Proof: Let y be a square root of $-x$ modulo P and $r \neq y, -y$. Suppose $(r + \sqrt{-x})^Q = u + v\sqrt{-x}$ with one of u, v equal to 0. Then $((r + \sqrt{-x})/(r - \sqrt{-x}))^Q$ is 1 or -1, and so is $((r + y)/(r - y))^Q$. Hence $(r + y)/(r - y)$ is a $2Q$th root of unity in $\mathbf{Z}_P^*$ not equal to 1 or -1. It follows, using Lemma 1, that the algorithm fails for at most $2Q$ values of r. Thus the probability of failure at each iteration is $2Q/(P - 1) = 1/2^{s-1}$.

ACKNOWLEDGEMENTS

This research has greatly benefited from conversations with Eric Bach, Elwin Berlekamp, Manuel Blum, Joan Boyar, Gilles Brassard, Sampath Kannan, and René Schoof.

REFERENCES

[1] L. Adleman, K. Manders, and G. Miller, "On taking roots in finite fields," presented at 18th IEEE Annual Symp. Foundations of Computer Science, Providence, RI, 1977.

[2] Eric Bach, Ph.D. Thesis, University of California at Berkeley, 1984.

[3] E. R. Berlekamp, "Factoring polynomials over large finite fields," *Math. Comput.*, vol. 24, no. 111, p. 713, July 1970.

[4] D. H. Lehmer, "Computer Technology Applied to the Theory of Numbers," in *Studies in Number Theory*, W. J. LeVeque, Ed. p. 117, Englewood Cliffs, N.J.: MAA, Prentice Hall, 1969.

[5] M. Rabin, "Probabilistic algorithms in finite fields," *Siam J. Comput.*, vol. 9, pp. 273–280, 1980.

[6] D. Shanks, "Five Number-Theoretic Algorithms," in *Proc. Second Manitoba Conf. Numerical Mathematics*, 1972.

Maximum Entropy as a Special Case of the Minimum Description Length Criterion

MEIR FEDER

***Abstract*—The Maximum Entropy (ME) and Maximum Likelihood (ML) criteria are the bases for two approaches to statistical inference problems. A new criterion, called the Minimum Description Length (MDL), has been recently introduced. This criterion generalizes the ML method so it can be applied to more general situations, e.g., when the number of parameters is unknown. It is shown that ME is also a special case of the MDL criterion; maximizing the entropy subject to some constraints on the underlying probability function is identical to minimizing the code length required to represent all possible i.i.d. realizations of the random variable such that the sample frequencies (or histogram) satisfy those given constraints.**

I. INTRODUCTION

A. Maximum Entropy and Maximum Likelihood

The Maximum Entropy (ME) and the Maximum Likelihood (ML) methods flow from two different philosophies for statistical inference. In both methods the "output," or the result of the inference process, is a choice of probability function which we believe (by those philosophies) represents best the behavior of the phenomena that we observe. To be more specific, let us describe the common situations leading to the choice of these methods.

Maximum Likelihood: Suppose we have an observation x. We assume that the probability function that describes x is characterized by some unknown parameter vector $\theta \in \Theta$; i.e., the probability function $p(\cdot)$ belongs to the set P_Θ, where

$$P_\Theta = \{ p(\cdot) | p(\cdot) = p(\cdot; \theta), \theta \in \Theta \}.$$

The ML criterion will choose $\hat{p}$ from P_Θ (which is equivalent to choosing θ in Θ) by

$$\hat{p} = \arg\max_{p \in P_\Theta} \log p(x) \left(\text{or } \hat{\theta} = \arg\max_{\theta \in \Theta} \log p(x; \theta) \right).$$

The basic limitation of the ML method is the need for modeling assumptions; i.e., we have to assume that we know the probability function up to a fixed number (usually much smaller than the length of the observations) of unknown parameters. Without those restrictions the ML method will break down; for example, if we allow any probability function, maximizing the likelihood will lead to the trivial (but unacceptable) result $p(\alpha) = \delta(\alpha - x)$, where $\delta(\cdot)$ is the Dirac delta function. As another example, if the number of the parameters is not fixed, then the ML will not work; the more parameters we choose, the larger the likelihood can be. This example has led (as will be seen later) to the introduction of the Minimum Description Length (MDL) criterion.

Maximum Entropy: Suppose we know that $p(\cdot)$ belongs to a set P, where this set is defined by the knowledge of some averages,

$$P = \{ p(x) | E_p[g(x)] = \bar{g} \}.$$

The given averages are the only information available on the underlying phenomena in the ME framework. The choice of probability function is then

$$\hat{p} = \arg\max_{p \in P} H(p) = \arg\max_{p \in P} \left[-\int_x p(x) \log p(x)\, dx \right].$$

We first note that in the ME framework we do not assume any model; the method will work even if the set P contains all possible probability functions. The constraints on the probability functions are derived from the data. On the other hand, the basic limitation of the direct ME method is that one cannot incorporate the information provided by a specific observation sequence. Usually in this case one calculates some sample averages and uses them as constraints on the entropy maximization. However, this approach does not use all available information about the phenomena, and we introduce errors in the inference process since the sample averages differ from the statistical averages.

B. Minimum Description Length Method

A new method for statistical inference, based upon the MDL criterion, has recently been introduced by Rissanen [1], [2], [3]. This method overcomes some of the limitations of the ML method. Stated simply, it is as follows.

Suppose we have a specific observation sequence x. The probability function that we will choose is such that the code length required to represent those observations (which is a function of the probability that we assign to the observations) is minimized.

This description (or code) length is influenced by two factors. If we know the probability distribution, the "ideal" code length [4] that is required to represent the specific observation is the (self-) information of that observation, i.e., $-\log p(x)$. (In case $p(x)$ is density, a term proportional to the precision ought to be added. We drop this term, however.) The second factor is the code length needed to represent the model or the parameters.

As was noted by Rissanen [2], the MDL criterion reduces to the ML criterion when the number of the parameters is fixed; this follows from the code length needed to represent the model being then fixed too.

Manuscript received August 8, 1985; revised February 24, 1986. This work was supported in part by the Advanced Research Projects Agency monitored by ONR Contract N00014-81-K-0742 NR-049-506, in part by NSF Grant ECS-8407285 and in part by Woods Hole Oceanographic Institution.

The author is with the Department of Electrical Engineering and Computer Science, Research Laboratory of Electronics, Rm. 36-605, Massachusetts Institute of Technology, Cambridge, MA 02139.

0018-9448/86/1100-0847$01.00 ©1986 IEEE

Rissanen, inspired by the work of Akaike [5] and the algorithmic notion of complexity, has suggested the MDL criterion be used to generalize the ML procedure to the case where the number of the parameters is also unknown. The intuitive notion is that since we can always increase the likelihood by allowing more parameters, there should be some term in the criterion that will penalize using too many parameters (overfitting). The minimum code-length criterion is adequate since now the description length needed to define the parameters depends on their number and on the precision with which they are written. This code length was calculated in [2]. If we take only the dominant term, the MDL criterion suggested by Rissanen [3] for this case yields the parameter estimates

$$\hat{n}, \hat{\theta} = \arg\min_{n,\theta} \left[-\log p(x; \theta) + \frac{1}{2} n \log N \right] = \arg\min_{n,\theta} I_\theta(x),$$

where n is the number of the parameters in θ and N is the length of the observation sequence. $I_\theta(x)$ is defined to be the (self-) information of the sequence x with respect to the given family [3]; see also [6] and [7] for related work in the context of universal coding.

The main result in this paper is that the ME criterion is also a special case of this new MDL criterion. To show this result we somewhat extend the MDL method. Suppose the given information about the underlying phenomena is not a single observation sequence but rather a set of such sequences. This type of information is available either by having several independent observation sequences or by having constraints that define a possible set of observation sequences. The MDL criterion for this type of information will suggest that we choose the probability distribution that minimizes the total code length when all the members of this set of possible observation sequences are encoded using the proposed distribution.

Now we can adapt the MDL criterion to the ME framework in which the given information about the underlying phenomena is in terms of constraints on the probability distribution. We claim that if we try to represent all possible observation sequences whose "histogram" or sample frequencies satisfy those constraints, the minimum code length is achieved if the probability function is the ME distribution.

II. Maximum Entropy as Minimum Description Length

In the ME framework the given information is the knowledge of some averages. Recall that the strong law of large numbers implies

$$E[g(x)] = \bar{g} \Leftrightarrow \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(x_i) = \bar{g} \text{ a.s.},$$

where x_i are i.i.d. observations, distributed as x. So this ME-type information is equivalent to the information that the observations lie in the set of all infinitely long sequences whose sample averages are the given averages.

Having the above argument in mind, we will show that maximizing the entropy subject to some constraints on the probability distribution is asymptotically equivalent to minimizing the code length needed to represent all the sequences whose histogram or sample frequencies satisfy the given constraints.

To clarify our argument let us start with a simple example. Suppose we want to estimate the probability of 1 (success) in a simple binary (Bernoulli) trial. We denote $p(1) = \theta$, $p(0) = 1 - \theta$. Suppose we do not have any observations and thus we only know that for any N trials we will perform we may observe any of the 2^N possible sequences of 1's and 0's.

Equipped with the MDL philosophy, we will choose θ so that all 2^N sequences can be represented by the shortest possible code. Now for each sequence x we need about

$$-\log p(x; \theta) = -\log\left[\theta^k (1-\theta)^{N-k}\right] \tag{1}$$

bits, where k is the number of 1's in x. We will denote by $\mathcal{X}$ the set of all such sequences and by $L(\mathcal{X})$ the total code length required to represent the whole set. Now for any k we have $\binom{N}{k}$ sequences with k 1's, so that

$$\begin{aligned} L(\mathcal{X}) &= \sum_{k=0}^{N} \binom{N}{k} \left[-\log \theta^k (1-\theta)^{N-k} \right] \\ &= -\left[\sum_{k=0}^{N} \binom{N}{k} k \right] \log \theta - \left[\sum_{k=0}^{N} \binom{N}{k} (N-k) \right] \log(1-\theta). \end{aligned} \tag{2}$$

Noting that

$$\sum_{k=0}^{N} \binom{N}{k} k = \sum_{k=0}^{N} \binom{N}{k} (N-k) = N 2^{N-1} \equiv \alpha,$$

we see that

$$L(\mathcal{X}) = -\alpha \log \theta - \alpha \log(1-\theta) \tag{3}$$

is minimized (unsurprisingly) by $\theta = 1/2$. Observe that this probability function is the same as that given by the ME principle (with no constraints) on the binary random variable.

Further note that we have ignored the term $1/2 \log N$ required to represent the code length for describing the single parameter θ because it has no effect on the minimizing value of θ.

We are ready now to prove the general claim, stated as the following theorem.

Theorem: Let X be a random variable that takes its values over the finite set $\{1, \cdots, m\}$. Let $x = x_1 x_2 \cdots x_N$ be a sample of N independent trials of X. Let $f_i(x) = k_i(x)/N$ be the frequencies of each outcome in this sample. (The vector $f(x) = [f_1(x), \cdots, f_m(x)]^T$ will be sometimes called the histogram of the sample). Let $\mathcal{F}$ be any fixed set of histograms. Let $\mathcal{X}_N$ be the set

$$\mathcal{X}_N = \{ x = x_1 \cdots x_N | f(x) \in \mathcal{F} \}.$$

The code length that results when the whole set $\mathcal{X}_N$ is encoded depends on $p = [p_1, \cdots, p_m]$, the probability assignment of X with which the code design is done, and it will be denoted by $L(\mathcal{X}_N, p)$. Then the probability that minimizes this length is given by

$$\hat{p}_N = \frac{\sum_{f \in \mathcal{F}} \frac{N!}{k_1! \cdots k_m!} f}{\sum_{f \in \mathcal{F}} \frac{N!}{k_1! \cdots k_m!}}. \tag{4}$$

Furthermore, if the entropy function H has a unique maximum in $\mathcal{F}$, then

$$\hat{p} = \lim_{N\to\infty} \hat{p}_N = \arg\max_{p \in \mathcal{F}} H(p) = \arg\max_{p \in \mathcal{F}} - \sum_{i=1}^{m} p_i \log p_i. \tag{5}$$

In other words, the "best" (in the MDL sense) probability is the ME probability subject to the given constraints on the histogram.

Proof: The probability of a sample $x_1 \cdots x_N$ depends on the relative frequencies (or the number of occurrences of each outcome) as

$$p(x) = \prod_{i=1}^{m} p_i^{k_i}.$$

So the code length required to represent this sample to within the term $m/2 \log N$ required to encode the probabilities p_i is

$$L(x) = -\log p(x) = - \sum_{i=1}^{m} k_i \log p_i. \tag{6}$$

Now there are $N!/k_1!\cdots k_m!$ possible sequences having the same frequencies or the same number of occurrences $\boldsymbol{k} = [k_1,\cdots,k_m]^T$. Since the constraints are only on the frequencies (or on $\boldsymbol{k}$) we can write the total code length

$$L(\mathcal{X}_N,\boldsymbol{p}) = \sum_{\text{admissible } \boldsymbol{k}} \frac{N!}{k_1!\cdots k_m!}\left(-\sum_{i=1}^{m} k_i \log p_i\right)$$
$$= -\sum_{i=1}^{m}\left(\sum_{f\in\mathcal{F}} \frac{N!}{k_1!\cdots k_m!} k_i\right)\log p_i. \tag{7}$$

We will denote

$$\beta_i = \sum_{f\in\mathcal{F}} \frac{N!}{k_1!\cdots k_m!} k_i. \tag{8}$$

The total code length is thus

$$L(\mathcal{X}_N,\boldsymbol{p}) = -\sum_{i-1}^{m} \beta_i \log p_i,$$

which is minimized (using Jensen's inequality) by

$$\hat{p}_{N,i} = \frac{\beta_i}{\sum_{l=1}^{m} \beta_l}. \tag{9}$$

Substituting (8) in (9) and recalling that $k_i/\sum_{l=1}^{m} k_l = f_i$ yields (4).

Again, we observe that if the set $\mathcal{F}$ includes all the possible distributions, we can conclude by symmetry that all β_i are equal; thus $\hat{p}_{N,i} = 1/m$; i.e., we get the uniform distribution, which is the ME distribution for this case of no constraints.

In general, we will get the ME distribution only in the limit as $N \to \infty$ as follows. It is easy to show (using Stirling's formula for factorials) that

$$S = \frac{N!}{k_1!\cdots k_m!} = e^{N[H(f_1,\cdots,f_m)+o(\log N/N)]}, \tag{10}$$

where $H(f_1,\cdots,f_m) = -\sum_{i=1}^{m} f_i \log f_i$ is the entropy associated with the frequencies $f_i = k_i/N$. Substituting (10) in (4) and taking the limit as $N \to \infty$ yields

$$\hat{\boldsymbol{p}} = \lim_{N\to\infty} \frac{\sum_{f\in\mathcal{F}} f e^{NH(f)}}{\sum_{f\in\mathcal{F}} e^{NH(f)}}. \tag{11}$$

Let us assume first that the function $H(\boldsymbol{f})$ has in $\mathcal{F}$ a single global maximum, at $\boldsymbol{f}_{\max}$. Now as $N \to \infty$

$$\lim_{N\to\infty} e^{-N[H(f_{\max})-H(f)]} = \begin{cases} 0 & \text{if } \boldsymbol{f} \neq \boldsymbol{f}_{\max} \\ 1 & \text{if } \boldsymbol{f} = \boldsymbol{f}_{\max} \end{cases}. \tag{12}$$

So we can write (11) as

$$\lim_{N\to\infty} \frac{\sum_{f\in\mathcal{F}} f e^{NH(f)}}{\sum_{f\in\mathcal{F}} e^{NH(f)}} = \lim_{N\to\infty} \frac{\sum_{f\in\mathcal{F}} f e^{-N[H(f_{\max})-H(f)]}}{\sum_{f\in\mathcal{F}} e^{-N[H(f_{\max})-H(f)]}} = \boldsymbol{f}_{\max}, \tag{13}$$

i.e., $\hat{\boldsymbol{p}}$ is the ME distribution. Q.E.D.

Note that in general if $H(\cdot)$ has several global maxima $\boldsymbol{f}_1,\cdots,\boldsymbol{f}_n$, then the result in (11) will be

$$\hat{\boldsymbol{p}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{f}_i.$$

We claim that the above theorem can be extended, following the same lines of proof, to the case where the random variable takes its values over a countably infinite set.

III. Conclusion

The desired outcome of this paper is to establish the MDL criterion as a general criterion for statistical inference. The relation to ME together with the motivation introduced by Rissanen [3] makes this method powerful and adequate for many statistical inference problems.

The advantage of using the MDL will be clear in situations where both the ML and the ME methods fail. Having the MDL criterion, we can, on the one hand, work with more general sets of probability distributions (not just distributions that are known up to a fixed number of parameters as in the ML method) and, on the other hand, take into consideration more general data (not just a specific sequence as in the ML method or some averages as in the ME method).

An example that we have in mind is the following. Suppose we observe an output of a system that implies that the system state is in a set $\mathcal{X}$ of states. We can construct a probability assignment for the state of the system by looking for a probability measure that will minimize the code length required to represent all the states in $\mathcal{X}$. This implementation, which cannot be solved by the ML since we have no *a priori* parameterized probability distributions in mind, and which cannot be solved by the ME method since we do not have direct constraints on the probability distribution, might be solved using the MDL criterion, and it is now under investigation.

Acknowledgment

The author would like to thank the associate editor Prof. T. L. Fine and the reviewers for their suggestions that helped to clarify this paper. The author also thanks Mati Wax, Bruce Musicus, Udi Weinstein, Taly Tishbi, and Amir Dembo for interesting discussions on this topic.

References

[1] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.
[2] ——, "A universal prior for integers and estimation by MDL," *Ann. Statist.*, vol. 11, no. 2, pp. 416–432, 1983.
[3] ——, "Universal coding, information, prediction and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 4, pp. 629–636, July 1984.
[4] J. Rissanen and G. G. Langdon, "Universal modeling and coding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 12–23, 1981.
[5] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.
[6] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.
[7] ——, "Minimax noiseless universal coding for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 211–215, 1983.

Synchronization of Binary Source Codes

BRUCE L. MONTGOMERY AND JULIA ABRAHAMS

***Abstract*—The problem of achieving synchronization for variable-length source codes is addressed through the use of self-synchronizing binary prefix-condition codes. Although our codes are suboptimal in the sense of**

Manuscript received January 14, 1985; revised October 16, 1985. This work was supported in part by NSF Grant ECS-8411623. This paper was presented at the IEEE International Symposium on Information Theory, Brighton, U.K., June 1985.

The authors are with the Department of Electrical and Computer Engineering, Carnegie-Mellon University, Pittsburgh, PA 15213.

IEEE Log Number 8609706.