

# Strong Optimality of the Normalized ML Models as Universal Codes and Information in Data

J. Rissanen

IBM Research Division

Almaden Research Center, DPE-B2/802

San Jose, CA 95120-6099, rissanen@almaden.ibm.com

1/3/2000

## Abstract

We show that the normalized maximum likelihood (*NML*) distribution as a universal code for a parametric class of models is closest to the negative logarithm of the maximized likelihood in the mean code length distance, where the mean is taken with respect to the worst case model inside or outside the parametric class. We strengthen this result by showing that, when the data generating models are restricted to be the most ‘benevolent’ ones in that they incorporate all the constraints in the data and no more, the bound cannot be beaten in essence by any code except when the mean is taken with respect to the data generating models in a set of vanishing size.

These results allow us to decompose the code of the data into two parts, the first having all the useful information in the data that can be extracted with the family in question and the rest which has none, and we obtain a measure for the (useful) information in data.

**Index Terms** - *MDL*-principle, minmax bounds, noise, relative redundancy, stochastic complexity, useful information

## 1 Introduction

In this paper the central issue examined is to show that the normalized maximum likelihood (*NML*) distribution of a parametric model class incorporates, intuitively

speaking, all the information in the data that can be extracted with the models in the class. We do this by considering two cases, first when the data have constraints not captured by the models in the class, and then when they have no other constraints than those expressible in terms of such models. We formalize the first case by considering the data as having been generated by a distribution outside the parametric model class. We show that the *NML* distribution solves the following generalization of a minmax problem in [3]

$$\inf_q \sup_g E_g \log \frac{f(X^n; \hat{\theta}(X^n))}{q(X^n)}, \quad (1)$$

where  $q$  and  $g$  range over the set of virtually all distributions,  $\hat{\theta}(x^n)$  is the maximum likelihood estimate of the parameters in the model class  $\{f(x^n; \theta)\}$  considered, and  $x^n = x_1, \dots, x_n$  denotes a data sequence. The same distribution solves also Shtarkov's minmax problem, [14],

$$\min_q \max_{x^n} \log \frac{f(x^n; \hat{\theta}(x^n))}{q(x^n)}. \quad (2)$$

A similar generalization of the ordinary minmax redundancy problem to the minmax *relative* redundancy problem

$$\min_q \max_g E_g \log \frac{f(X^n; \theta_g)}{q(X^n)} \quad (3)$$

was studied in [15], where the expectation is taken with respect to a distribution outside the model class satisfying certain conditions, and  $f(x^n; \theta_g)$  is the model in the class that is nearest to  $g$  in the Kullback-Leibler distance. The solution is given asymptotically by a modified Jeffreys' mixture. We can see that the two problems differ on the nature of the ideal target: In the second it is the ideal code length  $\log 1/f(x^n; \theta_g)$  of the unknown nearest model in the class to the data generating distribution  $g$ , while in the first it is the unreachable lower bound  $\log 1/f(x^n; \hat{\theta}(x^n))$  of the ideal code lengths.

We then consider the case where the parametric model class does capture all the constraints in the data. To formalize this we restrict the data generating models to certain most 'benevolent' ones. Even with such a restriction we show that the minmax bound is the same as when there are no restrictions on the data generating

models. Further, we strengthen this by showing that for any ideal code, identified with a density function, the mean code length, the mean taken with respect to the most benevolent distributions, cannot be significantly smaller than in the worst case, except for a vanishing subset of them.

These results have important implications beyond universal coding. The negative logarithm of the *NML* density function, which in [12] was defined to be the stochastic complexity, breaks up into two parts: The first is the code length of the optimal model as one of the set of (asymptotically) distinguishable models from the data, and the second is the code length of equivalent data sequences that have no further information about the optimal model. Here we are taking advantage of the idea of ‘distinguishable models’, which has been introduced in [1] and [2]. We are then justified to define the first part as the amount of (useful) *information* in the data that can be extracted with the model class at hand. Despite the overused notions of ‘information’ and ‘complexity’ it seems to be desirable to formally separate the two main constituents of the data, the latter playing the central role in data compression and the former in statistical modeling.

## 2 Two MinMax Problems

Consider a class of parametric density functions  $\mathcal{M}_\gamma = \{f(x^n; \theta, \gamma) : \theta \in \Omega\}$  defined on sequences  $x^n = x_1, \dots, x_n$  of real numbers, where  $\gamma$  is a structure index, such as the pair  $(p, q)$  of autoregressive and moving average orders in ARMA models, and the parameters  $\theta = \theta_1, \dots, \theta_k$ ,  $k$  depending on  $\gamma$ , range over a subset  $\Omega$  of the  $k$ -dimensional Euclidean space. In data compression problems the data are often just characters from a finite alphabet, and the models are probability mass functions  $P(x^n; \theta, \gamma)$ . In either case it is convenient to identify the negative logarithms  $-\log f(x^n; \theta, \gamma)$  and  $-\log P(x^n; \theta, \gamma)$  with *ideal* code lengths and the density functions and probability mass functions themselves with ideal codes. The basis of the logarithms is irrelevant, but for the sake of definiteness we use the natural logarithm.

It is clear that the models in the class  $\mathcal{M}_\gamma$  cannot express all the statistical properties in real world data, no matter how large  $n$  is. One way to mathematically model such a case is to let the data be generated with a distribution  $g(x^n)$  lying outside

the class such that it will give the data statistical properties different from those expressible with the models in the class. Consider the following problem

$$\inf_q \sup_{g \in G} E_g \ln \frac{f(X^n; \hat{\theta}(X^n), \gamma)}{q(X^n)}, \quad (4)$$

where  $\hat{\theta}(x^n)$  is the maximum likelihood estimate of the parameters in the class  $\mathcal{M}_\gamma$  for fixed  $\gamma$ ,  $q(x^n)$  is a distribution; i.e. an ideal code, and  $G$  is a class of all distributions such that  $E_g \ln(g(X^n)/f(X^n; \hat{\theta}(X^n), \gamma)) < \infty$ . This excludes the singular distributions, which do not specify properties in data that we wish to learn.

**Theorem 1** *Let  $\Omega$  in  $\mathcal{M}_\gamma$  be such that the integral*

$$C_n(\gamma) = \int_{\hat{\theta}(y^n) \in \Omega} f(y^n; \hat{\theta}(y^n), \gamma) dy^n \quad (5)$$

*is finite. The solution to the problem (4) is the universal NML model*

$$\hat{f}(x^n; \gamma) = \frac{f(x^n; \hat{\theta}(x^n), \gamma)}{C_n(\gamma)}, \quad (6)$$

*and clearly,*

$$E_g \ln \frac{f(X^n; \hat{\theta}(X^n), \gamma)}{\hat{f}(X^n; \gamma)} = \ln C_n(\gamma) \quad (7)$$

*for all  $g \in G$ .*

**Proof:** We have

$$\inf_q \sup_{g \in G} E_g \ln \frac{f(X^n; \hat{\theta}(X^n), \gamma)}{q(X^n)} \geq \sup_{g \in G} \inf_q E_g \ln \frac{f(X^n; \hat{\theta}(X^n), \gamma)}{q(X^n)}, \quad (8)$$

and

$$\begin{aligned} \sup_{g \in G} \inf_q E_g \ln \frac{f(X^n; \hat{\theta}(X^n), \gamma)}{q(X^n)} &\geq \inf_q D(g \| q) - D(g \| \hat{f}) + \ln C_n(\gamma) = \\ &= -D(g \| \hat{f}) + \ln C_n(\gamma). \end{aligned}$$

The last equality results with the choice  $q = g$ . The choice  $g = \hat{f}$  achieves the maximum  $\ln C_n(\gamma)$  of the right hand side, which equals the expectation in the left hand side of (8). The claims follow.

Notice that  $\hat{f}$  involves only the models in the class  $\mathcal{M}_\gamma$ , so that it can be computed and data encoded in the best manner with the models available in the class for the worst case data generating distribution. The set of parameters  $\Omega$  forms an essential part of the model class, and it affects the minmax bound  $\ln C_n(\gamma)$ . Its selection to insure the finiteness of the integral in (5) may involve additional hyperparameters. For gaussian models in the linear regression problems, however, these can be determined so that they will not affect the *MDL* criterion, which seeks to maximize  $\hat{f}(x^n; \gamma)$  over  $\gamma$ , [13].

Under certain conditions, Theorem 1 of [12]

$$\ln C_n(\gamma) = \frac{k}{2} \ln \frac{n}{2\pi} + \ln \int_{\Omega} \sqrt{|I(\theta)|} d\theta + o(1), \quad (9)$$

where  $I(\theta)$  is the Fisher information matrix

$$I(\theta) = \left\{ -E \frac{\partial^2 \ln f(X; \theta, \gamma)}{\partial \theta_i \partial \theta_j} \right\}.$$

A closely related formula for the minmax mean redundancy of Jeffreys' mixture for iid (independent identically distributed) processes was derived in [4].

The worst case bound  $\ln C_n(\gamma)$  clearly depends on the parametric class at hand, which leaves room for lots of codes, whose mean length is shorter than that obtainable with the *NML* model restricted to this class. Take for instance a family of nested subclasses  $\mathcal{M} = \bigcup_k \mathcal{M}_k$ , such as the set of polynomials of all degrees  $k$ ,  $\gamma = k$ , for curve fitting, with the normal distribution for the deviations. If we put  $g = q = \hat{f}(x^n; k)$  for  $k > m$ , then

$$E_g \ln \frac{f(X^n; \hat{\theta}(X^n), m)}{\hat{f}(X^n; k)} < \ln C_n(m).$$

This, of course, is possible because we are using codes with more powerful models than those in the model class  $\mathcal{M}_m$ .

Consider next the larger class of models  $\mathcal{M} = \bigcup_{\gamma \in \Gamma} \mathcal{M}_\gamma$ , where  $\Gamma$  is a set of the structure indexes such as  $k$  ranging from 0 to  $n$  in the preceding polynomial example. We assume  $\Gamma$  to be finite as often is the case. Consider the second minmax problem

$$\inf_q \sup_{g \in G} E_g \ln \frac{\hat{f}(X^n; \hat{\gamma}(X^n))}{q(X^n)}, \quad (10)$$

where  $\hat{\gamma}(x^n)$  is the maximum likelihood estimate of the structure index of the universals  $\hat{f}(X^n; \gamma)$  for  $\gamma \in \Gamma$ . With the same arguments as in the proof of Theorem 1 the solution to (10) is given by the *NML* model for  $\mathcal{M}$

$$\hat{f}(x^n) = \frac{\hat{f}(x^n; \hat{\gamma}(x^n))}{\mathcal{C}_n}, \quad (11)$$

where

$$\mathcal{C}_n = \sum_{\gamma \in \Gamma} \int_{\hat{\gamma}(y^n) = \gamma} \hat{f}(y^n; \hat{\gamma}(y^n)) dy^n, \quad (12)$$

and the minmax value itself is obtained for  $g = \hat{f}$  as  $\ln \mathcal{C}_n$ .

Because  $\mathcal{C}_n$  does not affect the particular  $\hat{\gamma}(x^n)$  obtained from the observed data  $x^n$  we see that so long as we are searching for the optimal universal model within the class  $\Gamma$  it can be found with the *MDL* principle

$$\min_{\gamma \in \Gamma} \ln 1/\hat{f}(x^n; \gamma) \quad (13)$$

without need to compute  $\mathcal{C}_n$  nor to add a code length for the minimizing  $\hat{\gamma}$ ; an application to denoising is in [13].

To conclude this section we mention that the solution to (4) provides another way to analyze the minmax relative redundancy, [15],

$$R_n(\gamma) = \min_q \max_{g \in \mathcal{G}} E_g \ln \frac{f(X^n; \theta_g, \gamma)}{q(X^n)}, \quad (14)$$

where  $f(X^n; \theta_g, \gamma)$  minimizes the KL distance

$$\min_{\theta \in \Omega} D(g(X^n) \| f(X^n; \theta, \gamma)). \quad (15)$$

The minimizing parameter  $\theta_g$  is assumed to exist and to be unique. We have

$$\min_q \max_{\theta} E_{\theta} \ln \frac{f(X^n; \theta, \gamma)}{q(X^n)} \leq R_n(\gamma) \leq \ln \mathcal{C}_n(\gamma) + \max_{g \in \mathcal{G}} E_g \ln \frac{f(X^n; \theta_g, \gamma)}{f(X^n; \hat{\theta}(X^n), \gamma)}, \quad (16)$$

where  $E_{\theta}$  is the expectation with respect to the model  $f(X^n; \theta, \gamma)$  in the class  $\mathcal{M}_{\gamma}$ . Since for iid processes we know the left hand side to be below  $\ln \mathcal{C}_n(\gamma)$  in (9) by  $k/2$ , [4], the determination of the asymptotic behavior of the relative redundancy

is reduced to the examination of conditions under which the last term is uniformly bounded both in  $\theta_g$  and  $n$ . If it, in fact, can be shown to converge to  $-k/2$ , the *NML* model would achieve asymptotically even the minmax relative redundancy. Notice that for iid processes both  $\hat{\theta}(x^n)$  and  $(-1/n)\sum_t \ln f(x_t; \hat{\theta}(x^n), \gamma)$  converge to  $\theta_g$  and  $-E_g \ln f(X; \theta_g, \gamma)$ , respectively, in  $g$ -probability one, [6] and [16].

### 3 Strong Lower Bound

In this section we study optimal code performance when all the statistical constraints in the data are captured by the class  $\mathcal{M}_\gamma$  of models. This has been done traditionally with the assumption that some model in the class generates the data with the mean redundancy as the performance measure, first in terms of the minmax mean redundancy, [5], and later in the stronger sense that only rarely can the minmax bound be beaten by any code, [9] and [8]. Since our performance measure is neither the mean redundancy nor the mean relative redundancy, we supplement these results with a different analysis.

Let  $\Omega$  be a compact set of parameters. We discuss the case where the data range over the real line so that the set of the maximum likelihood estimates  $\hat{\theta}(x^n)$  is the entire set  $\Omega$ ; the discrete case, where this is not true, is handled with obvious modifications. Let

$$X_{\hat{\theta}} = \{y^n : \hat{\theta}(y^n) = \hat{\theta}\}. \quad (17)$$

We have the factorizations

$$\begin{aligned} f(x^n; \theta) &= f(x^n, \hat{\theta}(x^n); \theta) = f(x^n | \hat{\theta}(x^n); \theta) p(\hat{\theta}(x^n); \theta) \\ f(x^n; \hat{\theta}(x^n)) &= f(x^n | \hat{\theta}(x^n)) p(\hat{\theta}(x^n)), \end{aligned} \quad (18)$$

where  $f(y^n | \hat{\theta}(x^n); \theta)$  is the density function with  $X_{\hat{\theta}(x^n)}$  as its support induced by  $f(y^n; \theta)$  and conditioned on  $\hat{\theta}(x^n)$ , and  $p(\hat{\theta}(x^n); \theta)$  is the marginal density function of the statistic  $\hat{\theta}(x^n)$ . We also dropped the repeated argument in  $f(x^n | \hat{\theta}(x^n); \hat{\theta}(x^n))$  and  $p(\hat{\theta}(x^n); \hat{\theta}(x^n))$  for simplicity. In these and the subsequent formulas we also omit the structure index  $\gamma$ , because it is fixed, and show it only when needed.

We mention in passing that for the exponential families the conditional density

function  $f(x^n|\hat{\theta})$  is uniform in its support, because, [11],

$$-\ln f(x^n; \hat{\theta}(x^n)) = -\ln f(x^n|\hat{\theta}(x^n)) - \ln p(\hat{\theta}(x^n)) = H(\hat{\theta}(x^n)), \quad (19)$$

where  $H(\hat{\theta}(x^n))$  is the empirical entropy, obtained by replacing  $\theta$  in the entropy function  $H(\theta)$  of  $X^n$  by the ML estimate. Since the right hand side and  $p(\hat{\theta}(x^n))$  depend on the data strings only through the maximum likelihood estimate the claim holds. Even for other model classes, whenever the asymptotic equipartition property holds, the conditional density function is approximately uniform over nearly all of the support for large  $n$ .

**Example:** In the Bernoulli class the factorization (18) is given by

$$P(x^n; \theta) = \frac{1}{\binom{n}{m}} \binom{n}{m} \theta^m (1 - \theta)^{n-m},$$

where  $\theta$  is the probability of symbol 1. The first factor is clearly a uniform probability function in the set  $X_{m/n}$  of all strings of length  $n$  with  $m$  1's. Replacing  $\theta$  by  $\hat{\theta}(x^n) = m/n$  we get  $H(m/n)$  as  $nh(m/n)$ , where  $h(p)$  is the binary entropy function, and (19) is seen to hold.

When the code performance is measured by the mean redundancy the ideal target for each parameter value  $\theta$  is  $\ln 1/f(x^n; \theta)$ , and the most 'benevolent' data generating distribution would of course be the distribution  $f(x^n; \theta)$  itself, which minimizes the Kullback-Leibler distance  $D(g||f(X^n; \theta))$ . In our formalism for each maximum likelihood estimate  $\hat{\theta}_0$  the ideal code with which the data sequence can be encoded is

$$f(y^n|\hat{\theta}_0) = \begin{cases} \frac{f(x^n; \hat{\theta}_0)}{p(\hat{\theta}_0)} & \text{for } x^n \in X_{\hat{\theta}_0} \\ 0 & \text{otherwise.} \end{cases} \quad (20)$$

For this code the most 'benevolent' data generating distribution is again the code itself  $g(x^n; \hat{\theta}_0) = f(y^n|\hat{\theta}_0)$ , with the ideal mean code length given by the entropy of the distribution. Indeed, given  $\hat{\theta}_0 = \hat{\theta}(x^n)$ , the remaining uncertainty about the string  $x^n$  is expressed optimally within the model class considered by  $f(y^n|\hat{\theta}_0)$ . This distribution appears also to capture the intent that it incorporates all the constraints in the data  $x^n$  that can be expressed in terms of the model class considered and no more.

**Example:** For the Bernoulli class take  $\hat{\theta}_0 = m/n$ . Then

$$g(x^n; m/n) = \frac{1}{\binom{n}{m}}.$$

More generally, for exponential families the minimizing density function is uniform in its support  $X_{\hat{\theta}_0}$  as we showed above.

We now have a parametric class of models  $\hat{G}_\gamma = \{g(x^n; \hat{\theta}) : \hat{\theta} \in \Omega\}$ , and we may consider the minmax problem

$$\min_q \max_{g \in \hat{G}_\gamma} E_g \ln \frac{f(X^n; \hat{\theta}(X^n), \gamma)}{q(X^n)} = \ln \hat{C}_n(\gamma), \quad (21)$$

where we reintroduced the structure index  $\gamma$ . We have the theorem

**Theorem 2** *The minimizing  $q$  in (21) is given by the NML density function*

$$\hat{q}(x^n) = \hat{f}(x^n; \gamma) = \frac{f(x^n; \hat{\theta}(x^n), \gamma)}{C_n(\gamma)}, \quad (22)$$

and the minmax bounds (21) and (5) are equal:

$$\hat{C}_n(\gamma) = C_n(\gamma) = \int_{\hat{\theta}(y^n) \in \Omega} f(y^n; \hat{\theta}(y^n), \gamma) dy^n = \int_{\Omega} p(\theta) d\theta. \quad (23)$$

Finally, the prior

$$\bar{w}(\hat{\theta}) = \frac{p(\hat{\theta})}{\int_{\Omega} p(\theta) d\theta} \quad (24)$$

maximizes the mean

$$\int_{\Omega} w(\hat{\theta}) E_g \ln \frac{f(X^n; \hat{\theta}(X^n), \gamma)}{g(X^n; \hat{\theta}) w(\hat{\theta})} d\hat{\theta} \quad (25)$$

where  $w(\hat{\theta})$  ranges over all priors on  $\Omega$ .

We use the term ‘prior’ for these distributions following tradition even though they need not have anything to do with prior knowledge in the Bayesians’ sense. In particular,  $\bar{w}$  could be called *canonical prior*, because it is constructed entirely in terms of the data and cannot incorporate any prior knowledge whatsoever.

**Proof:** Let  $Q(\hat{\theta})$  be the probability of  $X_{\hat{\theta}}$  under the density function  $q(x^n)$ . The minmax problem (21) is equivalent with

$$\min_q \max_{\hat{\theta}} \{D(f(X^n|\hat{\theta}, \gamma) \| q(X^n)/Q(\hat{\theta})) + \ln \frac{\bar{w}(\hat{\theta})}{Q(\hat{\theta})} + \ln C_n(\gamma)\}. \quad (26)$$

This is closely related to Shtarkov's minmax problem, (2), and the solution is found by similar arguments: The first term within the curly brackets is nonnegative for all  $q(x^n)$ , and so is the second for the maximizing  $\hat{\theta}$ , because at that point the ratio of the two distributions is not smaller than unity. Both terms vanish for the choice (22), which proves the first two claims.

To prove the last claim observe that (25) is  $\ln C_n(\gamma) - D(w \| \bar{w})$ , which reaches its maximum value  $\ln C_n(\gamma)$  for  $w = \bar{w}$ .

The next theorem generalizes the strong lower bound in [9] as well as those in [8] in that the expectation is taken with respect to density functions outside of the model class  $\mathcal{M}_\gamma$ . In [8] the proof relies heavily on the fact that the minmax bound is the channel capacity for the model class  $\mathcal{M}_\gamma$  in question, which requires the maximizing mixture density. Here, this is not the case nor can we use any mixture density at all, but because  $\hat{f}(x^n; \gamma) = f(x^n|\hat{\theta}(x^n), \gamma)\bar{w}(\hat{\theta})$  behaves like a mixture maximizing (25), which plays the role of channel capacity, the quite ingenious arguments in the cited reference still apply with modifications.

**Theorem 3** *Let the parameters of the model class  $\mathcal{M}_\gamma$  range over a subset  $\Omega$  of the euclidean space  $R^k$  such that  $p(\theta)$ , (18), satisfies*

$$0 < K < \frac{\inf_{\theta \in \Omega} p(\theta)}{\sup_{\theta \in \Omega} p(\theta)} \quad (27)$$

*for some positive constant  $K$ . If  $C_n(\gamma) \rightarrow \infty$  as  $n \rightarrow \infty$ , then for any  $q(x^n)$*

$$E_{g(\cdot; \theta)} \ln \frac{f(X^n; \hat{\theta}(X^n), \gamma)}{q(X^n)} > (1 - \epsilon) \ln C_n(\gamma) \quad (28)$$

*for every positive  $\epsilon$  and all  $\theta \in \Omega - B_n(q, \epsilon)$ , where the volume of the exceptional set  $B_n(q, \epsilon)$  goes to zero as  $n$  grows to infinity.*

The proof is given in Appendix A.

**Discussion.** Both of the assumptions in the theorem are satisfied if the maximum likelihood estimates  $\hat{\theta}(x^n)$  of the parameters in  $\mathcal{M}_\gamma$  at each interior point of  $\Omega$  satisfy the central limit theorem, and  $\Omega$  is such that the Fisher information is both bounded and bounded away from zero. In fact, then  $p(\hat{\theta})$  behaves like  $n^{k/2}\sqrt{|I(\hat{\theta})|}/(2\pi)^{n/2}$ , and (27) follows. To see that  $\ln C_n(\gamma) \rightarrow \infty$ , apply Theorem 1 in [9] to the model class  $\mathcal{M}_k$ , which gives

$$E_{f(\cdot, \theta)} \ln \frac{f(X^n; \theta, \gamma)}{\hat{f}(X^n; \gamma)} \geq (1 - \epsilon) \frac{k}{2} \ln n$$

for all  $\theta$  except in a set whose volume goes to zero as  $n \rightarrow \infty$ . This immediately implies that

$$\ln C_n(\gamma) \geq (1/2 - \epsilon) \frac{k}{2} \ln n + E_{f(\cdot, \theta, \gamma)} \ln \frac{f(X^n; \hat{\theta}(X^n), \gamma)}{f(X^n; \theta, \gamma)} \geq (1 - \epsilon) \frac{k}{2} \ln n,$$

which shows that  $\ln C_n(\gamma) \rightarrow \infty$ . We chose not to assume the central limit theorem, because the assumptions made suffice; for instance, the second condition is shown to hold for exponential families in [7] without the assumption of the central limit theorem.

Theorem 1 in the previous section dealt with the case where the data have regular features that are not captured by the models in the class  $\mathcal{M}_\gamma$ , and we found the *NML* density function among all density functions as ideal codes to minimize the mean excess code length, the mean taken with respect to the worst case model of almost any kind. Theorem 2, then, shows that the same minmax bound will result even when the data generating models are restricted to the most ‘benevolent’ ones which was our way to model the idea that the data have no other regular features than those that can be expressed in terms of the models in the class  $\mathcal{M}_\gamma$ . In other words, regardless of what other constraints the data might have, the *NML* density function guarantees the same worst case bound, which can be reduced only by designing codes with more powerful models than those in the class chosen. Finally, in the case where the parametric model class does capture all the constraints in the data, the third theorem strengthens the second by stating that the worst case data generating model is not a rare event: Nearly all of the ‘benevolent’ data generating models will be

almost as difficult to encode against no matter how the coding is done.

In conclusion, we have shown that it is not possible to encode data with a shorter mean code length than that obtained with the *NML* model unless we permit more elaborate models to be used in the code designs than those in the class  $\mathcal{M}_\gamma$ , which, in turn, are subject to their own lower bound.

## 4 Complexity and Information

For the model class  $\mathcal{M}_\gamma$  consider the decomposition

$$-\ln \hat{f}(x^n; \gamma) = -\ln f(x^n; \hat{\theta}(x^n), \gamma) + \ln C_n(\gamma). \quad (29)$$

We have defined  $-\ln \hat{f}(x^n; \gamma)$  in [12] to be the *stochastic complexity* of the data sequence as the shortest code length, relative to the model class considered, the idea of ‘shortest’ to be taken in a probabilistic sense. This has been amply justified in the present paper. By contrast, the interpretation of the two terms is not entirely clear, other than the first representing an ideal target and the second the logarithm of the necessary normalizing factor. A new and intuitively most pleasing interpretation of the term  $\ln C_n(\gamma)$  as well as the stochastic complexity itself has been obtained in [1] and [2] by differential geometric arguments without any appeal to code length. We sketch the process in a somewhat different way and establish the link with the coding theoretic interpretation of the previous sections to suit our purposes.

The idea is to consider the manifold of models  $f(y^n; \theta, \gamma)$ ,  $\theta$  ranging over  $\Omega$ , with a special metric induced by the Fisher information matrix  $I(\theta)$ . Consider the hyper ellipsoid  $(\theta - \theta_i)'I(\theta_i)(\theta - \theta_i) \leq d$  centered at  $\theta_i$ . Let  $\theta_i$  run through a finite set such that the largest curvilinear rectangles within these ellipsoids partition the parameter space by a nonuniform grid. Further, let  $d$  shrink as a function of  $n$  at such a rate that the volumes of the rectangles become  $V_i(n) = |I(\theta_i)|^{-1/2}(\frac{2\pi}{n})^{k/2}$  to within an asymptotically ignorable error due to the fact that the sides of the rectangles are not exactly straight lines. (If they were, the rectangles would not partition the space, because the adjacent ones would overlap slightly). This shrinking rate is critical in the sense that the models induced by the center points of two adjacent rectangles can be distinguished from an increasing sequence of data in such a manner that the

probability of making a mistake goes to zero as  $n$  goes to infinity, while for a faster shrinking rate the error probability does not go to zero. The startling outcome is that the logarithm of the number of models that *can* be so distinguished from data sequences  $x^n$  is given by the formula (9) for  $\ln C_n(\gamma)$ . In [1] it was defined as the *geometric complexity* of the model class considered.

We may now interpret the decomposition (29) as follows: The optimal code length for the data sequence results if we first encode the optimal model as one of the distinguishable models, each having equal probability  $1/C_n(\gamma)$ , after which we encode the data with the ideal code length given by the first term as one of the sequences  $y^n$  which get mapped by  $\hat{\theta}(\cdot)$  into the rectangle whose center point specifies the optimal model. Hence these equivalent sequences define models which cannot be distinguished from each other, and they add nothing to the ‘useful information’ represented by the optimal model. We are then justified to define  $\ln C_n(\gamma)$  to be the amount of *information* in the data that can be learned with the agreed model class.

When Shannon defined  $-\ln P(x^n)$  as the self information of a fixed probability distribution  $P$ , there was no reason to distinguish between information and complexity. By contrast, in our more general situation the two notions need be distinguished, and Shannon’s self information will have to be interpreted as complexity in our sense. After all, if we know the data generating distribution there is nothing further the data can teach us about the data generating machinery; the information in our sense is zero, as it should be.

### Appendix A.

Proof of Theorem 3. Dropping again the fixed structure index  $\gamma$  to simplify the notations, let  $B$  be the exceptional set  $B_n(q, \epsilon)$  in the theorem

$$B = \left\{ \hat{\theta} : \int g(x^n; \hat{\theta}) \ln \frac{f(x^n; \hat{\theta})}{q(x^n)} dx^n \leq (1 - \epsilon) \ln C_n \right\}, \quad (30)$$

and define

$$\begin{aligned} g_{\bar{w}}(x^n) &= f(x^n | \hat{\theta}) \bar{w}(\hat{\theta}) \\ g_{\bar{w}}^B(x^n) &= \begin{cases} f(x^n | \hat{\theta}) \bar{w}(\hat{\theta}) / \bar{w}(B) & \text{for } \hat{\theta} \in B \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

$$g_{\bar{w}}^{\bar{B}}(x^n) = \begin{cases} f(x^n|\hat{\theta})\bar{w}(\hat{\theta})/\bar{w}(\bar{B}) & \text{for } \hat{\theta} \in \bar{B} \\ 0 & \text{otherwise} \end{cases}$$

where  $\bar{B} = \Omega - B$ . We then have the identity

$$\begin{aligned} \int_{\Omega} d\hat{\theta} \int g_{\bar{w}}(x^n) \ln \frac{f(x^n; \hat{\theta})}{g_{\bar{w}}(x^n)} dx^n &\equiv \bar{w}(B) \int_B d\hat{\theta} \int g_{\bar{w}}^B(x^n) \ln \frac{f(x^n; \hat{\theta})}{g_{\bar{w}}^B(x^n)} dx^n \\ &+ \bar{w}(\bar{B}) \int_{\bar{B}} d\hat{\theta} \int g_{\bar{w}}^{\bar{B}}(x^n) \ln \frac{f(x^n; \hat{\theta})}{g_{\bar{w}}^{\bar{B}}(x^n)} dx^n \\ &+ h(\bar{w}(B)), \end{aligned} \quad (31)$$

where, we remind the reader,  $h(\cdot)$  denotes the binary entropy function.

Using the definition of the exceptional set  $B$  we can upper bound the integral in the first term of the right hand side as follows:

$$\int_B d\hat{\theta} \int g_{\bar{w}}^B(x^n) \ln \frac{f(x^n; \hat{\theta})}{g_{\bar{w}}^B(x^n)} dx^n \leq \int_B d\hat{\theta} \int g_{\bar{w}}^B(x^n) \ln \frac{f(x^n; \hat{\theta})}{q(x^n)} dx^n \quad (32)$$

$$\leq (1 - \epsilon) \ln C_n, \quad (33)$$

the first inequality following from Shannon's inequality. The integral in the second term of (31) is upper bounded by  $\ln C_n$ , because the restricted prior  $\bar{w}^{\bar{B}}(\theta)$  differs from  $\bar{w}(\theta)$ , which maximizes (25). Replacing the left hand side of (31) by  $\ln C_n$  we get with these upper bounds, as in [8], the inequality

$$\ln C_n \leq \bar{w}(B)(1 - \epsilon) \ln C_n + (1 - \bar{w}(B)) \ln C_n + h(\bar{w}(B)). \quad (34)$$

This reduces to

$$\ln \frac{1}{\bar{w}(B)} + \frac{1 - \bar{w}(B)}{\bar{w}(B)} \ln \frac{1}{1 - \bar{w}(B)} \geq \epsilon \ln C_n.$$

Since the maximum of the second term is 1  $\bar{w}(B) \rightarrow 0$  as  $\ln C_n \rightarrow \infty$ . If  $\bar{p}$  and  $\hat{p}$  are the infimum and the supremum, respectively, in (27), we have

$$\int_{\Omega} p(\hat{\theta}) d\hat{\theta} = C_n \leq \hat{p}|\Omega| < \bar{p}|\Omega|/K,$$

where  $|A|$  denotes the volume of a set  $A$ , and

$$|B| = \int_B d\hat{\theta} \leq \frac{1}{\bar{p}} \int_B p(\hat{\theta}) d\hat{\theta} = \frac{1}{\bar{p}} C_n \bar{w}(B) < \frac{|\Omega|}{K} \bar{w}(B). \quad (35)$$

Because  $\bar{w}(B)$  shrinks to zero as  $n$  grows to infinity, and so does  $|B|$ .

## References

- [1] Balasubramanian, V. (1996), ‘A Geometric Formulation of Occam’s Razor for Inference of Parametric Distributions’, *Princeton physics preprint PUPT-1588*, Princeton, NJ
- [2] Balasubramanian, V. (1996), ‘Statistical Inference, Occam’s Razor and Statistical Mechanics on the Space of Probability Distributions’, *Neural Computation*, **9**, No. 2, 349-268, 1997 <http://arxiv.org/list/nlin/9601>
- [3] Barron, A.R., Rissanen, J., and Yu, B. (1998), ‘The MDL Principle in Modeling and Coding’, special issue of *IEEE Trans. Information Theory* to commemorate 50 years of information theory, Vol. **IT-44**, No. 6, October 1998, pp 2743-2760
- [4] Clarke, B. S. and Barron, A. R. (1990), ‘Information-Theoretic Asymptotics of Bayes Methods’, *IEEE Trans. Information Theory*, Vol. **IT-36**, No. 3, 453-471, May 1990.
- [5] Davisson, L.D. (1973), ‘Universal Noiseless Coding’, *IEEE Trans. Information Theory*, Vol. **IT-19**, 783-795, November 1973
- [6] Grünwald, P.D. (1998), *The Minimum Description Length Principle and reasoning under Uncertainty*, PhD thesis, Institute for Logic, Language and Computation, Universiteit van Amsterdam, 296 pages
- [7] Li, L. and Yu, B. (2000), ‘Iterated Logarithmic Expansions of the Pathwise Code Lengths for Exponential Families’, *IEEE Trans. on Information Theory*, Vol. **IT-46**, Nr. 7, November 2000.
- [8] Merhav, N. and Feder, M. (1995), ‘A Strong Version of the Redundancy-Capacity Theorem of Universal Coding’, *IEEE Trans. Information Theory*, Vol. **IT-41**, No. 3, pp 714-722, May 1995.
- [9] Rissanen, J. (1984), ‘Universal Coding, Information, Prediction, and Estimation’, *IEEE Trans. Information Theory*, Vol. **IT-30**, Nr. 4, 629-636

- [10] Rissanen, J. (1986), 'Stochastic Complexity and Modeling', *Annals of Statistics*, Vol **14**, 1080-1100
- [11] Rissanen, J. (1989), *Stochastic Complexity in Statistical Inquiry*, World Scientific, New Jersey, 175 pages, (second edition)
- [12] Rissanen, J. (1996), 'Fisher Information and Stochastic Complexity', *IEEE Trans. Information Theory*, Vol. **IT-42**, No. 1, pp 40-47
- [13] Rissanen, J. (2000), 'MDL Denoising', *IEEE Trans. on Information Theory*, Vol. **IT-46**, Nr. 7, November 2000. Also <http://www.cs.tut.fi/~rissanen/>.
- [14] Shtarkov, Yu. M. (1987), 'Universal Sequential Coding of Single Messages', Translated from Problems of Information Transmission, Vol. 23, No. 3, 3-17, July-September 1987.
- [15] Takeuchi, Jun-ichi and Barron, Andrew R. (1998), 'Robustly Minimax Codes for Universal Data Compression', *The 21'st Symposium on Information Theory and Its Applications*, Gifu, Japan, December 2-5, 1998.
- [16] White, H. (1994), *Estimation, inference, and specification analysis*, Cambridge University press, Cambridge, UK, 349 pages