The Minimum Description Length Principle and Non-Deductive Inference

P. Grünwald
CWI, P.O. Box 4079, 1009 AB
Amsterdam, The Netherlands
Email: pdg@cwi.nl

1 Introduction

This short position paper is about the following claim involving the Minimum Description Length (MDL) Principle, a principle originally developed for inductive inference:

> *The scope of the MDL Principle is much broader than just inductive inference: we claim it can be profitably used in many areas of reasoning under uncertainty; specifically, it can be used as a direct means of, or at least a guideline for, providing reasonable abductive inferences.*

The space allowed here is much too short to really defend this claim; rather, we will say something about the Minimum Description Length Principle and how its relationship to the Maximum Entropy and Conditional-Independence-Based methods for uncertainty reasoning leads us to believe that it should be taken into consideration into any discussion about the similarities and differences of abduction and induction.

2 The MDL Principle

The Minimum Description Length (MDL) Principle [Rissanen, 1989] is a widely applicable method for doing *inductive inference*. As such it has been applied numerous times in the fields of statistics and machine learning, generally with very good results. Its theoretical properties have been well-investigated (see for example the works by J. Rissanen, who introduced MDL in its present form [Rissanen, 1989; 1996]).

The MDL Principle, which can be viewed as a precise mathematical formalization of 'Occam's Razor', is based on the following two fundamental ideas: first, it is recognized that if we are given some data, and we want to learn from this data, then this means that we are looking for the *regularities* in the data, i.e. the *laws* that seem to underly it. Second, it is claimed that *any* regularity in the data at hand can be used to *compress* this data, i.e. to describe it with a smaller number of bits than would be needed if all of the data were written down in a binary alphabet. From this second claim it is concluded that the more we can compress our data, the more knowledge we have about it. Specifically, if we want to explain our data in terms of a specific hypothesis, then the more such a hypothesis compresses the data, the better it explains it.

Most of the work done on the MDL Principle is actually about the *predictive* MDL Principle [Rissanen, 1989] in which a third, very important idea is added: if we learn something from a given set of data, then we may want to use it (the 'something' we learned) to predict future data, i.e. data that is not known to us during the learning process but will be made available later on. Inference should thus be done in a way that allows for predictions of future data. Moreover:

> *the ultimate test for an inductive inference procedure is whether it performs well in predicting future data.*

Interestingly, prediction and compression turn out to be closely related. One can prove that under quite general conditions, an hypothesis that allows for a lot of compression of the data at hand also allows for good prediction of future data and vice versa.

If we want to compress our data, we need a *code*: a 1-1 function mapping all possible sets of data to encoded versions (the 'descriptions') of them and back. Now it follows from a basic fact of information theory [Rissanen, 1989] that there is a 1-1 correspondence between codes and probability distributions: for any probability distribution P defined on data coming from some set of possible data ('outcomes') $\mathcal{O}$, there is a code C_P such that for any $D \in \mathcal{O}$, $L_{\mathrm{C}_P}(D) = -\log_2 P(D)$, i.e. the description length of D when the encoding is done using code C_P is equal to $-\log P(D)$. Conversely, for any code C defined on data sets D, there is a probability distribution P_{C} such that for all D, $P_{\mathrm{C}}(D) = 2^{-L_{\mathrm{C}}(D)}$. We can thus *identify* codes with probability distributions; this brings MDL into the realm of probability theory and statistics. Notice that a short code length implies a high probability and vice versa.

Usually, when applying MDL, we want to explain our data in terms of one or more hypotheses for this data. These hypotheses are then taken from a set of candidate hypotheses $\mathcal{H}$ (we will see later where this set comes from). Often, the candidate hypotheses are probabilistic in nature (i.e. any $H \in \mathcal{H}$ defines a probability distribution over the space of possible data sets). In such a case, any hypothesis $H \in \mathcal{H}$ can be used to encode any data set $D \in \mathcal{O}$ using $-\log_2 P(D|H)$ bits. Here $P(\cdot|H)$ stands for the probability distribution defined by H.

We note here that most (according to Rissanen: *all*) hypotheses that may ever be of interest to us can be cast in probabilistic terms, even if they do not explicitly refer to any probability at all. This perhaps somewhat surprising claim is quite forcefully defended in [Rissanen, 1989], chapter 2. Throughout this paper we assume that all our hypotheses are indeed probabilistic.

The question is then how to arrive at a mathematically precise definition of the MDL Principle. Actually, a completely satisfactory solution to this problem has only been found very recently [Rissanen, 1996]. The way it is defined there, the MDL Principle says the following:

Suppose we are given a set of candidate hypotheses $\mathcal{H}$ for a set of possible data $\mathcal{O}$. Then we should look for the code C^* that minimizes the additional number of bits needed to encode the data, compared to the number of bits needed by the hypothesis $\hat{H} \in \mathcal{H}$ that is optimal for the particular data $D \in \mathcal{O}$ that was observed.

We now clarify this seemingly very complicated definition. Our data D comes from some outcome space $\mathcal{O}$. For the moment, we are only concerned with the finite case: there are only a finite number of possibilities for D. Now for each particular outcome D, there is some $\hat{H} \in \mathcal{H}$ that is optimal for that D: among all hypotheses in $\mathcal{H}$, $\hat{H}$ gives the highest probability to D. Hence

$$\hat{H}(D) = \arg\max_{\hat{H} \in \mathcal{H}} P(D|H) \qquad (1)$$

We write $\hat{H}(D)$ to indicate that $\hat{H}$ depends on D. Since log and $-$ are monotonous transformations, (1) is equivalent to $\hat{H}(D) = \arg\min_{\hat{H} \in \mathcal{H}} -\log_2 P(D|H)$. In light of the discussion above, this means $\hat{H}$ is also the hypothesis that allows for the shortest encoding of the observed data D. For any H, we will write $L_H(D) := -\log_2 P(H|D)$; in other words, $L_H(D)$ is the length of the encoding of D when the encoding is done with the code that corresponds to $P(\cdot|H)$.

We now look for a code that, when compared to $\hat{H}$ uses as few *additional*[1] bits as possible to encode D – whatever the actually observed D will turn out to be. This leads to the following definition:

$$\mathrm{C}^* = \arg\min_{\mathrm{C}} \max_{D \in \mathcal{O}, H \in \mathcal{H}} L_{\mathrm{C}}(D) - L_{\hat{H}(D)}(D) \qquad (2)$$

Here C ranges over all possible codelengths for the outcome space $\mathcal{O}$. $L_{\mathrm{C}^*}(D)$, i.e. the length of data D when encoded using C^* is called the *Stochastic Complexity* of data D; it is the central notion in MDL theory. Note that $2^{-L_{\mathrm{C}^*}(D)}$ determines a probability distribution over all $D \in \mathcal{O}$ that, in a sense, gives *as much probability as possible* to any $D \in \mathcal{O}$. We define $P^*(D) = 2^{-L_{\mathrm{C}^*}(D)}$.

In the *predictive* MDL Principle, this P^* is used to arrive at a prediction of future outcomes of $\mathcal{O}$. It can be shown that, under some regularity conditions, P^* is optimal for prediction in the sense that, whatever the sequence of outcomes $D_1, \ldots, D_n$ will turn out to be, predicting according to P^* will yield a prediction error that is very close to the prediction error of the optimal $H \in \mathcal{H}$, i.e. the H that has lowest prediction error on $D_1, \ldots, D_n$ [Rissanen, 1996].

In MDL *hypothesis* selection, P^* is used to deduce the single $H \in \mathcal{H}$ that best explains the data. It turns out that – in most cases – the optimal H is the $H \in \mathcal{H}$ that minimizes the sum of a) the description length of H and b) the description length of data D when it is described with the help of H (i.e. $-\log_2 P(D|H)$).

MDL and Bayesian Statistics

It turns out that MDL is closely related to Bayesian Statistics; in many contexts, MDL and Bayesian Statistics allow for the same inferences. But interestingly – at least in its current form, [Rissanen, 1996] – MDL avoids one of the central problems of Bayesian Statistics, i.e. the problem of how to choose a good prior distribution over all the candidate hypotheses. In stead of taking a prior distribution that reflects 'initial degrees of belief' for the various candidate hypotheses in $\mathcal{H}$, MDL takes the prior distribution π such that the marginal probability $P(D) = \sum_{H \in \mathcal{H}} P(D|H)\pi(H)$ (in case $\mathcal{H}$ ranges over a continuous set, we replace the sum by an integral) is as close as possible to $P^*(D)$. This prior distribution turns out to be optimal in the sense that if it is used for prediction, it gives the lowest possible worst-case prediction error [Rissanen, 1996].

Several other methods of statistical inference can be seen as special cases of the MDL Principle; we focus on one of them: The *Maximum Entropy (ME) Principle.*

3 The ME Principle and Reasoning under Uncertainty

The ME Principle has been introduced by Jaynes [Jaynes, 1978] as a general means of reasoning in the presence of uncertainty. It can be seen as an extension

[1] It is very easy to show that there can be no code C that, for any $D \in \mathcal{O}$, uses as few or fewer bits to encode D than H does.

of Laplace's 'Principle of Indifference': if we have no reason to prefer any one out of N possibilities, we should attach to each of them the probability $1/N$, this being the 'least committal', or 'most inherently uncertain' probability distribution consistent with what we know. (In this case, what we know is that there are N possible outcomes). By means of the concept of entropy, Jaynes has shown how to extend this idea to deal with partial knowledge of the following sort (here E stands for 'expected value'):

$$E[f_i(X)] = d_i \quad \text{for some real } d_i. \tag{3}$$

The ME Principle says the following: if we are given an outcome space $\mathcal{O}$, the definitions of a finite number of functions f_i and values d_i and additional knowledge that the probability distribution we are looking for adheres to the constraints (3) for all i, then, from among all those probability distributions satisfying constraints (3), we should pick the one with the highest *entropy*. The entropy of a probability distribution P defined on a finite space of outcomes $\mathcal{O}$ (we are not concerned with the infinite case here) is defined as $-\sum_{x \in \mathcal{O}} P(x) \log P(x) = E[-\log P(x)]$. We will write the set of all known constraints of the form (3) as $\mathbf{K}$; if the constraints in $\mathbf{K}$ hold for a distribution P, we write $P \models \mathbf{K}$. We can now formally define the ME principle as: given outcome space $\mathcal{O}$ and constraints $\mathbf{K}$, one should prefer the distribution

$$P_{ME} = \arg \max_{P: P \models \mathbf{K}} E[-\log P(D)] \tag{4}$$

The Maximum Entropy Principle has seen several applications in AI-oriented uncertainty reasoning; see for example [Goldszmidt *et al.*, 1993; Grove *et al.*, 1994]. Typically, in these works ME is applied to *reasoning with defaults* : the default $a \rightsquigarrow b$ ('if a, then normally also b') is translated to a probabilistic statement $P(b|a) = 1 - \epsilon$ where ϵ is some very small number; i.e. given that a is true, b has a high probability. The set of constraints $\mathbf{K}$ then consist of a number of such statements. The preferred models of a logical theory are arrived at by first determining P_{ME} according to (4) and then picking the models with the highest P_{ME}-probability.

'Inevitability' of ME

In [Paris and Vencovská, 1990; Shore and Johnson, 1980] some theorems are proved which state that under certain, quite general assumptions, ME is the only consistent way to do plausible inference. These strong theorems notwithstanding, the ME Principle has always been hotly debated; ME's view on probability as a subjective degree of belief is unacceptable to many; also, there are several well-known problem domains where it yields counterintuitive conclusions, for example in the presence of disjunctive constraints (see below) [Goldszmidt *et al.*, 1993]. Furthermore, one of its chief shortcomings is that it *does not allow for learning* [Grove *et al.*, 1994]. The constraints of form (3) allow one to express knowledge that permit several forms of plausible reasoning, among which abductive, but not inductive reasoning, as is discussed in [Grove *et al.*, 1994; Paris and Vencovská, 1990].

4 MDL and ME

It turns out that under quite general conditions, the MDL Principle permits the same inferences as the ME Principle. However, in those cases that the author is aware of where the ME Principle fails, the MDL Principle still works. We give three simple examples:

Suppose our data consists of a finite data sequence of l a's and b's. That is all we know; we have no idea on whether there are any regularities in it or not, i.e. the set $\mathbf{K}$ is empty. The ME Principle (4) now says that we should attach probability $1/2^l$ to each sequence. Notice that this is equivalent to a probability distribution P over $\{a, b\}$ with $P(a) = P(b) = 1/2$ and the additional assumption that all outcomes are *probabilistically independent*, i.e. for all D_1, D_2, $P(D_1, D_2) = P(D_1)P(D_2)$.

In this case, the MDL Principle gives us the same, but for a different reason: it says we should pick the hypothesis H that is the solution of (2). Since we know nothing at all about what H's are possible and what not, we take the set of candidate hypotheses $\mathcal{H}$ to consist of *all* possible probability distributions over the set of outcomes of length n. Taking this $\mathcal{H}$, (2) yields a P^* that gives probability $1/2^l$ to each sequence, or equivalently $P^*(a) = P^*(b) = 1/2$; $P(D_1, D_2) = P(D_1)P(D_2)$. The code C^* corresponding to this P^* encodes any l-bit data sequence by an l-bit description – we have gained nothing, but also, we have *lost* nothing. For *some* possible data sequences, another code may be better. But also, *every* other code is worse than C^* on some data sequences: every code other than C^* has a worse worst-case behaviour than C^*.

In the presence of constraints things get more interesting. Consider the famous 'Brandeis Dice' example [Jaynes, 1978]: we are given a six-sided die and the additional information that the expected outcome of throwing the die is 4.5 rather than the more familiar 3.5, i.e. we have the constraint set $\mathbf{K} = \{E[X] = 4.5\}$. We also assume that the outcome of throwing the die at any time t is independent on the outcome of a throw at any other time t'. Now ME tells us to that among all the distributions $\{p_1, \ldots, p_6\}$ (here p_i stands for $P(X = i)$) with $E[X] = 4.5$, we should prefer the one given by (4). We get:

$$P_{ME} = \{p_1, \ldots, p_6\} =$$

$$\{0.05435, 0.07877, 0.11416, 0.16545, 0.23977, 0.34749\} \tag{5}$$

What does MDL tell us in this case? To find the solution of (2), we need to define the set of possible hypotheses $\mathcal{H}$. Actually, this set is given by the set of constraints $\mathbf{K}$! Having a constraint set $\mathbf{K}$ can be interpreted as: 'we know the constraints in $\mathbf{K}$ hold. So we are looking for *the best H out of the set of* all *hypotheses for which these constraints hold!'* This means that we can redefine (2) in terms of $\mathbf{K}$:

$$\mathrm{C}^* = \arg\min_{\mathrm{C}} \max_{D \in \mathcal{O}, H : H \models \mathbf{K}} L_{\mathrm{C}}(D) - L_{\hat{H}(D)}(D) \tag{6}$$

It turns out that in the dice example, MDL and ME again yield the same answer: $P^* = P_{ME}$. In general, one can prove that as long as all constraints are of the form (3), MDL and ME give the same answer. But if the constraints are of a more complex nature, MDL and ME can give different answers. For example, suppose again we have a six-sided die, but now we have a *disjunctive* constraint on the expected values: $\mathbf{K} = \{E[X] = 4.5 \vee E[X] = 2\}$. ME now tells us that we should still pick (5): among all $P \models \mathbf{K}$ it has highest entropy. MDL however tells us in this case to take P^* according to (2), which gives

$$P^*(X = i) = 1/2 \cdot P_1^*(X = i) + 1/2 \cdot P_2^*(X = i)$$

Here P_1^* is the MDL distribution for the constraint set $\mathbf{K} = \{E[X] = 4.5\}$, i.e. the distribution given by (5). P_2^* is the MDL distribution for the constraint set $\mathbf{K} = \{E[X] = 2\}$. P_2^* looks as follows:

$$P_2^* = \{p_1, \ldots, p_6\} = \\ \{0.5328, 0.4781, 0.1357, 0.0723, 0.0385, 0.0205\} \tag{7}$$

Now why should we prefer P^* over P_{ME}? To answer this, let us first look again at P_{ME} in the example without any disjunction: we note that P_{ME} is acceptable only to the subjectivist, Bayesian statistician: it can be interpreted as the least 'comitted' belief in what the outcomes of throwing the die will be; however, it can *not* be interpreted in the objectivist sense: p_i is *not* necessarily equal to the frequency with which outcome i will occur. To see this, observe that it could very well be the case that the real probability distribution underlying the die is $\{0, 0, 0, \frac{1}{2}, \frac{1}{2}, 0\}$ – we would then still have $E[X] = 4.5$, so the constraint would be satisfied. This is one of the facts which have made ME so controversial.

From the MDL point of view, we look at probabilities neither as beliefs nor as frequencies, but rather as (things isomorphic to) codelengths. It can be shown [Rissanen, 1989] that short codelengths (i.e. high probabilities) directly relate to good prediction capabilities: it is easy to prove that, if the given constraints on the probability distributions indeed hold, predicting outcomes using P_{MDL} minimizes the worst-case prediction error. This means that for any other probability distribution $P' \models \mathbf{K}$, there is a $P'' \models \mathbf{K}$ such that if P'' is the 'true' distribution generating the data, then predicting according to P' will result in higher prediction error than predicting according to P_{MDL}. Here 'prediction error' is measured in terms of the logarithmic loss [Rissanen, 1989]; a low logarithmic loss implies a low square prediction error, a notion that may be more familiar to the reader. In the presence of the disjunctive constraint, it may be possible that the 'true' distribution is such that $E[X] = 4.5$. In that case, using P_{ME} will yield a lower prediction error than using P^*. But it may also be the case that $E[X] = 2$, in which case P_{ME} will yield a much higher prediction error. Using P_{MDL}, it makes no difference whether actually $E[X] = 2$ or $E[X] = 4.5$. The prediction error for P_{MDL} in both cases will be equal, and in the former case better than for P_{ME} - the *worst-case* prediction error for P_{MDL} is much better than for P_{ME}.

We see that MDL gives us a new interpretation of probabilities as codelengths; this interpretation is philosophically interesting because of its links to Occam's Razor, but it is also actually *useful*: it allows us to predict well future data. Strangely, the criterion that an inference method should be able to predict well future data, while common-place in statistics and machine learning, is not often encountered in the literature on abduction.

5 Conditional Independence and Abduction

If one looks for the optimal joint probability distribution of a number of variables, and one knows nothing about any relationship between these variables, then both ME and MDL prescribe to pick a probability distribution in which *all variables are probabilistically independent*. A recent trend in statistics is to use graphs for modelling all the dependencies that one assumes to hold in a domain of variables. The 'graphical models' [Spiegelhalter *et al.*, 1993] which result contain as a special case the so-called 'Bayesian Networks' which have been used in many AI applications, most notably expert systems [Spiegelhalter *et al.*, 1993]. These Bayesian Network expert systems are full of 'independence assumptions' of the form 'once we know the value of variable A, variable B and C become independent'. According to the MDL Principle, one can make these independence assumptions even if the variables may really be dependent after all, but one just does not know it: if one does not know whether they are dependent or not, assuming they are independent leads to the lowest worst-case prediction error and thus seems the best thing to do. Hence, in our view, in Bayesian Network expert systems the MDL Principle is implicitly used all the time. As the basic reasoning steps performed in most expert systems are of an abductive nature, we think this shows that in some form, MDL is indeed being used in abduction already.

6 MDL, Induction and Abduction

We have focused here on the use of MDL in uncertain reasoning that was *not* of an inductive nature. This seemed to us important since many people think that MDL is there only for induction. The fact that MDL can be used for both says a lot about the relationship between abduction and induction. If we are given the possibility of writing a longer paper, we will discuss this in detail. In this longer paper we will also discuss how MDL/ME can be used for logical theories and qualitative knowledge. Though we have neglected this fact in this short paper, it should be stressed here that this, too, is possible [Goldszmidt *et al.*, 1993; Paris and Vencovská, 1990].

References

[Goldszmidt *et al.*, 1993] M. Goldszmidt, P. Morris, and J. Pearl. A maximum entropy approach to nonmonotonic reasoning. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 15(3):220–232, 1993.

[Grove *et al.*, 1994] A.J. Grove, J.Y. Halpern, and D. Koller. Random worlds and maximum entropy. *Journal of AI Research*, 2:33–88, 1994.

[Jaynes, 1978] E.T. Jaynes. Where do we stand on maximum entropy? In R.D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*, pages 15–118. MIT Press, Cambridge, MA, 1978.

[Paris and Vencovská, 1990] J.B. Paris and A. Vencovská. A note on the inevitability of maximum entropy. *International Journal of Approximate Reasoning*, 4(3):183–224, 1990.

[Rissanen, 1989] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific Publishing Company, 1989.

[Rissanen, 1996] J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Inf. Theory*, 1(42):40–47, 1996.

[Shore and Johnson, 1980] J.E. Shore and R.W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, IT-26:26–37, jan 1980.

[Spiegelhalter *et al.*, 1993] D.J. Spiegelhalter, A.P. Dawid, S.L. Lauritzen, and R.G. Cowell. Bayesian analysis in expert systems (with discussion). *Statistical Science*, 8:204–283, 1993.