# Symmetry and Information Theory[*]

Aleks Jakulin[†]

December 5, 2004

**Abstract**

Before information theory can be applied, we must postulate a particular model of the universe based on probability theory. We journey through the assumptions, advantages and disadvantages of the view. There are three kinds of symmetry or similarity in such a universe: symmetries between probabilities reveal ignorance, symmetries between events reveal indifference, and symmetries between properties reveal information.

## Contents

[*] submitted to Symmetry: Culture and Science

[†] Artificial Intelligence Laboratory, Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, SI-1001 Ljubljana, Slovenia. jakulin@acm.org

# 1 Introduction

In this paper we will discuss similarity and symmetry within the framework of information theory. The starting point in our discussion is a *model*, as it is usually understood in statistics. A model is how we formally describe a particular pattern. But a model is formal, and must be expressed in a particular *language*. For example, linear functions are a language, as are specifications of Turing machines. For a model to have any relation to the reality, it should agree with the *data*. The data are abstractions of measurements or of sensory experiences.

The data does not pretend to be general. It is the task of the model to generalize upon it. Let us consider an example: if the language is geometry, and the data are our experiences involving the sunset and the sunrise, the model will attempt to provide a simple set of geometric statements that explain the many observations. So, the model will be a geometric statement associating the Earth and the Sun as spheres placed in space. The Earth rotates around its axis with a cycle of 24 hours, and the sunrise is defined by the sun becoming visible to a point on the surface of the Earth. This model will be able to predict that the sun will rise tomorrow.

If the model is true, the data can be expected to be its measurements. Even if the model is not true, it can be understood as an approximate explanation of the data. The goal of modelling is to connect the data and the model by forming an expression in the language so that it will agree with the data. A model is thus consistent both with the language (otherwise it cannot be conveyed to others, and remains forever captured inside a black box, a mind or a computer) and with the data (otherwise it is a flight of fancy).

There are many phenomena in real world that are much more complex to model than the mechanics of celestial bodies. Consider the simple coin toss: of course we could measure the exact shape of the coin, the properties of the forces acting upon it, and predict the outcome of tossing it. But these characteristics are often not known to us. Still, we would like to predict the outcome. It is impossible to predict a single outcome, but we can predict how likely different outcomes are. A coin toss is expected to be equally likely to fall heads or tails. In fact, that is why it is used as a random outcome generator.

Coin toss is an example of a phenomenon where it is impossible for the data to be fully consistent with any model, either because we lack the data, or because the god tosses dice. Of course, the data could be perceived in a very restricted way ("Did the coin get tossed?"), by giving up the precision of our perception, ignoring the outcome. Nevertheless, the situations near the transition between tossing and not tossing are always good subject matter for paradoxes ("Does it count as a toss if the coin is intercepted in the air?"), so the solution is rarely perfect. Alternatively, there must be some way of accounting for probability as an aspect of the model, allowing for multiple outcomes of an identical experiment. A particular event is certain only if the probability is 1, and impossible if the probability is 0. Because both heads and tails are possible, their individual probability will lie somewhere in between. Only when such a model is built in the language of probability, we have the foundation for applying Shannon's theory of information (Shannon, 1948).

Shannon's entropy is based upon a model expressed in terms of probability. As such, it has little to do with thermodynamic entropy. If there is no probability in the model, the entropy will be zero. So, any non-trivial model to be considered with information theory should involve uncertainty. In the subsequent sections we will show how probabilistic models look like and what assumptions do they enforce. We will describe the notion of an *attribute* as something knowable about the reality, and then show how information theory helps question attributes and the relationships between them. In the end we will interpret symmetry as a geometric language, and discuss the nature of symmetry in models. Although all the notions of this paper are expressible with mathematical formalizations, our style will be conceptual and expository.

## 2 Models and Probability

We have described models, languages and data, and identified probability as a way of allowing for unpredictability. We will now provide a more specific account of the language of probability. This representation will be the foundation for interpreting any information-theoretic quantity. Namely, as information theory is built upon probability theory, we have to be aware of its limitations. We will also address some of the criticisms of probability. On the other hand, we will not discuss various interpretations of probability: our applications of probability are compatible with several interpretations, but there may definitely be interpretations incompatible with our applications.

Before Shannon's theory of information can be applied, we need to formalize the notion of a 'model'. To do this, we will use two concepts: the universe and the attribute. A universe is a collection of possibilities (sun, clouds, rain), while probability measures the likelihood of each of them (sun: 0.7, clouds: 0.2, rain: 0.1). On the other hand, an attribute wet/not-wet is a shortened projection of the universe (wet:(rain), not-wet:(sun,clouds)). An attribute is a property. Using attributes, we can condition a universe, split it into separate subuniverses, one for each value of the attribute (wet:(rain:1), non-wet:(sun:0.78, clouds:0.22)). Alternatively, we may marginalize a universe by collapsing all events that cannot be distinguished with the given set of attributes (wet:0.3, non-wet:0.7). The following subsections are intended to be an informal introduction to mathematical probability. A reader who desires a more formal approach should refer to other literature, such as (DeGroot and Schervish, 2002).

## 2.1 Universes

Most of Shannon's theory of information is based on the notion of a *probability mass function*, or briefly PMF. When we have several PMFs, we assure their cohesion by having them all derived from an underlying *universe*. The universe is a measure space $\langle S, \mathcal{E}, P\rangle$ based on a discrete set of elementary events $\mathcal{E} = \{e_1, e_2, \ldots, e_n\}$. The set of events is sometimes also referred to as sample space in probability theory, or an alphabet. Note that events may be letters, symbols, things, entities, objects in a bag, states of some machine, outcomes of an experiment, or words in a document: events are merely the carriers of distinction. The formal term for a universe along with probability is a *probability space*, but information theory refers to probability spaces with discrete events.

It is extremely important to note that our universe is a model. It is not necessarily a true model of reality, but of a partial view of reality. It is the goal of statistical mechanics to provide a good model of reality through probabilistic modelling, but we can use the same tools to model anything, such as patients entering a doctor's office. And in such circumstances there is little similarity between Shannon's entropy and Boltzmann's 'quantity called $H$' (Tolman, 1979) which refers to molecules of gas. In retrospect, it was not a good decision to call Shannon's entropy entropy: a more appropriate term would be neginformation.

In the universe, *probability* $P$ is a measure of each event. The probabilities for all these elementary events should sum up to 1: $\sum_i P(e_i) = 1$. Therefore, in every circumstance exactly one of the events should happen. The assumption that elementary events be mutually exclusive is sometimes found problematic, but is easily remedied. One frequent example is the case of the 'excluded middle'. Exactly one of $a$ and $\neg a$, where $\neg a$ signifies not-$a$, is true at the same time. For example, if $a$ signifies a full cup, and $\neg a$ an empty cup, this appears to be a problem. But it is not a problem of assumptions, but of the representation: saying that $\neg a$ marks an empty cup is incorrect, as a cup can be neither full nor empty. More appropriate would be a larger set of four events, based on $a$ signifying a full cup and $\neg a'$ an empty cup: $\{a \wedge \neg a', a \wedge a', \neg a \wedge a', \neg a \wedge \neg a'\}$, here $\wedge$ stands for logical conjunction. It is then the task of the probability to capture the semantic mutual exclusivity of emptiness and fullness: $P(a \wedge a') = 0$, but we could have excluded this joint event when defining the events.

Another problem with probability may be unforeseen circumstances. What happens if we get a broken cup: is it full or empty? Indeed, in some situations we need to create a ghost event $e_0$ which means 'something else' or 'something unforeseen'. Also, it would be incorrect to use a probability larger than 1 to describe an event that has happened several times: this is achieved by creating multiple events $\{a^1, a^2, a^3, \ldots\}$. As these events are mutually exclusive, we have 'invented' natural numbers.

The events and probabilities are considered to be pure and objective. We do not concern ourselves with the notion of an observer and the observed. If this is necessary, the act of observation should be included among the events. For example, if $a$ signifies the sun rising, and $b$ me observing the sunrise, the universe should be modelled as four events: $\{a \wedge b, \neg a \wedge b, a \wedge \neg b, \neg a \wedge \neg b\}$. If the model should allow for the truth of my claims about the sunrise, a further symbol $c$ would need to be combined with $a$ and $b$, and would signify what I claimed about the sunrise. Therefore, these three events capture the situation in its full scope: $a$ - truth, $b$ - what I see as true, and $c$ - what I say.

It is obvious that our model is of limited precision: we cannot break any event into its constituent parts - the events are atomic and internally indistinguishable. Should we want to do that, we would need a new sample space, a new universe. The universe is the embodiment of the notion of an ontology: the list of conceivable events along with the possibility, impossibility and probability of each one of them. If one prefers the language of logic, the set of events is the set of possible atomic statements, the probability of each event is their semantics, so that each resulting statement has a probability. The mathematical structure of a sample space generalizes upon this notion by allowing aggregations of elementary events.

The choice of the universe is in the eye of the beholder. The beholder only distinguishes those nuances that matter. There are unimaginably many possible states, but as beholders, we choose not to distinguish all of them. We might distinguish 37.001 and 37.002 as abstract numbers, but we would generally not distinguish them if they indicated the body temperature as one variable in medical diagnosis. On the other hand, 37.049 and 37.051 would be distinguished in the universe where rounding to the nearest number turned them into 37.0 and 37.1, but not in another universe where all numbers are rounded down. We avoid associated problems by allowing for a number of universes that model the same reality: ultimately the choice of the universe is an event like any other. Furthermore, we may have several probability measures for the same universe: each choice of a probability measure is an event. Finally, all that we truly require is that the probabilities are consistent within a particular universe, and that universes can be coalesced into a single universe which agrees with the above assumption of mutual exclusivity and completeness.

It is also possible to model dynamics with the concept of the universe. Given a static universe $\mathcal{E}$, the dynamic universe is a Cartesian product of the universe before and the universe after: $\mathcal{E}_{before} \times \mathcal{E}_{after}$. The implicit time of the dynamic universe is also discrete: 'before' and 'after' are distinctly separated. At the same time, the model is unable to account for its possible changes through time: it is necessarily invariant with respect to translations in time. The invariance of some kind, with respect to moments, types of cups, translations in time or something else, facilitates the repeatability of a particular event. Multiplicity or repeatability of occurrence of an event, or at least belief in the occurrence of an event is what is needed to speak about probability. A 'thick time' model of the universe would be $\mathcal{E}_0 \times \cdots \times \mathcal{E}_{now}$, but only ignorance or multiplicity of universes (multiverse) would allow probability.

The data $\mathcal{D}$ is represented as a multiset of events, or as a set of *instances* or measurements: a single event may have happened several times and so corresponds to several instances, just the same temperature can be obtained through several acts of measurement. This means that the universe may not distinguish every pair of instances, either due to ignorance or intentional disregard. There is no ordering of instances, unless the order is a part of each event. Many possible probability measures are consistent with a given set of data: the only requirement is that each instance has non-zero probability.

It is possible to *learn* the probability from the data, too: we can seek the probability assignments that make the data as likely as possible (Fisher, 1912). Or, more generally, we can use the Bayes rule to assign probabilities to different probability measures consistent with the data, e.g. (Good, 1965; Jaynes, 2003), thereby creating a universe of probability measures. In some cases it is necessary to interpret the data probabilistically, especially with unreliable sensors or with real-valued measurements. The temperature reading of 37.0 degrees Celsium may be interpreted as an observation that the true temperature has a uniform distribution between 37.05 and 37.15 degrees Celsium: an additional source of uncertainty. Not to get bogged down in this complexity, we will always consider a single universe with a single probability measure. However, if this universe is nested as an event within another universe, so will every statement or conclusion based on it.

## 2.2 Attributes

We have interpreted the universe as a formalization of everything that can be distinguished. There is no structure in the universe, it is a mere structureless list. We will now consider the notion of an *attribute* $A$ as a construct built on top of the universe. We will start with a binary attribute, the simplest of all, whose *range* $\Re_A$ is $\{0, 1\}$. The binary attribute $A$ is a function $A : \mathcal{E} \to \Re_A$. Thus, for each event, we will know whether the attribute took the value of 0 or 1. The attribute merges all the states of the universe into those that have the value of 0 and those that have the value of 1: the attribute values are mutually exclusive. By summing all the corresponding event probabilities, we can obtain the attribute value probabilities. We can also envision a universal attribute whose range is the universe itself: the universe itself is then the original attribute, the alphabet. [1]

There are a few arguments against attributes. First, fuzzy logic (Zadeh, 1965) disagrees with the notion of an attribute which takes a single crisp value for each event. Instead, fuzzy logic recommends using grades of membership of an attribute value for each event. We will attempt to do the same in the existing framework by introducing the notion of a 'perception' or a *sensor*. The sensor is unreliable, and may or may not react to a particular event. But this can easily be handled within our notion of events and attributes. As earlier, we will include the sensor reading $\{s, \neg s\}$ into the universe, obtaining four events: $\{a \wedge \neg s, a \wedge s, \neg a \wedge s, \neg a \wedge \neg s\}$. If the sensor is precise, $P(a \wedge \neg s)$ and $P(\neg a \wedge s)$ will be low. Nevertheless, there is a good reason why sensors should not always be precise: consider $a$ indicating the height of $183.2321\ldots$ centimeters and the $s$ signifying 'tall': there is a good reason for working with clumpier $s$ rather than with $a$. Of course, if we have several heights and several sensors, the situation of sensors 'tall' and 'very tall' both taking the value of 1 for the same objective height is perfectly possible. When there are $k$ mutually exclusive binary attributes, meaning that for each event in the universe there is exactly one of them taking the value of 1, we may replace them all with a single $k$-ary attribute with the range $\{1, 2, \ldots, k\}$. This is a major gain in economy, but it is contingent upon mutual exclusivity.

Another common assumption is the invariance of a sensor: it should remain the same for all instances, in the same way as an event is atomic. This assumption is not always realistic: there may be drift through time (Widmer and Kubat, 1996), old sensors may not be the same as new sensors and consequences once upon the time are no longer the same. A systematic deviation from this assumption cannot be captured by the model, and the resulting model will carry some uncertainty because of that. The solution lies in introducing a sensor's sensor, indicating if the sensor is new, old or broken. And one can continue by including the sensor of a sensor's sensor.

In other circumstances, the value of the attribute might be unknown or undefined. Assume patients coming to a physician, so each patient is an event. For some patients, the body temperature is known, but for others it is not. Technically, the attribute's range must then include 'not known' as one of its values. Even in the binary case, we can imagine the range {'1', '0 or unknown'}. Alternatively, we may create a binary attribute {'temperature known', 'temperature unknown'}, and *condition* the universe to contain only those patients whose temperature is known. In that conditional universe, the temperature is always known. This conditioning is always implicit: the patients themselves are conditioned on the binary attribute 'The event is a

[1] More formally, an attribute is essentially a random quantity, and each attribute value corresponds to an element of the event space in probability theory. The attributes whose range is the set of real numbers $\mathbb{R}$ are sometimes referred to as random variables, and that is why we are using the term 'attribute' and not 'random variable'.

patient coming to a physician.' in the first place. The probabilities in each branch of a conditional universe sum up to 1.

The second kind of operation is *marginalization.* Here, we take a set of attributes, for example $\{A, B, C\}$, and collapse all events that cannot be distinguished with these attributes into elementary ones. For example, if we marginalize away all colors except for two $A : \{\text{black}, \text{not-black}\}$ and $B : \{\text{white}, \text{not-white}\}$, every color will be mapped either to black, white or gray (not-black and not-white). Furthermore, zero probability is put in effect for each combination $\langle a, b, c\rangle$ of attributes' values, $\langle a, b, c\rangle \in \Re_A \times \Re_B \times \Re_C$, that cannot be found in the universe (such as 'black and white'). In the example of physician's patients, the attribute 'astrological signs' has been marginalized away and is not known or used by the physician (presumably). On the other hand, 'body temperature' is usually marginalized away in the discourse of an astrologer. In all, contemporary medicine generally assumes that all people are equal, and this assumption both allows generalizing from one patient to others, but also prevents distinguishing specific characteristics of patients. Some theories of probability claim that that probability is purely a result of marginalization and a consequence of the fact that the causes are not known.

In all, we see that attributes can be seen as projections of the universe, as views of the universe. Marginalization serves as integration, as merging of events, and probability reflects this merging. On the other hand, conditioning creates separate universes, each of them with a consistent definition of probability. The universe serves as a unified foundation for defining the relationships between attributes, and in turn, these attributes serve as means for characterizing the events. It is possible to construct or remove attributes as deemed suitable, and these attributes will transform the perception of the universe.

## 3 Information Theory

In the previous section we have described the three crucial elements needed to discuss entropy: the universe $\mathcal{E}$, the probability $P$ and the attributes $A, B, C, \ldots$. We can now begin to disentangle the model with information theory. We will show the connection between entropy, investment and growth. In the second subsection, we will justify other information-theoretic expressions through questions we can ask about the truth. We will use the results and expressions of (Shannon, 1948).

### 3.1 Entropy as a Bound on Growth

We will now consider the definition of entropy through gambling, following a popular information theory textbook (Cover and Thomas, 1991). Assume that we are playing a game, trying to predict what event will take place. We start with $K$ coins, and place a bet on each of the events in the universe, expressing it as a proportion of $K$. So for event $e_i$, our bet is $b(e_i)$, while $\sum_{e\in\mathcal{E}} b(e) = 1$. We now let some event $e'$ happen, and our gain is $MKb(e')$, where $M$ is the maximum reward multiplier: had we bet everything on $e'$, $b(e') = 1$, our funds would increase $M$-fold. Therefore, our funds multiply by $Mb(e')$.

Clearly, we would achieve the highest growth of funds by putting all the money on the single most likely event, but would also lose everything if that event did not happen. Alternatively, we minimize the chances by betting on every outcome equally, but if there are too many possible events, we would be losing in every game. It can be shown that the maximum rate of growth out of all possible betting portfolios is achieved by betting proportionally to event probabilities, so that $P(e) = b(e)$, and this is called the Kelly gambling scheme. The doubling rate of the horse race using the proportional gambling is $\log_2 M + \sum_{e\in\mathcal{E}} P(e)\log_2 P(e)$. It is easy to see that for an omniscient player the game is worth playing only if $\log_2 M > H(\mathcal{E})$, or in other words, if the logarithm of the rewards exceeds the *Shannon* or *information entropy* of the universe:

$$H(\mathcal{E}) := -\sum_{e\in\mathcal{E}} P(e)\log_2 P(e) \tag{1}$$

Of course, it is impossible to stop playing with reality.

Realistic observers, however, are not omniscient, and their portfolio $b$ deviates from the true distribution $P$. For them, the doubling rate is $\log_2 M - H(\mathcal{E}) - D(P\|b)$, where $D(P\|b)$ is the Kullback-Leibler divergence or relative entropy between the truth $P$ and their belief $b$:

$$D(P\|q) := \sum_{e\in\mathcal{E}} P(e)\log_2 \frac{P(e)}{q(e)} \tag{2}$$

It is important to understand that a linear change either in entropy or in KL-divergence corresponds to a linear change in the rate of growth. Entropy is the minimum rate of growth for an omniscient predictor. Furthermore, the rate of growth or demise is essentially linked with the ability to place bets well. The same conclusion is valid also if a different $M$ is specified for each event $m(e)$, only the $\log_2 M$ would be replaced by $\sum P(e)\log_2 m(e)$.

That we refer to money should not be a distraction. Instead of money, we could refer to food, energy, and even information. Plants successfully photosynthesize because they correctly predict that there will be sunlight coming vertically down from the sky, a coal-powered power plant successfully operates because the engineers correctly predict that heated air will expand. A scientist successfully publishes papers if she correctly predicts that some of her experiments will demonstrate something new.

We can express entropy and divergence in the terms of loss functions. Consider that the player whose betting portfolio is $q$. He suffers the loss of

$-\log_2 q(e)$ in the case of event $e$. This means that we have a loss function $S(e, q) = -\log_2 q(e)$: the less the player bet, the more he lost. This specific loss function is used in data compression, where we pay each symbol proportionally to the logarithm of the probability with which we predicted it. Data compression programs, such as zip and unzip, are nothing else than successful gamblers.

The *expected loss* is the expectation of player's loss. The player is using an imperfect model of reality with $q$ instead of the true probability $P$. The Shannon entropy corresponds to the minimum expected loss, suffered by the omniscient player: $H(\mathcal{E}) = \inf_q E_{e \sim P}\{S(e, q)\}$. The KL-divergence thus corresponds to the player's expected loss beyond the omniscient player's: $D(P\|q) = E_{e \sim P}\{S(e, q) - S(e, P)\}$. We could also understand these expressions as definitions of entropy and divergence based on some loss function. Entropy and divergence are just specific definitions of loss and gain, and we may introduce quantities corresponding to entropy and divergence with different definitions of this loss (Grünwald and Dawid, 2004). Of course, not all properties would be retained.

Probability can be defined also for continuous universes with real-valued events. Unfortunately, such a definition of the universe is inappropriate for Shannon entropy. A different concept of *differential entropy* is then defined as $h(X) = \int P(x) \log_2 P(x) dx$, but its properties differ from those of Shannon entropy. For example, the differential entropy may be zero or negative. The KL-divergence, on the other hand, works satisfactorily also with continuous universes.

## 3.2 Queries and Contexts

Let us now focus on three attributes, $A$, $B$ and $C$ in the context of some creature. The range of $A$ is $\Re_A = \{\text{happy}, \text{unhappy}\}$, $B$'s range covers possible actions that our focal creature can take, while $C$ is a sensor of its environment. The creature cannot know $A$, is unsure about $B$, but does know $C$. The default uncertainty of our creature depends solely on $A$, and can be quantified by $H(A) = -\sum_{a \in \Re_A} P(a) \log_2 P(a)$. This implies that the creature performs no intervention different than usual: it does not mean that no action is taken, just that the creature does not think about it. If the creature does decide to think about taking an action, the uncertainty is reduced to $H(A|B) = H(A, B) - H(B)$. Conditional entropy is always lower or equal to the unconditional entropy. If it is equal, being conscious about $B$ did not reduce the creature's uncertainty about $A$ and its effort was wasteful. The unfortunate situation of the consequence being independent of the action is possible: turning the steering wheel left or right is quite independent of being happy after the drive.

A more promising conditional entropy is $H(B|C)$, the entropy of the action in the context of the sensor. Indeed, it is a lot easier to decide whether to turn left or right based on the sensor than on nothing else: the entropy of the 'action' variable $B$ is lower, meaning that the uncertainty about the action is lower given the information about the surroundings. It is also possible to be more specific about conditional entropy. We could ask about the uncertainty of the action in a particular context: $H(B|C =$ standing in front of a red light$)$ is low because there is little the creature can do. On the other hand, $H(B|C =$ slow truck in front, other lane empty$)$ requires some thought whether to overtake or not. If we do not specify the value of the context, the conditional entropy is the expectation over the possible contexts $H(B|C) = \sum_c P(c)H(B|C = c)$.

The conditional entropy of $A$ given $C$ is not necessarily low. Of course, there are situations that are inherently more ambiguous and uncertain, and some that we know well. But the key connection between $A$ and either of $B$ and $C$ is through the *interaction* between $B$ and $C$: we are happy, if our action was appropriate for the circumstances and unhappy otherwise. Whether the action was good is only learned through $A$, and can be formed as a question "Which action $b$ in the context of $c$ will make us happy with the highest probability, or unhappy with the lowest probability?" The entropy can also be seen a measure of how difficult it is to make a decision: the larger the difference between these two probabilities, the lower the entropy. If we are to seek a new sensor $D$, the most desirable one would minimize $H(A|B, C, D)$: it would allow the best prediction of happiness along with the existing actions and sensors. Simply minimizing $H(A|D)$ would ignore the contributions of $D$ beyond those already provided by $B$ and $C$.

### 3.3 Similarity as Information

Conditional entropy manifests how the uncertainty may be decreased by consideration of other attributes. We treat all satisfaction, action and knowledge as attributes. Our true bets are placed on satisfaction, but we cannot bet on satisfaction directly - we can only place bets on our actions, and hope that these will translate to good bets on the outcomes. We do not place bets randomly, but we let perceptions $C$ inform us on how to place bets. Furthermore, an active creature may introduce new attributes to decrease the entropy of making decisions. But with all the possible conditional entropies, how to make sense of reality? *Information* performs this deed by decomposing the entropy.

Earlier, we mentioned that the lower the $H(A|B)$, the better the attribute $B$ in reducing $A$'s entropy. It was already Shannon who proposed *mutual information*, defined as $I(A; B) = H(A) + H(B) - H(A, B)$ for the purpose of understanding the relationship between $A$ and $B$. Mutual information is symmetric with respect to the ordering or roles of individual attributes: $I(A; B) = I(B; A)$, and can be used to reconstruct conditional entropy: $H(A|B) = H(A) - I(A; B)$, $H(B|A) = H(B) - I(A; B)$. It can also be interpreted as the loss caused by assuming independence between $A$ and $B$: $I(A;B) = D(P(A,B)\|P(A)P(B))$. Finally, mutual information forms the backbone of *Rajski's distance* (Rajski, 1961), a metric on attributes defined as:

$$d_R(A,B) := 1 - \frac{I(A;B)}{H(A,B)} \tag{3}$$

The higher the mutual information, ceteris paribus, the closer the two attributes should be in the cognitive space of the universe and the probability measure on it. Rajski's distance is always in the interval $[0,1]$, as mutual information is never larger than the entropy of either attribute, and the entropy of either attribute cannot be larger than the joint entropy of them both. It is notable that similarity is no longer postulated but instead inferred through the notions of information theory. A good similarity measure will be the one that will enable us to predict well (Baxter, 1997), and with mutual information we tie similarity between $A$ and $B$ directly to how well $A$ is predicted from $B$ and vice versa.

It is possible to carry the entropy decomposition approach further, and to speak of 3-way or 4-way interactions. This course of work has been pioneered by (Quastler, 1953) and (McGill, 1954), independently at first, but later together (McGill and Quastler, 1955). Their definition of *interaction information* or is best understood as a lattice structure of the entropy space (Han, 1975), while visualizations of this space are described by (Jakulin and Bratko, 2004). All such $k$-way interactions are symmetric with respect to the ordering of attributes.

The notion of the context is important here. For example, $A$ and $B$ may not be dependent by default. However, in some context $C$, they may turn out to be dependent. In such cases, we may employ the notion of conditional mutual information: $I(A;B|C) = H(A|C) + H(B|C) - H(A,B|C)$. We now have a context-dependent notion of similarity.

Sometimes, two attributes become more dependent in a context, a situation revealed by positive interaction information. In such a case, we speak about positive interactions or synergies. For example, the employment of a person and criminal behavior are not particularly dependent attributes (most unemployed people are not criminals, and many criminals are employed), but adding the knowledge of whether the person has a new sports car suddenly makes these two attributes dependent: it is a lot more frequent that an unemployed person has a new sports car if he is involved in criminal behavior; the opposite is also true: it is somewhat unlikely that an unemployed person will have a new sports car if he is not involved in criminal behavior.

On the other hand, two attributes become less dependent inside the context $C$. We detect this with negative interaction information. For example, the attributes of weather, rain and lightning, are dependent because they occur together. But the attribute storm interacts negatively with them, since it reduces their dependence. Storm explains a part of their dependence. Should we wonder whether there is lightning, the information that there is rain would contribute no information if we already knew that there is a storm. In a specific context two attributes may be perfectly independent. This is an indication of conditional independence, meaning that any kind of dependence or correlation between them can be reduced to the fact that they share the context. The view that the context may be seen as the cause underlies certain theories of causality (Pearl, 2000; Spirtes, Glymour and Scheines, 2000).

## 4 Symmetry as a Language

Let us now try to provide a universe for a geometric space. Let us assume the familiar Cartesian geometry. The purpose of our endeavor is to be able to represent a geometric figure as a data set in the universe. The events of the universe could simply be points in space. A figure is a set of points, and can be represented as a sequence of events, one for each point. We are also able to compute the entropy of a particular model of figures. Our attributes can represent lines or shapes. Consequently, we can compute the entropy in the context of a particular line.

Unfortunately, the probabilities need to sum up to 1. Because points are infinitesimal, it will take infinitely many points to describe an object of finite area. Most of practical data sets will thus be infinite, and this is impractical. Let us then assume that the size of the point is specified as a constant $d$, restricting some minimum resolution, and disregarding everything smaller than that. But the precision of the point's displacement is still infinitesimal, and the specification of any single point's position itself will still take infinitely many bits of information.

A common assumption is that the coordinates are multiples of $d$ as well. Now we have pixels with integer coordinates, the fundamental representation in computer graphics. The problems due to having to round coordinates to integers are referred to as aliasing. But without restricting the dimensions of the universe, we are still unable to represent a figure. In computer graphics the dimensions of the window, viewport or an image are indeed restricted. A similar solution has been used by Boltzmann to model gas: all that we are interested in is the integer number of different states in an integer number of discrete sections of a container.

Is there any way around this? One answer lies in symmetry. Even if no two points are the same, we can identify "sameness" between assemblages of points at different positions in the space. Symmetry is nontrivial equality (Petitjean, 2003): there is symmetry between spatial objects $X$ and $Y$ if $X$ equals a transformation of $Y$. Therefore, given a collection of points, symmetries form an algebraic *language* for describing geometry. We can say: 'Draw a point. Transform the point into a line. Copy the line three times and obtain the square.' Of course, symmetry of shape is just one kind of a transformation. Color symmetries imply that different shapes have the same color. Distance symmetries imply that the distances between shapes are equal, and the shapes are equidistant. Relative distances can be symmetric too: all articles have the titles on the top of the first page.

Fractal image compression (Barnsley, 1988; Wohlberg and de Jager, 1999) is based on the realization that an image can be represented algebraically as an iterated function system (IFS) whose fixed point is close to the original image. Note that in this image compression universe, each image is an event, not each point. The iterated function system is a union of contractive affine maps. The fixed point or the basin of attraction is unique, through the Banach fixed point theorem. However, the image is no longer represented as a matrix of pixels, but instead as a finite set of affine maps, each with its parameters (scaling, rotation, translation). Furthermore, the fixed point has infinite resolution, unlike any quantization of the image. With the ubiquity of self-similarity in nature, such a representation yields realistic representations of real world images. Or perhaps our own perception too focuses on discovering symmetry and self-similarity and is thus unable to distinguish the fixed point from the reality.

The major practical problem of fractal image compression is how to acquire the parameters of the IFS that match a specific image well. The collage theorem (Barnsley, 1988) shows that only the first few iterations of the mappings are sufficient to judge the overall quality of the attractor. A further breakthrough that put fractal image compression into practice was the realization that it is easier to apply transforms to parts of the image than to the image as a whole.

In summary, fractal image compression is based on representing an image purely in terms of its symmetries. Because compression 'works', the image can be compressed into fewer bits than by ordinary quantization into pixels. If fewer bits are required for the same set of images, betting in the universe of iterated function systems yields lower entropy than betting in the universe of quantized pixels. Recently, image compression algorithms based on wavelets have won over fractal image compression, and this means that currently betting in the space of smooth images is better than betting in the space of self-similar ones. Some day, perhaps, both symmetry and smoothness will be included in the betting portfolio of compression algorithms.

Lossy image compression algorithms are all mapping a set of images that all look equivalent to human perception into the same sequence of bits. A possibly large number of images are thus symmetric with respect to our perception. This is a reminder that both smoothness and self-similarity are means used by human visual system to reduce the entropy of the real world. This does not mean that symmetry and smoothness are necessarily aspects of the real world but that slight asymmetry and roughness are often invisible to our brain. Furthermore, if our brain functions in terms of symmetry and smoothness to reduce entropy, it will find those images, situations and objects that are symmetric and smooth preferable to others. The preference may arise purely from lower cost of processing, perhaps explaining what the principles of aesthetics refer to.

# 5 Entropy and Symmetry

## 5.1 Invariance and Symmetric Universes

Symmetry can also refer to algebraic entities, not just to geometric ones. For example, through Galois' notion of permutation group, symmetry is an invariance regarding a group of transformations (Mostow, 1996; Brassard, 1998). With the permutation group we ignore the ordering of letters in a string, and all that matters is the frequency of individual letters. Or, with the group of rotations, we disregard the orientation of an object, only considering its shape. Probability too is born from symmetries: we do not care about when and how the event happens, but only about how likely it is to happen, or about the number of times it happened.

However, there is also a different kind of symmetry here. What is an event? The structure of the universe is undoubtedly simpler than the structure of reality. All observations of the same event are indistinguishable. Similarly, Boltzmann did not distinguish between various arrangements of gas molecules within a specific microstate. The set of events of the universe can be seen as a result of applying some set of transformations to reality and each event describes all the states of reality that get mapped to it. The partitioning of reality into events of the universe has established an explicit mapping from reality into events, and implies the assumption of symmetries in reality. The fewer the events, the more symmetry we have assumed in the universe.

If we start with a particular universe, we can modify it by establishing further symmetries between individual events. In a particular universe $\mathcal{E}$, we could collapse two events $e_1$ and $e_2$ into a single $e_{1+2}$ with the probability $P(e_{1+2}) = P(e_1) + P(e_2)$. This could be interpreted as an increase in symmetry, as we distinguish less than before. But what would happen to Shannon entropy? Because logarithm is a monotonically increasing function, it is trivial to see that $P(e_1)\log_2 P(e_1) \leq P(e_1)\log_2(P(e_1) + P(e_2))$ and that $P(e_2)\log_2 P(e_2) \leq P(e_2)\log_2(P(e_1)+P(e_2))$, therefore also $P(e_2)\log_2 P(e_2) + P(e_1)\log_2 P(e_1) \leq (P(e_1) + P(e_2))\log_2(P(e_1) + P(e_2))$. *Increasing symmetry of the universe by collapsing events entails decreasing Shannon entropy.*

## 5.2 Maximum Entropy and Symmetric Models

There is another kind of symmetry that refers to probabilities. Imagine that the probabilities form the universe. If the same probability, such as 0.1, appears for several events, we can say that there is some symmetry within the model. The universe with probabilities $\{a : 0.25, b : 0.25, c : 0.25, d : 0.25\}$ is more symmetric than $\{a : 0.2, b : 0.3, c : 0.4, d : 0.1\}$ because it only has a single probability. This means that the model is symmetric, even if the reality is not. For example, in the common model of the coin toss, we do distinguish the heads from tails, but by saying that the probability of both heads and tails is 0.5, we have established symmetry between these two probabilities. In inference it is desirable to establish such symmetries through the *principle of insufficient reason*, or the Bayes-Laplace postulate (Bernardo and Smith, 2000): in absence of evidence to the contrary, all possibilities should have the same initial probability. In a more general context with the events being expressed as infinitesimal points in some space, symmetries can be established between areas with the same probability density.

Let us focus on the simple example of probabilities of two discrete events within some larger model, and assume that $p_1 > p_2$. We could equalize these probabilities by moving them closer to one another: $d = k(p_2 - p_1), p'_1 = p_1 + d, p'_2 = p_2 - d$. However, since the logarithm's first derivative is a monotonically decreasing function, the entropy cannot increase by equalization. This was noted already by (Shannon, 1948). Therefore, *increasing similarity or symmetry among probabilities in the model reflects in increasing Shannon entropy of the universe.* This principle underlies the maximum entropy principle (Jaynes, 2003), which states that among several models of probability, one should choose the model that yields the highest entropy. Such a model is the least pretentious, contains the fewest dependencies, and has the most symmetric probabilities.

But if the model itself is considered a universe, symmetries between probabilities allow us to represent it in a more compact way. Imagine that the model is a sequence of events, and each event is a probability. There would be such an alphabet of probabilities: $\{0, 0.25, 0.5, 0.75, 1\}$. In a symmetric model, certain probabilities will be more likely than others, and so the *descriptions of* a symmetric model have lower Shannon entropy than the descriptions of an asymmetric model. Yet, the universe described by a symmetric model has higher Shannon entropy than a universe described by an asymmetric model. It is helpful to think of probabilities of probabilities in this context.

Most of statistical modelling is about balancing between these two symmetries: the attempt to provide a sharp and symmetric view of reality, to reduce entropy of the universe and to maximize the predictive power of the model. On the other hand, the goal is to provide simple, smooth and symmetric probabilistic models. There are dangers to either approach. A sharp picture of reality might be overidealistic and unrealistic. The problem that ensues is *overfitting*: the model may be overconfident. On the other hand, smoothing the probabilistic model draws a pessimistic view of reality as unpredictable and disorganized, yet the model of such disorganized reality is surprisingly smooth and organized. This problem is referred to as *underfitting*.

In maximum entropy inference, we start with sharp constraints that should minimize our entropy, then investigate all the models that conform to the constraints, and pick the maximum entropy and thus the most symmetric one amongst them all. In classical statistics, we pick a maximum entropy distribution and then discover sharp parameters that minimize the entropy: it is well-known that most distributions in statistics are maximum entropy models given some constraint (Kapur, 1990): the normal distribution results from constraining the mean and the variance, the uniform distribution results from constraining the bounds, etc. Therefore, most practical statistical methods are balanced in the sense that they combine both entropy maximization and minimization.

# 6 Conclusion

Several authors have recently investigated the relationship between symmetry and information theory, but the interpretations often oppose one another. For example, both (Lin, 2001) and (Petitjean, 2003) note that symmetry is closely associated with similarity. However, increasing symmetry can be interpreted both as decreasing or increasing entropy (Lin, 2001). Our framework agrees with both interpretations, and adds another one.

**Symmetries hidden inside events:** By making assumptions that simplify views of reality, we also postulate symmetry and similarity in the data (Sect. 5.1). Henceforth, the Shannon entropy of the universe is reduced. For example, instead of allowing any wavelength for a photon ($\mathcal{E} = \{\text{blue}, \text{yellow}, \text{green}, \text{violet}, \ldots\}$), we can restrict the photons to just three kinds: red, green and blue, and through this symmetry, we arrive to a simpler universe $\mathcal{E}'$. With this, our indifference about the subtypes of events has entered the mode, and we have decreased the Shannon entropy of the universe: $H(\mathcal{E}') \leq H(\mathcal{E})$.

**Symmetries between probabilities:** The language may enforce symmetry and similarity in the models it describes (Sect. 5.2). This way, the probability of two events is assumed to be identical, just as in the coin toss. Similarly, both tails of the Gaussian distribution are assumed to be identical. The resulting simplicity of the model increases the Shannon entropy of the universe, making it seem harder to predict than it really is. On the other hand, if we consider each probability in the model to be an event, e.g., $\mathcal{P}(\mathcal{E}) = \{0, 0.5, 1\}$, collapsing two unique probabilities $p_1$ and $p_2$ into $p'_1 = p'_2 = (p_1 + p_2)/2$ can decrease the resulting entropy of the model $H(\mathcal{P}'(\mathcal{E}))$, but increase the entropy of the universe $H(\mathcal{E}|\mathcal{P}')$. In general, symmetries between probabilities indicate some kind of ignorance, the inability to differentiate between the likelihoods of events.

**Symmetries between attributes:** Two attributes provide information about one another when their mutual information is high (Sect. 3.3). This can be interpreted as symmetry or similarity between them. For example, because of symmetry in reality, the binary smoking attribute and the binary lung cancer attribute have high mutual information. We exploit these symmetries by being able to infer that a smoker is more likely to fall ill with lung cancer, but also that lung cancer victims are often smokers. Neither the entropy of the universe nor the entropy of model is affected by the symmetries between attributes. Instead, choosing and constructing useful attributes for prediction is guided by the knowledge of symmetries between them.

## References

Barnsley, M. (1988). *Fractals Everywhere.* Academic Press.

Baxter, J. (1997). The canonical distortion measure for vector quantization and approximation. In *ICML 1997* (pp. 39–47). Morgan Kaufmann.

Bernardo, J. M. & Smith, A. F. M. (2000). *Bayesian Theory.* Chichester: Wiley.

Brassard, L. (1998). *The Perception of the Image World.* PhD thesis, Simon Fraser University, Burnaby, BC, Canada.

Cover, T. M. & Thomas, J. A. (1991). *Elements of Information Theory.* Wiley Series in Telecommunications. New York: Wiley.

DeGroot, M. H. & Schervish, M. J. (2002). *Probability and Statistics* (third Ed.). Addison-Wesley.

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, *41*, 155–160.

Good, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, Volume 30 of *Research Monograph.* Cambridge, Massachusetts: M.I.T. Press.

Grünwald, P. D. & Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, *32(4)*.

Han, T. S. (1975). Linear dependence structure of the entropy space. *Information and Control*, *29*, 337–368.

Jakulin, A. & Bratko, I. (2004). Quantifying and visualizing attribute interactions: An approach based on entropy. `http://arxiv.org/abs/cs.AI/0308002` v3.

Jaynes, E. T. (2003). In G. L. Bretthorst (Ed.), *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press.

Kapur, J. N. (1990). *Maximum Entropy Models in Science and Engineering*. John Wiley & Sons.

Lin, S.-K. (2001). The nature of the chemical process. *International Journal of Molecular Sciences*, *2*, 10–39.

McGill, W. J. (1954). Multivariate information transmission. *Psychometrika*, *19(2)*, 97–116.

McGill, W. J. & Quastler, H. (1955). Standardized nomenclature: An attempt. In Quastler, H. (Ed.), *Information Theory in Psychology: Problems and Methods* (pp. 83–92). Glencoe, Illinois: The Free Press.

Mostow, G. D. (1996). A brief survey of symmetry in mathematics. *Proc Natl Acad Sci USA*, *93(25)*, 14233–14237.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. San Francisco, CA, USA: Cambridge University Press.

Petitjean, M. (2003). Chirality and symmetry measures: A transdisciplinary review. *Entropy*, *5(3)*, 271–312.

Quastler, H. (1953). The measure of specificity. In H. Quastler (Ed.), *Information Theory in Biology*. Urbana: Univ. of Illinois Press.

Rajski, C. (1961). A metric space of discrete probability distributions. *Information and Control*, *4*, 373–377.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423, 623–656.

Spirtes, P., Glymour, C. & Scheines, R. (2000). *Causation, Prediction, and Search* (2nd Ed.). New York, N.Y.: MIT Press.

Tolman, R. C. (1979). *The Principles of Statistical Mechanics*. New York: Dover.

Widmer, G. & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning*, *23(1)*, 69–101.

Wohlberg, B. & de Jager, G. (1999). A review of the fractal image coding literature. *IEEE Transactions on Image Processing*, *8(12)*, 1716–1729.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*, 338–353.