5.5 Hypothesis Identification by Minimum Description Length

We can formulate scientific theories in two steps. First, we formulate a set of possible alternative hypotheses, based on scientific observations or other data. Second, we select one hypothesis as the most likely one. Statistics is the mathematics of how to do this. A relatively recent paradigm in statistical inference was developed by J.J. Rissanen and by C.S. Wallace and his coauthors. The method can be viewed as a computable approximation to the noncomputable approach in Section 5.2 and was inspired by it. In accordance with Occam's dictum, it tells us to go for the explanation that compresses the data the most.

Minimum description length (MDL) principle. Given a sample of data and an effective enumeration of the appropriate alternative theories to explain the data, the best theory is the one that minimizes the sum of

- the length, in bits, of the description of the theory;
- the length, in bits, of the data when encoded with the help of the theory.

The idea of a two-part code for a body of data D is natural from the perspective of Kolmogorov complexity. If D does not contain any regularities at all, then it consists of purely random data, and there is no hypothesis to identify. Assume that the body of data D contains regularities. With help of a description of those regularities (a model) we can describe the data compactly. Assuming that the regularities can be represented in an effective manner (that is, by a Turing machine), we encode the data as a program for that machine. Squeezing all effective regularity out of the data, we end up with a Turing machine representing the meaningful regular information in the data together with a program for that Turing machine representing the remaining meaningless randomness of the data. This is the intuition.

Formally, assume that our candidate theories are effective computation procedures, that is, Turing machines. While with the Kolmogorov complexity we represent the amount of information in the individual objects, here we are interested in the division of this information into a two-part code, as in Section 2.1.1. First, the "valuable" information representing regularities that are presumably usable (the model part of length $K(T)$), followed by the "useless" random part of length $C(D|T) = l(p)$ as in Equation 2.1 on page 99. However, it is difficult to find a valid mathematical way to force a sensible division of the information at hand in a meaningful part and a meaningless part. One way to proceed is suggested by the analysis below.

MDL is based on striking a balance between regularity and randomness in the data. All we will ever see are the data at hand (if we know more, then in fact we possess more data, which should be used as well). The best model or explanatory theory is taken to be the one that optimally uses regularity in the data to compress. "Optimally" is used in the sense of maximal compression of the data in a two-part code, that is, a model and a description from which the data can be reproduced with help of the model. If the data are truly random, no compression is possible and the optimum is reached for the "empty" model. The empty model has description length 0. If the data are regular, then compression is possible, and using the MDL principle (or the Kolmogorov complexity approach) identifies the optimal model.

We call such a "model," or "theory," a "hypothesis." With a more complex description of the hypothesis H, it may fit the data better and therefore decrease the misclassified data. If H describes all the data, then it does not allow for measuring errors. A simpler description of H may be penalized by an increase in misclassified data. If H is a trivial hypothesis that contains nothing, then all data are described literally and there is no generalization. The rationale of the method is that a balance between these extremes seems to be required.

Ideally, the description lengths involved should be the shortest effective description lengths, the prefix complexities, which however cannot be effectively computed. (This obviously impedes actual use. In practice, one needs to consider computable approximations to shortest descriptions, for example by restricting the allowable approximation time.) The code of the shortest effective self-delimiting descriptions, the prefix complexity code, gives the least expected code-word length—close to the entropy (pages 181, 231 or Section 8.1)—and moreover compresses the regular objects until all regularity is squeezed out. All shortest effective descriptions are completely random themselves, without any regularity whatsoever. Kolmogorov complexity can be used to develop a theory of (idealized) minimum description length reasoning. We rigorously derive and justify this Kolmogorov complexity based form of minimum description length, "ideal MDL," via the Bayesian approach using the universal distribution $\mathbf{m}(\cdot)$ of Section 4.3.1 as the particular prior distribution over the hypotheses. This leads to a mathematical explanation of correspondences and differences between ideal MDL and Bayesian reasoning, and in particular it gives some evidence under what conditions the latter is prone to overfitting while the former is not.

The analysis of both hypothesis identification by ideal MDL in this section, and of prediction in Section 5.2.2, shows that maximally compressed descriptions give good results on data samples that are random with respect to probabilistic hypotheses. These data samples form the overwhelming majority and occur with probability going to one when the length of the data sample grows unboundedly. That is, both for hypothesis identification and prediction, data compression is provably optimal but for a subset of (hypothesis, data sample) pairs of vanishing probability.

5.5.1 Derivation of MDL from Bayes's Rule

Let us see how we can rigorously derive an ideal version of MDL from first principles, *in casu* Bayes's rule as given by Equation 5.1. While this does not rigorously prove anything about MDL in its applied forms, the relations with Bayesian inference we establish for ideal MDL are corroborated by empirical evidence for applied MDL.

Recall Bayes's Rule, Sections 1.6, 1.10, 5.2:

$$\Pr(H|D) = \frac{\Pr(D|H)P(H)}{\Pr(D)}.$$

Here H is a hypothesis, P is the prior probability of the hypotheses and D is the set of observed data. In this equation we are only concerned with finding the H that maximizes $\Pr(H|D)$ with D and P fixed. Taking the negative logarithm of both sides of the equation, this is equivalent to *minimizing* the expression $-\log \Pr(H|D)$ over H:

$$-\log \Pr(H|D) = -\log \Pr(D|H) - \log P(H) + \log \Pr(D). \tag{5.15}$$

Since the probability $\Pr(D)$ is constant under varying H, this means we want to find an hypothesis H that minimizes

$$-\log \Pr(D|H) - \log P(H). \tag{5.16}$$

In applied forms of MDL one roughly speaking interprets these negative logarithms $-\log P(x)$ as the corresponding Shannon-Fano (or Huffman) code-word lengths. But why should one use the Shannon-Fano code (or Huffman code) and no other code reaching an expected code word length equal to the entropy? In particular, ignoring feasibility, why not use the objective shortest effective code, the shortest effective descriptions with code-word length set equal to the prefix complexities. This also has an expected code-word length equal to the entropy (pages 181, 231 or Section 8.1), but additionally, the shortest program compresses the object by effectively squeezing out and accounting for all regularities in it. The resulting code word is maximally random, that is, it has maximal prefix complexity.

For now let us assume that H and D are expressed as natural numbers or finite binary strings. To obtain the ideal MDL principle it suffices to replace the probabilities involved in Equation 5.16 by the *universal probability* $\mathbf{m}(\cdot)$ of Theorem 4.3.1 on page 247. The analysis of the *conditions* under which this substitution is *justified*, or conversely, how application of ideal MDL is equivalent to Bayesian inference using *admissible* probabilities, is deferred to the next section. Therefore, under conditions to be established below, we substitute according to

$$\begin{aligned} \log P(H) &= \log \mathbf{m}(H) + O(1), \qquad (5.17)\\ \log \Pr(D|H) &= \log \mathbf{m}(D|H) + O(1). \end{aligned}$$

By Theorem 4.3.3 on page 253, we have

$$\log \mathbf{m}(H) = -K(H) \pm O(1),$$
$$\log \mathbf{m}(D|H) = -K(D|H) \pm O(1),$$

where $K(\cdot)$ is the prefix complexity. Therefore, using the substitution of Equation 5.17 we can replace the sum of Equation 5.16 by the sum of the minimum lengths of effective self-delimiting programs that compute descriptions of H and $D|H$. That is, we look for the H that minimizes

$$K(D|H) + K(H), \tag{5.18}$$

which is the code-independent, recursively invariant, absolute form of the MDL principle.

The term $-\log P(D|H)$ is also known as the *self-information* in information theory and the *negative log-likelihood* in statistics. It can be regarded as the number of bits it takes to redescribe or encode D with an ideal code relative to H.

Example 5.5.1 If we replace all P-probabilities in Equation 5.15 by the corresponding $\mathbf{m}$-probabilities, we obtain in the same way by Theorem 4.3.3

$$K(H|D) = K(H) + K(D|H) - K(D) + O(1).$$

In Theorem 3.9.1 on page 232 it is shown that symmetry of information holds for individual objects in the following sense:

$$\begin{aligned} K(H,D) &= K(H) + K(D|H, K(H)) + O(1) \\ &= K(D) + K(H|D, K(D)) + O(1). \end{aligned} \tag{5.19}$$

Substitution gives $K(H|D) = K(H,D) - K(D)$, up to an $O(\log K(H,D))$ additive term. The term $K(D)$ is fixed and doesn't change for different H's. Minimizing the left-hand term, $K(H|D)$ can then be interpreted as

> **Alternative formulation MDL principle.** Given a hypothesis space $\mathbf{H}$, we want to select the hypothesis H such that the length of the shortest encoding of D together with hypothesis H is minimal.

◇

The discussion seems to have arrived at its goal, but a danger of triviality lurks nearby. Yet it is exactly the solution how to prevent trivialities, which gives us the key to the very meaning of ideal MDL, and by extension some insight in applied versions.

Since it is not more difficult to describe some object if we get more conditional information, we have $K(D|H,K(H)) \leq K(D|H) + O(1)$. Thus, by Equation 5.19 the quantity in Equation 5.18 satisfies

$$K(H) + K(D|H) \geq K(H,D) + O(1) \geq K(D) + O(1),$$

with equalities for the trivial hypothesis $H_0 = D$ or $H_0 = \emptyset$. At first glance this would mean that the hypothesis H_{mdl} that minimizes the sum of Equation 5.18 could be set to D or $\emptyset$, which is absurd in general. However, we have only derived the validity of Equation 5.18 under the condition that Equation 5.17 holds. The crucial part of our justification of MDL is to establish precisely when Equation 5.17 is valid.

5.5.2 The Role of Universal Probability

It is well known, see Example 1.11.2, that the so-called Shannon-Fano code for an ensemble of source words distributed according to probability Q is a prefix code E_Q with $l(E_Q(x)) = -\log Q(x)$ satisfying

$$\sum_x Q(x)l(E_Q(x)) = \min_{E'}\{\sum_x Q(x)l(E'(x)) : E' \text{ is a prefix code}\} + O(1),$$

that is, it realizes the least expected code-word length among all prefix codes (the entropy of $Q(\cdot)$ by Shannon's Noiseless Coding Theorem). Therefore, the H which minimizes Equation 5.16, that is,

$$l(E_{\Pr(\cdot|H)}(D)) + l(E_P(H))$$

minimizes the sum of two prefix codes which both have shortest *expected* code word lengths.

But there are many prefix codes which have expected code-word length equal to the entropy. Among those prefix codes there is one which gives the *shortest effective code word* to each individual source word: the prefix code with code word length $K(x)$ for object x. In ideal MDL we want minimize the sum of the effective description lengths of the *individual* elements H, D involved. This means using the *shortest effective description lengths*, as in Equation 5.18. However, we cannot simply replace negative logarithms in Equation 5.16 by corresponding $K(\cdot)$ terms. We can only do so if Equation 5.17 holds.

To satisfy Equation 5.17 we are free to make the new *assumption* that the prior probability $P()$ in Bayes's rule Equation 5.1 is fixed as $\mathbf{m}()$. Whether this can be justified or not is a question which we address in Section 5.5.7.

However, we *cannot assume* that the probability $\Pr(\cdot|H)$ equals $\mathbf{m}(\cdot|H)$. Namely, as explained at length in Section 5.1.3, probability $\Pr(\cdot|H)$ may be totally determined by the hypothesis H. Depending on H therefore, $l(E_{\Pr(\cdot|H)}(D))$ may be *very* different from $K(D|H)$. This holds especially for "simple" data D that have low probability under assumption of hypothesis H.

Example 5.5.2 Let us look at a simple example evidencing this discrepancy. Suppose we flip a coin of unknown bias n times. Let hypothesis H and data D be defined by:

$$\begin{aligned} H &:= \text{ Probability of "heads" is } \tfrac{1}{2}, \\ D &:= \underbrace{hh\ldots h}_{n \text{ times "}h\text{"(eads)}} \end{aligned}$$

Then we have $\Pr(D|H) = 1/2^n$ and

$$l(E_{\Pr(\cdot|H)}(D)) = -\log \Pr(D|H) = n.$$

In contrast, $K(D|H) \leq \log n + 2 \log\log n + O(1)$. ◇

5.5.3 Fundamental Inequality

The theory dealing with randomness of individual objects states that under certain conditions $-\log \Pr(D|H)$ and $K(D|H)$ are close. The first condition is that $\Pr(\cdot|\cdot)$ be a recursive function. That is, it can be computed to any required precision for each argument D and conditional H. Then we can appeal to the following known facts: Firstly, by Example 4.3.3 on page 249 following Theorem 4.3.1 on page 247,

$$\mathbf{m}(D|H) \geq 2^{-K(\Pr(\cdot|H))} \Pr(D|H).$$

Therefore,

$$\log \frac{\mathbf{m}(D|H)}{\Pr(D|H)} \geq -K(\Pr(\cdot|H)) \geq -K(H) + O(1). \tag{5.20}$$

The last inequality arises since from H we can compute $\Pr(\cdot|H)$ by assumption on $\Pr(\cdot|\cdot)$.

Secondly, if the data sample D is sufficiently *random* with respect to the recursive distribution $\Pr(\cdot|H)$ (with respect to H therefore) in the sense of Martin-Löf (and only if it is so), we have

$$\log \frac{\mathbf{m}(D|H)}{\Pr(D|H)} \leq 0, \tag{5.21}$$

where $\kappa_0((D|H)|\Pr(\cdot|H)) = \log(\mathbf{m}(D|H)/\Pr(D|H))$ is a "universal sum P-test" as in Theorem 4.3.5, page 258. The overwhelming majority of D's is random in this sense because for each H we have

$$\sum_D \Pr(D|H) 2^{\kappa_0((D|H)|\Pr(\cdot|H))} = \sum_D \mathbf{m}(D|H) \leq 1,$$

since $\mathbf{m}(\cdot|H)$ is a probability distribution. For D's that are random in the appropriate sense, Equations 5.20 and 5.21 mean by Equation 5.17 that

$$K(D|H) - K(\Pr(\cdot|H)) \leq -\log \Pr(D|H) \leq K(D|H).$$

Above, by assuming that the a priori probability $P(H)$ of hypothesis H is in fact the universal probability, we obtained $-\log P(H) = \mathbf{m}(x)$. However, we do not need to make this assumption. For recursive $P(\cdot)$, we can analyze the situation when H is random in the required sense with respect to $P(\cdot)$. The first inequality below holds since $\mathbf{m}(\cdot)$ majorizes $P(\cdot)$ this way; the second inequality expresses the assumption of randomness of H,

$$\mathbf{m}(H) \geq 2^{-K(P)} P(H),$$
$$\log(\mathbf{m}(H)/P(H)) \leq 0,$$

where $K(P)$ is the length of the shortest self-delimiting program for the reference universal prefix machine to simulate the Turing machine computing the probability density function $P : \mathcal{N} \to [0,1]$. That is, it is the shortest effective self-delimiting description of P. Then,

$$K(H) - K(P) \leq -\log P(H) \leq K(H). \tag{5.22}$$

Altogether, we find

$$\begin{aligned} K(D|H) + K(H) - \alpha(P,H) \leq & -\log \Pr(D|H) - \log P(H) \\ & \leq K(D|H) + K(H), \end{aligned} \tag{5.23}$$

with

$$\alpha(P,H) = K(\Pr(\cdot|H)) + K(P),$$

and we note that $K(\Pr(\cdot|H)) \leq K(H) + O(1)$. We call Equation 5.23 the **Fundamental Inequality (FI)** because it describes the fundamental relation between Bayes's Rule and MDL in mathematical terms. It is left for us to interpret it.

5.5.4 Validity Range of FI

We begin by stressing that Equation 5.23 holds only in case simultaneously H is P-random and D is $\Pr(\cdot|H)$-random. What is the meaning of this?

H is P-random means that the true hypothesis must be "typical" for the prior distribution P in the sense that it must belong to all effective majorities (sets on which the majority of P-probability is concentrated). In Example 4.3.10 on page 261 it is shown that this is the set of H's such that $K(H) \approx -\log P(H)$. In case $P(H) = \mathbf{m}(H)$, that is, the prior distribution equals the universal distribution, then for *all* H we have $K(H) = -\log P(H)$, that is, all hypotheses are random with respect to the universal distribution.

Let us look at an example of a distribution where some hypotheses are random, and some other hypotheses are nonrandom. Let the possible hypotheses correspond to the binary strings of length n, while P

is the uniform distribution that assigns probability $P(H) = 1/2^n$ to each hypothesis $H \in \{0,1\}^n$. Then $H = 00\ldots0$ has low complexity $K(H) \leq \log n + 2\log\log n$. However, $-\log P(H) = n$. Therefore, by Equation 5.22, H is not P-random. If we obtain H by n flips of a fair coin, then with overwhelming probability we will have that $K(H) = n + O(\log n)$, and therefore $-\log P(H) \approx K(H)$ and H is P-random.

D is $\Pr(\cdot|H)$-random means that the data are random with respect to the probability distribution $\Pr(\cdot|H)$ induced by the hypothesis H. Therefore, we require that the sample data D are "typical," that is, "randomly distributed" with respect to $\Pr(\cdot|H)$. If, for example, $H = (\mu, \sigma)$ induces the Gaussian distribution $\Pr(\cdot|(\mu,\sigma)) = N(\mu,\sigma)$ and the data D are concentrated in a tail of this distribution, like $D = 00\ldots0$, then D is atypical with respect to $\Pr(\cdot|H)$ in the sense of being nonrandom because it violates the $\Pr(\cdot|H)$-randomness test Equation 5.21.

Note that an hypothesis satisfying FI is a prefix complexity version of the Kolmogorov minimal sufficient statistic of Section 2.2.2. This shows the connections between MDL, Bayes's rule, and the Kolmogorov minimal sufficient statistic.

5.5.5 If Optimal Hypothesis Violates FI

The only way to violate the Fundamental Inequality is that either D is not $\Pr(\cdot|H)$-random and by Equation 5.21, $-\log \Pr(D|H) > K(D|H)$, or that H is not P-random and by Equation 5.22, $-\log P(H) > K(H)$. We give an example of the first case:

We sample a polynomial $H_2 = ax^2 + bx + c$ at n arguments chosen uniformly at random from the interval $[0,1]$. The sampling process introduces Gaussian errors in the function values obtained. The set of possible hypotheses is the set of polynomials. Assume that all numbers involved are of fixed bounded accuracy.

Because of the Gaussian error in the measuring process, with overwhelming probability the only polynomials H_{n-1} that fit the sample precisely are of degree $n-1$. Denote the data sample by D. Now, this hypothesis H_{n-1} is likely to minimize $K(D|H)$ since we just have to describe the n lengths of the intervals between the sample points along the graph of H_{n-1}. However, for this H_{n-1} data sample D is certainly not $\Pr(\cdot|H_{n-1})$-random, since it is extremely unlikely, and hence atypical, that D arises when sampling H_{n-1} with Gaussian error. Therefore, Equation 5.21 is violated, which means that $-\log \Pr(D|H_{n-1}) > K(D|H_{n-1})$, contrary to what we used in deriving the Fundamental Inequality. With prior probability $P(\cdot) := \mathbf{m}(\cdot)$, which means $-\log P(\cdot) = K(\cdot) + O(1)$, this moreover violates the Fundamental Inequality.

In contrast, with overwhelming likelihood H_2 will show the data sample D random to it. That being the case, the Fundamental Inequality holds.

Now what happens if H_{n-1} is the true hypothesis and the data sample D by chance is as above? In that case the Fundamental Inequality is violated and Bayes's Rule and MDL may each select very different hypotheses as the most likely ones, respectively.

5.5.6 If Optimal Hypothesis Satisfies FI

Given data sample D and prior probability P, we call a hypothesis H *admissible* if the Fundamental Inequality Equation 5.23 holds. Restriction to the set of *admissible* hypotheses excludes setting $K(D|H) + K(H) \approx K(D)$ for trivial hypothesis (like $H = \emptyset$ or $H = D$, which are not admissible).

Theorem 5.5.1 *Let $\alpha(P, H)$ as in Equation 5.23 be small. Then Bayes's Rule and ideal MDL are optimized (or almost optimized) by the same hypothesis among the admissible H's. That is, there is one admissible H that simultaneously (almost) minimizes both $-\log \Pr(D|H) - \log P(H)$ (selection according to Bayes's Rule) and $K(D|H) + K(H)$ (selection according to MDL).*

Proof. The smallness of $\alpha(P, H)$ means that both the prior distribution P is simple, and that the probability distribution $\Pr(\cdot|H)$ over the data samples induced by hypothesis H simple. In contrast, if $\alpha(P, H)$ is large, which means that either of the mentioned distributions is not simple, for example when $K(\Pr(\cdot|H)) = K(H)$ for complex H, then there may be some discrepancy. Namely, in Bayes's Rule our purpose is to maximize $\Pr(H|D)$, and the hypothesis H that minimizes $K(D|H) + K(H)$ also maximizes $\Pr(H|D)$ up to a $2^{-\alpha(P,H)}$ multiplicative factor. Conversely, the H that maximizes $\Pr(H|D)$ also minimizes $K(D|H) + K(H)$ up to an additive term $\alpha(P, H)$. That is, with

$$\begin{aligned} H_{\text{mdl}} &:= \text{minarg}_H \{K(D|H) + K(H)\}, \\ H_{\text{bayes}} &:= \text{maxarg}_H \{\Pr(H|D)\}, \end{aligned} \tag{5.24}$$

we have

$$\begin{aligned} 2^{-\alpha(P,H)} &\leq \frac{\Pr(H_{\text{mdl}}|D)}{\Pr(H_{\text{bayes}}|D)} \leq 1, \\ \alpha(P, H) &\geq K(D|H_{\text{mdl}}) + K(H_{\text{mdl}}) - K(D|H_{\text{bayes}}) - K(H_{\text{bayes}}) \geq 0. \end{aligned} \tag{5.25}$$

□

Therefore, if $\alpha(P, H)$ is small enough and Bayes's rule selects an admissible hypothesis, and so does ideal MDL, then both criteria are (almost) optimized by both selected hypotheses.

5.5.7 What MDL Does

We can now assess what prior distributions and assumptions about the relation between the data sample and selected hypothesis MDL assumes.

That is, how we can translate MDL in terms of Bayes's Rule. Identifying application of MDL with application of Bayes's rule on some prior distribution P, we must assume that given D, the Fundamental Inequality is satisfied for H_{mdl} as defined in Equation 5.24. This means that H_{mdl} is P-random for the prior distribution P used. One choice to guarantee this is to choose

$$P(\cdot) := \mathbf{m}(\cdot)(= 2^{-K(\cdot)}).$$

This is a valid choice even though $\mathbf{m}$ is not recursive, since the latter requirement arose from the requirement that $\mathbf{m}(\cdot)/P(\cdot)$ be enumerable, which is certainly guaranteed by choice of $P(\cdot) := \mathbf{m}(\cdot)$. This choice of prior distribution over the hypotheses is an objective and recursively invariant quantified form of Occam's razor: simple hypotheses H (with $K(H) \ll l(H)$) have high probability, and complex or random hypotheses H (with $K(H) \approx l(H)$) have low probability, namely, $2^{-l(H)}$.. This choice of prior distribution is most convenient, since the randomness test $\log \mathbf{m}(H)/P(H) = 0$ for *each* hypothesis H. This means that all hypotheses H are random with respect to distribution $\mathbf{m}(\cdot)$. It is easy to verify the following.

Theorem 5.5.2 *Let $\alpha(P,H)$ in de FI Equation 5.23 be small (for example $\alpha = O(1)$). With prior $P(\cdot)$ set to $\mathbf{m}(\cdot)$, the Fundamental Inequality Equation 5.23 is satisfied iff data sample D is $\Pr(\cdot|H_{\mathrm{mdl}})$-random.*

With the chosen prior and data sample D, the required $\Pr(\cdot|H_{\mathrm{mdl}})$-randomness of D constrains the domain of hypotheses from which we can choose H_{mdl}. Hence we can interprete ideal MDL as an application of Bayes's Rule with as prior distribution the universal distribution $\mathbf{m}(\cdot)$ and selection of a hypothesis H_{mdl} which shows the given data sample random to it in the precise sense of $\Pr(\cdot|H_{\mathrm{mdl}})$-randomness of individual objects as developed in Section 2.4.

Since the notion of individual randomness incorporates all effectively testable properties of randomness, application of ideal MDL will select the simplest hypothesis which balances the $K(D|H)$ and $K(H)$ and also shows the data sample D random (as far as we are able to tell) with respect to the selected hypothesis H_{mdl}.

This is the "reason" why the hypothesis selected by ideal MDL is not simply the one that perfectly fits the data. With some amount of overstatement on can say that if one obtains perfect data for a true hypothesis, then ideal MDL interprets these data as data obtained from a simpler hypothesis subject to measuring errors. Consequently, in this case ideal MDL is going to give you the *false simple* hypothesis and *not* the *complex true* hypothesis.

- Ideal MDL only gives us the true hypothesis if the data satisfy certain conditions relative to the true hypothesis. Stated differently: there are only data and no true hypothesis for ideal MDL. The principle simply obtains the hypothesis that is suggested by the data and it assumes that the data are random with respect to the hypothesis.

We have now provided an explanation why and when MDL works within the theory of algorithmic information that agrees with the analysis performed in the theory of probability, namely, that the MDL estimates of the data-generating models are consistent, except for data in small probability. And this is the essence of the requirement of the data being random, relative to the best model, in condition FI, Equation 5.23.

It is only to within terms of order O(1) that the MDL and the Bayesian techniques are equivalent. In modern forms of MDL one departs from the straight correspondence with Bayes's rule and takes $-\log[\mathbf{m}(D|\hat{H})/\sum_{D'}\mathbf{m}(D'|\hat{H})]$, instead of $-\log[\mathbf{m}(D|\hat{H})]$, where $\hat{H}(D)$ is the minimizing hypothesis and the summation runs through all data D' such that $\hat{H}(D')=\hat{H}(D)$, [J.J. Rissanen, *IEEE Trans. Inform. Theory*, IT-42:1(1996), 40–47]. The probability in the denominator gets absorbed by the term O(1), however, but for smaller amounts of data it does make a difference.

5.5.8 Applying Minimum Description Length

Unfortunately, the function $K(\cdot)$ of the hypotheses H is not computable (Section 3.4). For practical applications one must settle for easily computable approximations. One way to do this is as follows: First encode both H and $D|H$ by a simply computable bijection as a natural number in $\mathcal{N}$. Assume we have some standard procedure to do this. Then consider a simple self-delimiting description of x. For example, x is encoded by $x' = 1^{l(x)}0l(x)x$. This makes $l(x') = \log x + 2\log\log x + 1$, which is a simple upper approximation of $K(x)$; see Section 3.2. Since the length of code-word sets of prefix-codes corresponds to a probability distribution by Kraft's Inequality (page 74), this encoding corresponds to assigning probability $2^{-l(x')}$ to x. In the MDL approach, this is the specific usable approximation to the universal prior. In the literature we find a more precise approximation that, however, has no practical meaning. For convenience, we smooth our encoding as follows.

Definition 5.5.1 Let $x \in \mathcal{N}$. The *universal MDL prior* over the natural numbers is $M(x) = 2^{-\log x - 2\log\log x}$.

In the Bayesian interpretation the prior distribution expresses one's prior knowledge about the "true" value of the parameter. This interpretation may be questionable, since the used prior is usually not generated by repeated random experiments. In Rissanen's view, the parameter is generated by the selection of the class of hypotheses and it has no inherent meaning. It is just one means to describe the properties of the data. The selection of H that minimizes $K(H) + K(D|H)$ (or Rissanen's approximation thereof) allows one to make statements about the data. Since the complexity of the models plays an important part, the parameters must be encoded. To do so, we truncate them to a finite precision and encode them with the prefix-code above. Such a code happens to be equivalent to a distribution on the parameters. This may be called the universal MDL prior, but its genesis shows that it expresses no prior knowledge about the true value of the parameter. See [J.J. Rissanen, *Stochastic Complexity and Statistical Inquiry*, World Scientific, 1989]. Above we have given a validation of MDL from Bayes's Rule, which holds irrespective of the assumed prior, provided it is recursive and the hypotheses and data are random.

Example 5.5.3 In statistical applications, H is some statistical distribution (or model) $H = P(\theta)$ with a list of parameters $\theta = (\theta_1, \ldots, \theta_k)$, where the number k may vary and influence the (descriptional) complexity of θ. (For example, H can be a normal distribution $N(\mu, \sigma)$ described by $\theta = (\mu, \sigma)$.) Each parameter θ_i is truncated to finite precision and encoded with the prefix-code above.

The data sample consists of n outcomes $\mathbf{y} = (y_1, \ldots, x_n)$ of n trials $\mathbf{x} = (x_1, \ldots, x_n)$ for distribution $P(\theta)$. The data sample D in the above formulas is given as $D = (\mathbf{x}, \mathbf{y})$. By expansion of conditional probabilities we have therefore

$$\Pr(D|H) = \Pr(\mathbf{x}, \mathbf{y}|H) = \Pr(\mathbf{x}|H) \cdot \Pr(\mathbf{y}|H, \mathbf{x}).$$

In the argument above we take the negative logarithm of $\Pr(D|H)$, that is,

$$-\log \Pr(D|H) = -\log \Pr(\mathbf{x}|H) - \log \Pr(\mathbf{y}|H, \mathbf{x}).$$

Taking the negative logarithm in Bayes's rule and the analysis of the previous section now yields that MDL selects the hypothesis with highest inferred probability satisfying $\mathbf{x}$ is $\Pr(\cdot|H)$-random and $\mathbf{y}$ is $\Pr(\cdot|H, \mathbf{x})$-random. Thus, Bayesian reasoning selects the same hypothesis as MDL does, provided the hypothesis with maximal inferred probability causes $\mathbf{x}, \mathbf{y}$ to satisfy these randomness requirements.

Under certain general conditions, J.J. Rissanen has shown that with k parameters and n data (for large n) Equation 5.16 is minimized for hypotheses H with θ encoded by

$$-\log P(H) = \frac{k}{2} \log n$$

bits. This is called the *optimum model cost* since it represents the cost of the hypothesis description at the minimum description length of the total.

As an example, consider a Bernoulli process $(p, 1-p)$ with p close to $\frac{1}{2}$. For such processes $k = 1$. Let the outcome be $x = x_1 x_2 \ldots x_n$. Set $f_x = \sum_{i=1}^n x_i$. For outcome x with $C(x) \geq n - \delta(n)$, the number of 1's will be (by Lemma 2.3 on page 159)

$$f_x = n/2 \pm \sqrt{\frac{3}{2}(\delta(n) + c)n/\log e}.$$

With $\delta(n) = \log n$, the fraction of such x's in $\{0,1\}^n$ is at least $1 - 1/n$ and goes to 1 as n rises unboundedly. Hence, for the overwhelming number of x's the frequency of 1's will be within

$$2^{-\frac{1}{2}\log n}$$

of the value $\frac{1}{2}$. That is, to express an estimate to parameter p with high probability it suffices to use a precision of $\frac{1}{2}\log n$ bits. It is easy to generalize this example to arbitrary p. ◇

Example 5.5.4 In biological modeling, we often wish to fit a polynomial f of unknown degree to a set of data points

$$D = (x_1, y_1), \ldots, (x_n, y_n),$$

such that it can predict future data y given x. Even if the data did come from a polynomial curve of degree, say, two, because of measurement errors and noise, we still cannot find a polynomial of degree two fitting all n points exactly. In general, the higher the degree of fitting polynomial, the greater the precision of the fit. For n data points, a polynomial of degree $n-1$ can be made to fit exactly, but probably has no predicting value. Applying ideal MDL we look for $H_{\mathrm{mdl}} := \mathrm{minarg}_H\{K(\mathbf{x}, \mathbf{y}|H) + K(H)\}$.

Let us apply the ideal MDL principle where we describe all $(k-1)$-degree polynomials by a vector of k entries, each entry with a precision of d bits. Then the entire polynomial is described by

$$kd + O(\log kd) \text{ bits.} \tag{5.26}$$

(We have to describe k, d, and account for self-delimiting encoding of the separate items.) For example, $ax^2 + bx + c$ is described by (a, b, c) and can be encoded by about $3d$ bits. Each datapoint (x_i, y_i) that needs to be encoded separately with precision of d bits per coordinate costs about $2d$ bits.

For simplicity assume that probability $\Pr(\mathbf{x}|H) = 1$ (because $\mathbf{x}$ is prescribed). To apply the ideal MDL principle we must trade the cost of hypothesis H (Equation 5.26) against the cost of describing $\mathbf{y}$ with help of H and $\mathbf{x}$. As a trivial example, suppose that $n-1$ out of n datapoints fit a polynomial of degree 2 exactly, but only 2 points lie on any polynomial of degree 1 (a straight line). Of course, there is a polynomial of degree $n-1$ that fits the data precisely (up to precision). Then the Ideal MDL cost is $3d + 2d$ for the 2nd degree polynomial, $2d + (n-2)d$ for the 1st degree polynomial, and nd for the $(n-1)$th degree polynomial. Given the choice among those three options, we select the 2nd degree polynomial for all $n > 5$.

A more sophisticated approach, accounting for the average encoding cost of exceptions, assumes that the data are Gaussian distributed. Consider polynomials f of degree at most $n-1$ that minimize the error

$$\text{error}(f) = \sum_{i=1}^{n} (f(x_i) - y_i)^2. \tag{5.27}$$

This way we find an optimal set of polynomials for each $k = 1, 2, \ldots, n$. To apply the MDL principle we must trade the cost of hypothesis H (Equation 5.26) against the cost of describing $D|H$.

To describe measuring errors (noise) in data it is common practice to use the normal distribution. In our case this means that each y_i is the outcome of an independent random variable distributed according to the normal distribution with mean $f(x)$ and variance, say, constant. For each of them we have that the probability of obtaining a measurement y_i, given that $f(x)$ is the true value, is of the order of $\exp(-(f(x) - y_i)^2)$. Considering this as a value of the universal MDL probability, this is encoded in $s(f(x)-y_i)^2$ bits, where s is a (computable) scaling constant. For all experiments together we find that the total encoding of $D|f, \mathbf{x}$ takes $s \cdot \text{error}(f)$ bits. The MDL principle thus tells us to choose a k-degree function f_k, $k \in \{0, \ldots, n-1\}$, that minimizes (ignoring the vanishing $O(\log kd)$ term) $kd + s \cdot \text{error}(f_k)$. ◇

Example 5.5.5 In this example we apply the MDL principle to infer decision trees. We are given a set of data, possibly with noise, representing a collection of examples. Each example is represented by a data item in the data set, which consists of a tuple of *attributes* followed by a binary *Class* value indicating whether the example with these attributes is a positive or negative example.

Figure 5.3 gives a small sample set. The columns in the table describe attributes that are weather conditions. The rows are examples that represent weather conditions in relation to some "unspecified occurrences." The last column classifies the examples as positive or negative, where "P" means that it happened and "N" means that it did not happen. We would like to obtain good predictions for such occurrences by compressing the data. Our task can now be explained as a communication

No.	Attributes				Class
	Outlook	*Temperature*	*Humidity*	*Windy*	
1	overcast	hot	high	not	N
2	overcast	hot	high	very	N
3	overcast	hot	high	medium	N
4	sunny	hot	high	not	P
5	sunny	hot	high	medium	P
6	rain	mild	high	not	N
7	rain	mild	high	medium	N
8	rain	hot	normal	not	P
9	rain	cool	normal	medium	N
10	rain	hot	normal	very	N
11	sunny	cool	normal	very	P
12	sunny	cool	normal	medium	P
13	overcast	mild	high	not	N
14	overcast	mild	high	medium	N
15	overcast	cool	normal	not	P
16	overcast	cool	normal	medium	P
17	rain	mild	normal	not	N
18	rain	mild	normal	medium	N
19	overcast	mild	normal	medium	P
20	overcast	mild	normal	very	P
21	sunny	mild	high	very	P
22	sunny	mild	high	medium	P
23	sunny	hot	normal	not	P
24	rain	mild	high	very	N

FIGURE 5.3. Sample data set

problem between Alice, who observed the data in Figure 5.3, and Bob. Alice and Bob both know the four parameters (outlook, temperature, humidity, windy) and their attributes. Alice wishes to send Bob the information in the table, using as few bits as possible. Alice and Bob have to agree in advance on an encoding technique to be used.

Alice and Bob do not know in advance which table they have to transmit. The simplest strategy for Alice is to transmit the complete table in Figure 5.3 to Bob literally. There are 24 rows. Each row has four attributes and one Class value. Three attributes have three alternative values each; the other attribute and Class have two alternative values each. Then this requires $24(3 \log_2 3 + 2) = 24(3 \times 1.585 + 2) \approx 162.12$ bits. Or Alice can agree with Bob beforehand about a fixed order of enumerating all $3 \times 3 \times 2 \times 3 = 54$ combinations of attributes, and then just send the last column of 54 bits, supplying arbitrary Class values for the 30 rows missing from the table in Figure 5.3. These methods use no data compression.

If Alice is clever enough to find some regularity in the data, like "the Class value is 'N' iff it rains," then Alice needs only a few bits to transmit this sentence to Bob, and Bob can use this rule to reconstruct the complete table with all the combinations of attributes with $3 \times 3 \times 2 \times 3 = 54$ rows.

Let us say that Alice and Bob have agreed to use a decision tree. A decision tree that is consistent with the data can be viewed as a classification procedure. The internal nodes in the tree are *decision nodes.* Each such node specifies a test of a particular attribute; the possible answers are labeled on the arcs leaving the decision node. A leaf of the tree specifies a class to which the object that passed the attribute tests and arrived at this leaf belongs. Given data as in Figure 5.3, we can construct many different decision trees of various sizes that agree with the data. Two such trees are given in Figures 5.4 and 5.5. The tree in Figure 5.4 is imperfect since it makes an error on row 8; the tree in Figure 5.5 classifies all of the sample set correctly. The tree in Figure 5.5 is the smallest perfect decision tree for the data in Figure 5.3.

Some data, for example noise, may not obey the predicting rule defined by the decision tree. One usually has a choice between using a small *imperfect* tree that classifies some data falsely or a big *perfect* tree that correctly classifies all given data. Alice can use a smaller imperfect tree or the bigger perfect tree. The tree in Figure 5.5 grows much bigger just because of a single (perhaps noisy) example (row 8), and Alice may find that it is more economical to code it separately, as an *exception.*

The goal is often to construct a decision tree that has the smallest error rate for classifying unknown future data. Is the *smallest perfect decision tree* really a good predictor? It turns out that in *practice* this is not the case. Due to the presence of noise or inadequacy of the given attributes, selecting a perfect decision tree "overfits" the data and gives generally poor predictions. Many ad hoc rules have been suggested and used for overcoming this problem.

The MDL principle appears to provide a solution and generally works well in practice. Essentially, given the data sample *without* the class val-

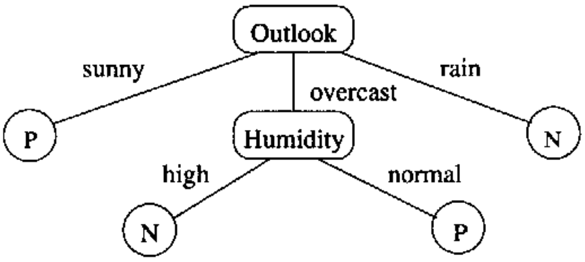

FIGURE 5.4. Imperfect decision tree

ues, we look for a reasonably small tree such that most data are correctly classified by the tree. We encode the inconsistent data as exceptions. We minimize the sum of

- the number of bits to encode a (not necessarily perfect) decision tree T, that is, the model that gives the $-\log P(H)$ term; and
- the number of bits to encode (describe) the examples (x_i, y_i) in the sample $(\mathbf{x}, \mathbf{y})$ that are inconsistent with T, given the entire data sample $\mathbf{x}$ without the class values $\mathbf{y}$. This gives the $-\log P(\mathbf{y}|H, \mathbf{x})$ term.

We have to provide a coding method. This is important in applications, since it determines where the optimum is. If the encoding of trees is not efficient, then we may end up with a very small tree (with relatively large depth), and too many examples become exceptions. An inefficient encoding of the exceptions would result in overly large trees. In both cases, the prediction of unseen data is affected. The reader should realize that the choice of cost measure and encoding technique cannot be objective. One can encode a tree by making a depth-first traversal. At each internal node, write down the attribute name in some self-delimiting form followed by its edge label. At a leaf write down the class value. If the tree is not perfect, then the data that do not fit in the tree are encoded separately as exceptions (in some economical way using the provided total data sample without the class values).

Coding the Tree It is desirable that the smaller trees be represented by shorter encodings. Alice can make a depth-first traversal of the tree in Figure 5.4, and accordingly she writes down

1 Outlook 0 P 1 Humidity 0 N 0 P 0 N.

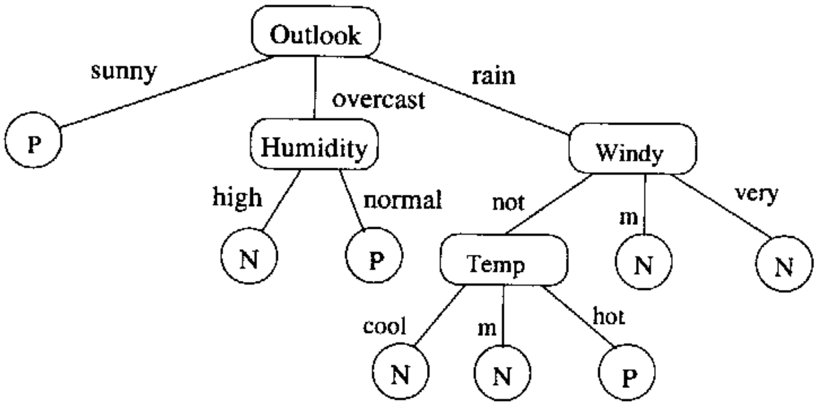

FIGURE 5.5. Perfect decision tree

For the tree in Figure 5.5, she writes down

1 Outlook 0 P 1 Humidity 0 N 0 P 1 Windy 0 N 0 N
1 Temperature 0 N 0 N 1 P.

Alice uses a "1" to indicate that the next node in the depth-first search is an internal node and then writes the corresponding attribute; Alice writes a 0 followed by "N" ("P") if she meets a leaf with value "N" ("P"). Representing "N" or "P" requires only one bit. Representing attribute "Outlook" at the root level requires 2 bits since there are 4 possibilities. Encoding the next-level attributes requires log 3 bits since there are only 3 choices left ("Outlook" is already used). She can use just one bit for "Windy" and one bit for "Temperature" (in fact, this one bit is not needed). Thus, the smaller (but imperfect) tree requires 13.585 bits; the bigger (but perfect) tree requires 25.585 bits.

Coding the Exceptions Since the decision tree in Figure 5.4 is not perfect, we need to indicate where the exceptions are. In this case there is a single exception. The most straightforward way is to indicate its *position* among all 54 possible combinations of attributes. This costs $\log 54 \approx 5.75$ extra bits.

Thus, the encoding using the decision tree in Figure 5.4 uses 19.335 bits; the encoding using the decision tree in Figure 5.5 uses 25.585 bits. The MDL principle tells us to use the former method, which is also much shorter than the 54-bit plain encoding.

Procedures for computing decision trees have been implemented by J.R. Quinlan and R. Rivest [*Inform. Computation*, 80(1989), 227–248]. Computing the absolute minimum decision tree is NP-complete, as shown by T. Hancock, T. Jiang, M. Li, and J. Tromp, *Inform. Comput.*, 126:2(1996), 114–122. They have shown that approximating minimum decision trees is also NP-hard, even approximation to within some polynomial factor. Consequently, approximation heuristics have to be used. See also K. Yamanishi, A Randomized Approximation of the MDL for Stochastic Models with Hidden Variables, Proc. 9th ACM Comput. Learning Conference, ACM Press, 1996; and V. Vovk, Learning about the parameter of the Bernoulli Model, *J. Comput. System Sci.*, to appear.

◇

Example 5.5.6 **(Alternative MDL-Like Principle)** In the above interpretation of MDL we essentially look for a hypothesis H minimizing $K(D|H)+K(H)$. This always satisfies

$$K(D|H) + K(H) \geq K(D).$$

An incorrect interpretation of the way we used MDL in Example 5.5.5 on page 364 is sometimes confused with MDL. In the new approach the idea is that we define $E := D - D_H$, where D_H is the data set classified according to H. We want to minimize

$$K(H, E) \approx K(H) + K(E|H)$$

over H. That is, E denotes the subset of the data sample D that are *exceptions* to H in the sense of being "not covered" by H. We want to find H such that the description of H and the exception data E *not covered* by H are together minimized. Note that in this case always

$$\min_H \{K(H) + K(E|H)\} \leq K(\emptyset) + K(D) = K(D),$$

in contrast to the standard interpretation of MDL above. This incarnation of MDL is not straightforwardly derived by our approach above. We may interpret it that we look for the shortest description of an accepting program for the data consisting of a classification rule H and an exception list E. While this principle sometimes gives good results, application may lead to absurdity as the following shows:

In many problems our data sample D consists of only positive examples, as when we want to learn (a grammar for) the English language given a corpus of data D like the *Oxford Dictionary*. Then according to our new MDL rule the best hypothesis is the trivial grammar H generating *all* sentences over the alphabet. Namely, this grammar gives $K(H) = O(1)$ independent of D and also $E := \emptyset$. Consequently,

$$\min_H \{K(H) + K(E|H)\} = K(H) = O(1),$$

which is absurd. The principle is vindicated and reduces to the standard one in the context of interpreting $D = (\mathbf{x}, \mathbf{y})$ as in Example 5.5.3 on page 362, with $\mathbf{x}$ fixed as in "supervised learning." This is a correct application as in Example 5.5.5 on page 364. We want to find H minimizing

$$K(H) + K(\mathbf{y}|H, \mathbf{x}) + K(\mathbf{x}|H),$$

which is the same as minimizing

$$K(H) + K(\mathbf{y}|H, \mathbf{x}),$$

provided we take $K(\mathbf{x}|H)$ constant. Now, $K(\mathbf{y}|H, \mathbf{x})$ corresponds to $K(E|H)$ if we ignore the constant $\mathbf{x}$ in the conditional. ◇

Example 5.5.7 **(Maximum Likelihood)** The *maximum likelihood* (ML) principle says that for given data D, one should select the hypothesis H that maximizes

$P(D|H)$ or equivalently, minimizes $-\log P(D|H)$. In case of finitely many hypotheses, this is a special case of the MDL principle with the hypotheses distributed uniformly (all have equal probability). The principle has many admirers, is supposedly objective, and is due to R.A. Fisher. ◇

Example 5.5.8 **(Maximum Entropy)** In statistics there are a number of important applications where the ML principle fails, but where the maximum entropy principle has been successful, and conversely.

In order to apply Bayes's Rule, we need to decide what the prior probabilities $p_i = P(H_i)$ are, subject to the constraint $\sum_i p_i = 1$ and certain other constraints provided by empirical data or considerations of symmetry, probabilistic laws, and so on. Usually these constraints are not sufficient to determine the p_i's uniquely. E.T. Jaynes proposed to select the prior by the *maximum entropy* (ME) principle.

The ME principle selects the estimated values $\hat{p}_i$ that maximize the entropy function

$$H(p_1, \ldots, p_k) = -\sum_{i=1}^{k} p_i \ln p_i, \tag{5.28}$$

subject to

$$\sum_{i=1}^{k} p_i = 1 \tag{5.29}$$

and some other constraints. For example, consider a loaded die, $k = 6$. If we do not have any information about the die, then using the principle of indifference, we may assume that $p_i = \frac{1}{6}$ for $i = 1, \ldots, 6$. This actually coincides with the ME principle, since $H(p_1, \ldots, p_6) = -\sum_{i=1}^{6} p_i \ln p_i$, constrained by Equation 5.29, achieves its maximum $\ln 6 = 1.7917595$ for $p_i = \frac{1}{6}$ for all i.

Now suppose it has been experimentally observed that the die is biased and the average throw gives 4.5, that is,

$$\sum_{i=1}^{6} i p_i = 4.5. \tag{5.30}$$

Maximizing the expression in Equation 5.28, subject to the constraints in Equations 5.29 and 5.30, gives the estimates

$$\hat{p}_i = e^{-\lambda i} (\sum_j e^{-\lambda j})^{-1}, \quad i = 1, \ldots, 6,$$

where $\lambda = -0.37105$. Hence,

$$(\hat{p}_1, \ldots, \hat{p}_6) = (0.0543, 0.0788, 0, 1142, 0.1654, 0.2398, 0.3475).$$

The maximized entropy $H(\hat{p}_1, \ldots, \hat{p}_6)$ equals 1.61358. How dependable is the ME principle? Jaynes has proven an "entropy concentration theorem" that, for example, implies the following: In an experiment of $n = 1000$ trials, 99.99% of all 6^{1000} possible outcomes satisfying the constraints of Equations 5.30 and 5.29 have entropy

$$1.602 \leq H\left(\frac{n_1}{n}, \ldots, \frac{n_6}{n}\right) \leq 1.614,$$

where n_i is the number of times the value i occurs in the experiment. We show that the Maximum Entropy principle can be considered as a special case of the MDL principle, as follows:

Consider the same type of problem. Let $\theta = (p_1, \ldots, p_k)$ be the prior probability distribution of a random variable. We perform a sequence of n independent trials. Shannon has observed that the real substance of Formula 5.28 is that we need approximately $nH(\theta)$ bits to record the sequence of n outcomes. Namely, it suffices to state that each outcome appeared $n_1, \ldots, n_k$ times, respectively, and afterwards give the index of which one of the

$$\binom{n}{n_1, \ldots, n_k} = \frac{n!}{n_1! \cdots n_k!} \tag{5.31}$$

possible sequences D of n outcomes actually took place. For this no more than

$$k \log n + \log \binom{n}{n_1, \ldots, n_k} + O(\log \log n) \tag{5.32}$$

bits are needed. The first term corresponds to $-\log P(\theta)$, the second term corresponds to $-\log P(D|\theta)$, and the third term represents the cost of encoding separators between the individual items. Using Stirling's approximation of $n! \sim \sqrt{2\pi n}(n/e)^n$ for the quantity of Equation 5.31, we find that for large n, Equation 5.32 is approximately

$$n\left(-\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}\right) = nH\left(\frac{n_1}{n}, \ldots, \frac{n_k}{n}\right).$$

Since k and n are fixed, the least upper bound on the minimum description length, for an arbitrary sequence of n outcomes under certain given constraints 5.29 and 5.30, is found by maximizing the term in Equation 5.31 subject to said constraints. This is equivalent to maximizing the entropy function 5.28 under the constraints.

Viewed differently, let S_θ be the set of outcomes with values $(n_1, \ldots, n_k)$, with $n_i = np_i$, corresponding to a distribution $\theta = (p_1, \ldots, p_k)$. Then due to the small number of values (k) in θ relative to the size of the sets, we have

$$\log \sum_\theta d(S_\theta) \approx \max_\theta \{\log d(S_\theta)\}. \tag{5.33}$$

The left-hand side of Equation 5.33 is is the minimum description; the right-hand side of Equation 5.33 is the maximum entropy. ◇

5.6 History and References

The material on Epicurus can be found in E. Asmis [*Epicurus Scientific Method*, Cornell University Press, 1984]. The elegant paper "The use of simplicity in induction," by J.G. Kemeny [*Phil. Rev.*, 62(1953), 391–408] contains predecessors to the ideas formulated in this chapter. Bayes's formula originates from Thomas Bayes's "An essay towards solving a problem in the doctrine of chances" [*Phil. Trans. Roy. Soc.* 25 (1763) 376–398. (*Ibid.* , 54(1764) 298–310, R. Price (Ed.))] posthumously published by his friend Richard Price. Properly speaking, Bayes's Rule as given in the text is not due to Bayes. P.S. Laplace stated Bayes's Rule in its proper form and attached Bayes's name to it in *A philosophical essay on probabilities* (1819). In his original memoir, Bayes assumes the uniform distribution for the prior probability and derives $P(H_i|D) = P(D|H_i)/\sum_i P(D|H_i)$. This formula can be derived from Bayes's Rule in its present form by setting all $P(H_i)$ equal. Bayes did not state the result in its general form, nor did he derive it through a formula similar to Bayes's Rule. The books by B. de Finetti [*Probability, Induction, and Statistics,* John Wiley & Sons, 1972], I.J. Good [*Good Thinking,* University of Minnesota Press, 1983], P.S. Laplace [*Ibid.*], R. von Mises [*Probability, Statistics and Truth,* Macmillan, 1939], and T.L. Fine [*Theories of Probability,* Academic Press, 1973] contain excellent discussions on the Bayesian and non-Bayesian views of inductive reasoning.

The idea of using Kolmogorov complexity in inductive inference, in the form of using a universal prior probability, is due to R.J. Solomonoff [*Inform. Contr.*, 7(1964), 1–22, 224–254]. Solomonoff's original definition of prior probability is problematic through the use of the C-version of the Kolmogorov complexity instead of the prefix complexity (as used here). Inductive inference, using $\mathbf{M}$ as universal prior, is due to R.J. Solomonoff [*IEEE Trans. Inform. Theory,* IT-24(1978), 422–432]; see also [T.M. Cover, 'Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin,' Tech. Rept. 12, 1974, Statistics Dept,