for information nor for surprise, because $c \subseteq d$ does not imply $\tilde{c} \subseteq \tilde{d}$, nor $\vec{c} \subseteq \vec{d}$. From Example 2.6 we can easily create an example where $c \subset d$ and $\mathcal{N}(c) > \mathcal{N}(d)$, but $\mathcal{I}(c) < \mathcal{I}(d)$. However, we get monotonicity of $\mathcal{I}$ for tight descriptions because of Proposition 2.6.

Counterexamples against monotonicity of surprise are easy to find. For example, for an equally distributed discrete random variable X, clearly $\widetilde{X} \subset X^{\geq}$, but $\mathcal{S}(\widetilde{X}) = 0$, whereas $\mathcal{S}(X^{\geq}) > 0$.

The other important property of classical information is its subadditivity: $\mathcal{I}(c \cap d) \leq \mathcal{I}(c) + \mathcal{I}(d)$. This will be shown in the next chapter (Proposition 3.5). It is quite easy to see that the novelty $\mathcal{N}$ as defined here does not have this property, i.e., in general, $\mathcal{N}(c \cap d) \not\leq \mathcal{N}(c) + \mathcal{N}(d)$. An example for this can be obtained by considering a description d and its complement d^c. (See also Exercise 7).) Also surprise does not have this property (see Exercise 16)). We will see in the next section that information has this property.

In order to obtain both properties, monotonicity and subadditivity, for both, information and novelty, the definitions given on the level of descriptions in this chapter are not sufficient. We have to elevate these definitions to a higher level, which is done in Part III of the book.

The other possibility is to consider only descriptions with particular properties, for example, tight or complete descriptions. For complete descriptions, information and novelty coincide and clearly have both properties. This is the framework of classical information theory.

2.6 Information and Surprise of a Random Variable

For a measurable mapping $X\colon \Omega \to M$ with values in some finite or countable set M, we may consider the corresponding description $\widetilde{X}$, which is only concerned with the values of X on Ω and defined by

$$\widetilde{X}(\omega) := \{\omega' \in \Omega\colon X(\omega') = X(\omega)\}.$$

We see that the description $\widetilde{X}$ is always complete and that our above definition of $\tilde{d}$ for a description d is just a special case of the definition of $\widetilde{X}$. With the aid of the complete description $\widetilde{X}$, we can define the *information* contained in a random variable X as $\mathcal{I}(X) := \mathcal{N}(\widetilde{X})$. For later reference we restate this definition.

Definition 2.15. For a discrete random variable X we define its (average) *information content* as

$$\mathcal{I}(X) := \mathcal{N}(\widetilde{X}).$$

Remark: Usually information is defined for partitions or for discrete random variables (Shannon and Weaver 1949; Ash 1965; Gallager 1968; Billingsley 1978). Since partitions correspond to complete descriptions in our terminology and since for any random variable X its description $\widetilde{X}$ is complete, this definition coincides with the usual one.

Given an arbitrary random variable $X: \Omega \to \mathbb{R}$ it may happen that $p[X = x] = 0$ for any $x \in \mathbb{R}$. In such a case $\mathcal{N}(\widetilde{X})$ would be infinite and we would need a different workable definition (see Chap. 11.4). Therefore it is sometimes useful to consider other descriptions concerning X, different from $\widetilde{X}$. For example, one may be interested in the largeness or the smallness of the values of X. This leads to the definitions of the descriptions $X^{\geq}$ and $X^{\leq}$. Or one may be interested in the values of X only up to a certain limited accuracy. This leads to the definition of X^{ϵ}.

Definition 2.16. For a random variable $X: \Omega \to \mathbb{R}$ we define the descriptions

$$X^{\geq}(\omega) = \{\omega' \in \Omega: X(\omega') \geq X(\omega)\} \text{ and}$$

$$X^{\leq}(\omega) = \{\omega' \in \Omega: X(\omega') \leq X(\omega)\} \text{ and}$$

$$X^{\epsilon}(\omega) = \{\omega' \in \Omega: |X(\omega') - X(\omega)| < \epsilon\} \text{ for any } \epsilon > 0 \text{ and } \omega \in \Omega.$$

Proposition 2.18. *For a random variable X we have $\mathcal{N}(X^{\geq}) = \mathcal{S}(X^{\geq})$ and $\mathcal{N}(X^{\leq}) = \mathcal{S}(X^{\leq}) = \mathcal{N}(-X^{\geq})$.*

Definition 2.17. For a random variable X we define the *surprise* of X as

$$S(X) := \mathcal{N}(X^{\geq}).$$

This definition provides a simple relation between the surprise of a description d and the surprise of its novelty N_d (see Definition 2.5).

Proposition 2.19. *For any description d we have*

$$S_d(\omega) = \mathcal{N}([N_d \geq N_d(\omega)]) \text{ and therefore } \mathcal{S}(d) = \mathcal{N}(N_d^{\geq}) = \mathcal{S}(N_d).$$

2.7 Technical Comments

Here we introduce information, novelty, and surprise as expectation values of appropriate random variables. For Shannon information this idea was occasionally used (e.g., Khinchin 1957), but it was always restricted to *partitions*, i.e., to *complete descriptions* in our terminology. The more general idea of an arbitrary description, although quite natural and simple, has never appeared in the literature.

In this exposition I have adopted the strategy to disregard propositions of zero probability, because this provides an unrestricted application domain for the ideas introduced. As mentioned in Chap. 1, this more general approach entails some technical difficulties involved in some of our definitions and propositions, mostly

concerning measurability and nonempty sets of probability 0. These difficulties are dealt with in some of the footnotes. Another possibility would have been to develop everything for discrete probability spaces first, assuming $p(\omega) \neq 0$ for every $\omega \in \Omega$, and extend it to continuous spaces later. This is often done in elementary treatments of information theory.

The new concept of *surprise* will provide a bridge from information theory to statistical significance. In earlier papers (Palm 1981), I have called it *normalized surprise*.

The concept of *novelty* was first introduced in Palm (1981) by the name of "evidence." The problem here, is to find yet another word, which has not too many different connotations. Today I believe that "novelty" is the more appropriate word, for such reasons. Proposition 2.14 gives the classical definition of information (Shannon and Weaver 1949).

The concept of a consequential description (Def. 2.11) and the following propositions 2.5 to 2.9 are perhaps a bit technical. These ideas are taken up again in Part IV.

2.8 Exercises

1) For the descriptions given in Examples 2.1, 2.5–2.7 determine their completion, their tightening,their novelty, their information, and their surprise.
2) Let $\Omega = \{0, \ldots, 999\}$ and consider the following random variables describing these numbers:

$$X(\omega) := \text{ first digit of } \omega,$$
$$Y(\omega) := \text{ last digit of } \omega,$$
$$Z(\omega) := \text{ number of digits of } \omega, \text{ for every } \omega \in \Omega.$$

What are the corresponding descriptions, what is the information content of X, Y, and Z, and what is the corresponding surprise (assuming equal probabilities for all thousand numbers)?
3) Measuring the height of a table by means of an instrument with an inaccuracy of about 1 mm can be described by two different descriptions on $\Omega = [500, 1500]$ (these are the possible table heights in mm):

$$d_1(\omega) := \Omega \cap [\omega - 0.5, \omega + 0.5] \text{ and}$$
$$d_2(\omega) := [i, i+1] \text{ for } \omega \in [i, i+1) \text{ for } i = 500, 501, \ldots, 1499.$$

What is the completion in these two cases and what is the average novelty, information, and *surprise* (assuming a uniform distribution of table heights on Ω)?

4) Prove Proposition 2.10 on page 24.
5) Is it true that $\mathcal{I}(X) = \mathcal{I}(X^2)$ for any random variable X?
6) Give an example for a pair (X, Y) of two random variables where
$$\widetilde{(X,Y)} = \widetilde{X} \cap \widetilde{Y} = \widetilde{X \cdot Y}.$$
7) Let $\Omega = \{1, \ldots, n\}$ and $X = \mathrm{id}\colon \Omega \to \mathbb{R}$. For equal probabilities on Ω, what is $\mathcal{N}(X^{\geq})$, $\mathcal{N}(X^{\leq})$, $\mathcal{N}(\widetilde{X})$? What is the limit for $n \to \infty$ in each of the three cases? Observe that $\widetilde{X} = X^{\leq} \cap X^{\geq}$. For which values of n is $\mathcal{N}(X^{\leq} \cap X^{\geq}) > \mathcal{N}(X^{\leq}) + \mathcal{N}(X^{\geq})$?
8) Prove Proposition 2.16 on page 29.

   Hint: *Remove the constraint* $\sum_{i=1}^{n} p(\widetilde{A}_i) = 1$ *by expressing* $p(\widetilde{A}_n)$ *as a function of* $p(\widetilde{A}_1), \ldots, p(\widetilde{A}_{n-1})$. *Then compute a local extremum by setting the derivatives of* $\mathcal{I}(d)$ *to* 0.
9) Given a probability space (Ω, Σ, p) and the events $A, B \in \Sigma$. We say

   - A supports B, if $p(B|A) > p(B)$
   - A weakens B, if $p(B|A) < p(B)$

   If A supports B, which of the relations "supports" and "weakens" hold for the following expressions?

   a) A and B^c
   b) B and A
   c) A^c and B^c
   d) B^c and A

10) Let c, d be descriptions. We say

   - c supports d, if $p(c \cap d) > p(c) \cdot p(d)$
   - c weakens d, if $p(c \cap d) < p(c) \cdot p(d)$
   - c is independent of d, if $p(c \cap d) = p(c) \cdot p(d)$

   Let $\Omega = \{1, \ldots, 6\}$. Give examples for c and d such that they

   a) Weaken each other
   b) Support each other and
   c) Are independent of each other

11) Determine the tightening of the descriptions in Example 2.1 and 2.6.
12) Is it possible that $\mathcal{N}(d^{\cap}) > \mathcal{N}(\tilde{d})$? If yes, give an example; if no, give a proof.
13) Given a description d on (Ω, Σ, p). The function $P\colon \Omega \to [0, 1]$ defined by $P(\omega) = p(d(\omega))$ is a random variable. Can you give an example for a description d for which

   a) $p[P \leq x] = x$,
   b) $p[P \leq x] = x^2$,
   c) $p[P \leq x] = \frac{1}{2} + \frac{x}{2}$

   for every $x \in [0, 1]$?

14) Determine all complete, all directed, and all tight descriptions on $\Omega = \{1, 2, 3\}$.
15) Let $\Omega = \{1, \ldots, 8\}$. Let $c(i) = \{1, 2, 3, 4\}$ for $i = 1, 2, 3, 4$, and $c(i) = \Omega$ for $i = 5, 6, 7, 8$, and $d(i) = \{2, \ldots, 6\}$ for $i = 2, \ldots, 6$, $d(i) = \Omega$ for $i = 1, 7, 8$. Calculate $\mathcal{N}$, $\mathcal{S}$, and $\mathcal{I}$ for c, d, and $c \cap d$.
16) Let $\Omega = \{1, \ldots, 6\}$, $c(1) = c(2) = \{1, 2\}$, $c(i) = \Omega$ for $i = 3, \ldots, 6$, and $d(1) = d(6) = \Omega$, $d(i) = \{2, \ldots, 5\}$ for $i = 2, \ldots, 5$. Calculate $\mathcal{N}$, $\mathcal{S}$, and $\mathcal{I}$ for c, d, and $c \cap d$.

References

Ash, R. B. (1965). *Information theory*. New York, London, Sidney: Interscience.

Billingsley, P. (1978). *Ergodic theory and information*. Huntington, NY: Robert E. Krieger Publishing Co.

Boltzmann, L. (1887). Über die mechanischen Analogien des zweiten Hauptsatzes der Thermodynamik. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, *100*, 201–212.

Brush, S. G. (1966). *Kinetic theory: Irreversible processes, Vol. 2*. New York: Pergamon Press, Oxford.

Bapeswara-Rao, V. V., & Rao, M. B. (1992). A three-door game show and some of its variants. *The Mathematical Scientist*, *17*, 89–94.

Clausius, R. J. E. (1865). Über verschiedenen für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie. *Annales de Physique*, *125*, 353–400.

Gallager, R. G. (1968). *Information theory and reliable communication*. New York, NY, USA: John Wiley & Sons, Inc.

Gardner, M. (1969). *The unexpected hanging and other mathematical diversions.* Simon and Schuster: New York.

Gardner, M. (1959). Mathematical games column. *Scientific American*.

Gillman, L. (1992). The car and the goats. *American Mathematical Monthly*, *99*(1), 3–7.

Granberg, D., & Brown, T. A. (1995). The Monty hall Dilemma. *Personality and Social Psychology Bulletin*, *21*(7), 711–723.

Khinchin, A. (1957). *Mathematical foundations of information theory*. New York: Dover Publications, Inc.

Palm, G. (1981). Evidence, information and surprise. *Biological Cybernetics*, *42*(1), 57–68.

Selvin, S. (1975). On the Monty Hall problem [Letter to the editor]. *The American Statistician*, *29*(3), 134.

Seymann, R. G. (1991). Comment on let's make a deal: The player's Dilemma. *The American Statistician*, *45*(4), 287–288.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. USA: University of Illinois Press.

http://www.springer.com/978-3-642-29074-9

Novelty, Information and Surprise

Palm, G.

2012, XXIV, 248 p., Hardcover
ISBN: 978-3-642-29074-9