

# The Competitive Complexity Ratio

R. A. Stine and D. P. Foster<sup>1</sup>

Department of Statistics

The Wharton School of the

University of Pennsylvania

Philadelphia, PA 19104-6302

`www-stat.wharton.upenn.edu/~bob`

*Abstract* — **The competitive complexity ratio is the worst case ratio of the regret of a data-driven model to that obtained by a model which benefits from side information. The side information bounds the sizes of unknown parameters. The ratio requires the use of a variation on parametric complexity, which we call the unconditional parametric complexity. We show that the optimal competitive complexity ratio is bounded and contrast this result with comparable results in statistics.**

## I. INTRODUCTION

Stochastic complexity measures the ability of a family of models to represent an observed data sequence  $Y = (Y_1, \dots, Y_n)$ . Stochastic complexity is the length of the minimax code for  $Y$  obtained by a member of the family. The resulting code for  $Y$  may be divided into two parts. One part encodes the data. For parametric models, say  $M_\theta$ , the data is encoded using the maximum likelihood model  $M_{\hat{\theta}}$ , where  $\hat{\theta}$  is the MLE of the parameters. The other part of the code, whose length is known as the parametric complexity, is the focus of our interest. This portion of the code represents the model itself. Because models with many parameters typically have large complexity, this part of the code guards against over-fitting when stochastic complexity is used for model selection.

Parametric complexity is a property of the model class  $M_\theta$  and is invariant of  $Y$ . It thus serves as a uniform measure of the complexity of  $M_\theta$ , one that does not depend upon nuances of the observed sequence. Because of this uniform assessment of model complexity, stochastic complexity permits a refined version of model selection using the minimum description length (*MDL*) criterion. *MDL* selects among competing

models based on the length of a uniquely decodable prefix code for the observed data, picking the model that obtains the shortest code. Since stochastic complexity gives the length of the best code for each model class (in a minimax sense), it defines the basis for comparing different models using *MDL*.

In many common settings, however, the parametric complexity is infinite. For example, the parametric complexity is unbounded in the normal location problem unless one restricts the size of the unknown mean. The competitive complexity ratio avoids this problem by considering codes which benefit from such restrictions. The restrictions themselves are an integral component of the competitive analysis.

In the next section, we briefly review the definition of stochastic complexity. We then consider the Gaussian location problem and introduce the competitive complexity ratio. We show that the best complexity ratio is bounded in the normal location problem. The bound is a solution of a numerical integration in general, but simplifies nicely in a simplified context. We then extend these results to multivariate problems and close with a short discussion.

## II. STOCHASTIC COMPLEXITY

Early versions of Rissanen's *MDL* model selection criterion [6] assess the ability of a model to represent data using the length of a two-part code. Let  $Y$  denote  $n$  observations with probability distribution  $P_{\theta_p}(y)$  which is indexed by some  $p$  dimensional parameter vector  $\theta_p \in \Theta_p \subset \mathbf{R}^p$ . As shown by Rissanen, it is most efficient in this type of coding to round the maximum likelihood estimator  $\hat{\theta}_p = \hat{\theta}_p(Y)$  to order  $O(1/\sqrt{n})$ , corresponding to an integer grid position  $\tilde{z}_p$  within  $\Theta_p$ . (Throughout, we will use ' $\sim$ ' to denote rounded values or properties of rounded values.) In the orthogonal case, the resulting vector  $\tilde{z}_p$  encodes each element of  $\hat{\theta}_p$  as a whole number of standard errors from the origin. The *idealized* length

<sup>1</sup>This work was supported by NSF Grant DMS-9704809

of the two-part code obtained by the  $p$  dimensional model is then

$$L(Y, p) = \ell(p) + \ell_s(\tilde{z}_p) + \log \frac{1}{P_{\hat{\theta}_p}(Y)} + \delta, \quad (1)$$

where  $\ell(p)$  is the length of a prefix code for the dimension  $p$ ,  $\ell_s(\tilde{z}_p)$  denotes the length of a ‘spiraling’ prefix code for the rounded vector of  $z$  scores [6], and  $\delta$  denotes a small remainder due to rounding  $\hat{\theta}_p$  to standard error scale. This form of the *MDL* criterion selects the model class that obtains the shortest code for the data, choosing the dimension  $p$  which minimizes  $L(Y, p)$ . All logs here and in what follows are to base 2 unless otherwise distinguished. The idealized code length is real valued and avoids the issue of quantization (see [1]).

Since *MDL* selects the model class obtaining the shortest code, the coding method must be efficient. Two-part codes such as the one just described, however, are not Kraft tight. The implicit codebook reserves symbols which will not be used. Once the receiver of the code decodes the dimension  $p$  and recovers  $\tilde{z}_p$  from the first part of the code, the set of possible values for the data  $Y$  becomes restricted to those values for which  $\hat{\theta}_p$  rounds to  $\tilde{z}_p$ . The resulting dependence implies that the data can be coded using fewer than  $\log 1/P_{\hat{\theta}_p}(Y)$  bits. Rissanen [7], for example, illustrates the calculations in the Bernoulli case. Although the effects are typically small and perhaps not important in data compression, such differences are important in model selection since the choice among models is often decided by just a few bits.

Stochastic complexity replaces these two-part codes with a tight, one-part code that no longer specifies a parameter value. Stochastic complexity encodes the data using the so-called normalized maximum likelihood (NML) distribution [11]. This distribution is formed by finding the integrating constant (whose log is known as the parametric complexity),

$$C_{n,p} = \int_Y P_{\hat{\theta}_p(Y)}(Y) dY, \quad (2)$$

that makes  $g(Y) = P_{\hat{\theta}_p(Y)}(Y)/C_{n,p}$  a density. The range of integration in (2) is over all possible  $Y$ , and we assume for the moment that this integral is finite. In regular problems, the parametric complexity (2) has a particularly nice asymptotic form [8]

$$\log C_{n,p} = \frac{p}{2} \log \frac{n}{2\pi} + \log \int_{\Theta_p} |I(\theta_p)|^{1/2} d\theta_p + o(1), \quad (3)$$

where  $I(\theta_p)$  is the asymptotic Fisher information matrix

$$I_{ij}(\theta_p) = \lim_{n \rightarrow \infty} -\frac{1}{n} \frac{\partial^2 \log P_{\theta_p}(Y)}{\partial \theta_{p,i} \partial \theta_{p,j}}.$$

The leading summand of (3) motivates the common association of *MDL* with the Bayesian information criterion *BIC* since it suggests a parameter penalty which grows logarithmically in  $n$ . (This association is spurious; see [5].) The idealized length of the resulting one-part code for  $Y$ , or stochastic complexity, using the  $p$  dimensional model  $P_{\theta_p}(Y)$  is then

$$S_p(Y) = \log C_{n,p} + \log \frac{1}{P_{\hat{\theta}_p(Y)}(Y)}.$$

Compared to the length of a two-part code, stochastic complexity replaces the lengths of the prefixes  $\ell(p)$  and  $\ell_s(\tilde{z}_p)$  in (1) by the log of an integral, the parametric complexity. Thus parametric complexity avoids the choice of a prefix code for the discretized parameter and the need to find the conditional density of  $Y$  given  $\tilde{z}_p$ . Further, the absence of a rounded estimate simplifies the comparison of models because it avoids the need to consider the complex quadratic patterns induced by rounding [5].

### III. THE COMPETITIVE COMPLEXITY RATIO

To introduce the competitive complexity ratio, we consider encoding a scalar location model for Gaussian data. We first consider the impact of parameter constraints on the parametric complexity. For this section, we assume  $Y_i \sim N(\mu, \sigma^2)$ ,  $i = 1, \dots, n$ , with  $\mu$  unknown and  $\sigma^2$  given. The likelihood function is

$$P_\mu(Y) = \frac{e^{-\sum (Y_i - \mu)^2 / 2}}{(2\pi\sigma^2)^{n/2}}.$$

In general, the parameter space for  $\mu$  is unbounded, and the integral (2) which defines the parametric complexity is infinite. To reinforce its role in defining the parameter space, we denote a constraining interval for  $\mu$  by

$$\Theta_{[a,b]} = \{\mu : a \leq \mu \leq b, -\infty < a \leq b < \infty\}.$$

Under the condition that  $a \leq \mu \leq b$ , the maximum likelihood estimator of  $\mu$  is not  $\bar{Y} = \sum_i Y_i/n$ , but is restricted to this range,

$$\hat{\mu} = \begin{cases} a, & \bar{Y} < a, \\ \bar{Y}, & a \leq \bar{Y} \leq b, \text{ and} \\ b, & \bar{Y} > b. \end{cases}$$

With bounds on the parameter space, the integration is finite and the parametric complexity is well-defined. Following [1], the parametric complexity is most easily found by using the sufficiency of  $\bar{Y}$  for  $\mu$ . The distribution of  $Y$  factors into

$$P_\mu(Y) = P(Y|\bar{Y})h_\mu(\bar{Y})$$

where  $P(Y|\bar{Y})$  is the conditional distribution of  $Y$  given  $\bar{Y}$  (which is free of  $\mu$  by sufficiency) and  $h_\mu(\bar{Y})$  is the distribution of  $\bar{Y}$ ,

$$h_\mu(\bar{y}) = \left(\frac{n}{2\pi\sigma^2}\right)^{1/2} e^{-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2}.$$

Since  $P(Y|\bar{Y})$  is a density, the parametric complexity normalizes just the maximum likelihood density for the sufficient statistic. When  $\hat{\mu} = \bar{Y}$ , this density reduces to a constant,

$$h_{\bar{Y}}(\bar{Y}) = \left(\frac{n}{2\pi\sigma^2}\right)^{1/2}. \quad (4)$$

Integrating over all sequences, the parametric complexity under the constraint  $\mu \in \Theta_{[a,b]}$  is

$$C_n(\Theta_{[a,b]}) = \int_Y P_{\hat{\mu}}(Y) dY \quad (5)$$

$$= \int_a^b h_m(m) dm + \int_{-\infty}^a h_a(m) dm + \int_b^{\infty} h_b(m) dm \quad (6)$$

$$= 1 + \frac{1}{\sqrt{2\pi}} \frac{b-a}{\sigma/\sqrt{n}}. \quad (7)$$

This calculation is one larger than similar expressions in various papers of Rissanen and coauthors (such as the review [1]) because  $\mu$ , not the sample mean  $\bar{Y}$ , is restricted to  $\Theta_{[a,b]}$ . Since the data are unrestricted, we term  $\log C_n(\Theta_{[a,b]})$  the *unconditional* parametric complexity of this model given  $\mu \in \Theta_{[a,b]}$ . The assumption that  $\mu \in \Theta_{[a,b]}$  constrains the MLE, not the data, so that the range for  $Y$  in (5) is unrestricted. Subsequently, we use the term *conditional* parametric complexity written as  $C_n(\bar{Y}|\Theta_{[a,b]})$  to refer to the integral in (6) which does not include the boundary contribution,

$$\begin{aligned} C_n(\bar{Y}|\Theta_{[a,b]}) &= \int_a^b h_m(m) dm = C_n(\Theta_{[a,b]}) - 1 \\ &= \frac{1}{\sqrt{2\pi}} \frac{b-a}{\sigma/\sqrt{n}}. \end{aligned} \quad (8)$$

This notation reinforces the distinction that the data are constrained in the definition of the conditional parametric complexity.

When combined with the code length for the data, the unconditional parametric complexity gives the total code length, or stochastic complexity of  $Y$ . Given the constraint  $a \leq \mu \leq b$ , the stochastic complexity is

$$\begin{aligned} L(Y, \Theta_{[a,b]}) &= \log \frac{C_n(\Theta_{[a,b]})}{P_{\hat{\mu}}(Y)} \\ &= \log C_n(\Theta_{[a,b]}) \\ &\quad + \log \frac{P_{\bar{Y}}(Y)}{P_{\hat{\mu}}(Y)} + \log \frac{1}{P_{\bar{Y}}(Y)} \end{aligned} \quad (9)$$

The log of the likelihood ratio or observed relative entropy

$$\log \frac{P_{\bar{Y}}(Y)}{P_{\hat{\mu}}(Y)} = \frac{\log e}{2} \frac{n(\bar{Y} - \hat{\mu})^2}{\sigma^2}$$

measures the increase in code length that occurs when  $\bar{Y}$  falls outside the parameter space for which the code is designed. As long as  $\hat{\mu}$  is near  $\bar{Y}$ , the cost in bits for enforcing such constraints is small. For example, if  $\bar{Y} > b$ , the increase in code length is a multiple of the squared  $z$  statistic for testing  $H_0 : \mu = b$ .

One means to bound the parametric complexity in this model is to incorporate bounds as part of the code itself. This approach is reminiscent of a two-part code. The first part of the code indicates  $\Theta_{[a,b]}$ , and the second part encodes the data under this constraint. As usually implemented, however, the prefix gives a range for the observed statistic  $\bar{Y}$  rather than the parameter  $\mu$ , and the subsequent code uses the conditional parametric complexity. The structure of the first part varies in how the region for  $\bar{Y}$  is specified. For example, in [1] the region is defined as  $\bar{Y}^2 \leq R^2$  with the prefix encoding  $R^2$  or perhaps  $\log R^2$ . Alternatively, one might constrain  $\bar{Y}$  on a standardized scale as  $\bar{Y}^2 \leq r^2 \sigma^2/n$  as in [9]. Taking a different approach, one can follow the logic leading to the NML density and perform a further normalization over the parameter space [10].

Rather than consider various means of incorporating information about the parameter space  $\Theta_{[a,b]}$  directly into the code, we instead consider a competitive analysis of how well a realizable code fares when compared to a code that *knows* features of the true parameter space. We formulate this information about the parameter space as a collection of ‘experts’ that define sets that contain the process mean  $\mu$ . Let  $\mathcal{A}$  denote a collection of intervals of  $\mathbf{R}$  where each interval  $A \in \mathcal{A}$  has finite length,  $\lambda(A) < \infty$ . Intuitively, these intervals represent the advice from various ‘experts’ in the following sense. When coding  $Y$ , each interval  $A \in \mathcal{A}$  implies a code length  $L(Y, A)$  as given by (9); this is the number of bits required to encode  $Y$  under the assumption  $\mu \in A$ . The ‘advice’ of the best expert produces the shortest code for  $Y$ ,

$$L^*(Y, \mathcal{A}) = \inf_{A \in \mathcal{A}} L(Y, A),$$

and obtains the minimal regret,

$$\begin{aligned} R^*(Y, \mathcal{A}) &= L^*(Y, \mathcal{A}) - \log \frac{1}{P_{\bar{Y}}(Y)} \\ &= \inf_{A \in \mathcal{A}} \log C_n(A) + \log \frac{P_{\bar{Y}}(Y)}{P_{\hat{\mu}_A}(Y)} \\ &= \inf_{A \in \mathcal{A}} \log \left( 1 + \frac{\lambda(A)}{\sqrt{2\pi\sigma^2/n}} \right) \\ &\quad + \frac{\log e}{2} \frac{n(\bar{Y} - \hat{\mu}_A)^2}{\sigma^2}, \end{aligned} \quad (10)$$

where the MLE is  $\hat{\mu}_A \in A$ . Since the optimal regret depends on  $Y$  only through its mean  $\bar{Y}$ , we will also write it as  $R^*(\bar{Y}, \mathcal{A})$ .

Now consider the task of coding  $Y$  without the assistance of such an expert. Let  $L(Y, \alpha)$  and  $R(Y, \alpha)$  denote the length and regret, respectively, obtained by a uniquely decodable prefix code  $\alpha$  that does not benefit from the advice of such experts. To compare this coding procedure to that obtained through the use of experts, consider the worst-case ratio of the regret of  $\alpha$  to the regret of the code provided by the best expert,

$$\rho_n(\alpha, \mathcal{A}) = \sup_Y \frac{R(Y, \alpha)}{R^*(Y, \mathcal{A})}.$$

We call  $\rho_n$  the *competitive complexity ratio*. If  $\rho_n(\alpha, \mathcal{A})$  is bounded for every  $n$ , we shall say that the coding procedure  $\alpha$  provides a *universal code for this model class* with respect to the collection  $\mathcal{A}$  of experts. Were we to define  $\rho_n$  using the full code lengths  $L$  and  $L^*$ , the ratio would not be discriminating since the likelihood component  $\log 1/P_{\bar{Y}}(Y) = O(n)$  would dominate the comparison of code lengths for finite-dimensional models. Given a class of experts, one prefers coding procedures for which  $\rho_n$  is small. The competitive ratio  $\rho_n$  has some intuitive properties with regard to the set of experts. In particular,  $\rho_n$  increases with the collection of experts. If we let  $\alpha$  denote a coding procedure and  $\mathcal{A}_1$  and  $\mathcal{A}_2$  two sets of experts, then it follows that

$$\mathcal{A}_1 \subset \mathcal{A}_2 \Rightarrow \rho_n(\alpha, \mathcal{A}_1) \leq \rho_n(\alpha, \mathcal{A}_2). \quad (11)$$

It is not possible to obtain bounded competitive regret in the sense of  $\rho_n$  for arbitrary classes of experts. For example, for any  $0 < \delta < \infty$ , let

$$\mathcal{B}_\delta = \cup_{x \in \mathbf{R}} \{y \in \mathbf{R} : |y - x| < \delta\},$$

denote the set of arbitrarily translated balls of radius  $\delta$ . Clearly, when the radius is small, say  $\delta < \sigma/\sqrt{n}$  (the standard error of  $\bar{Y}$ ) these experts — truly more like oracles in this case — essentially reveal the value of the MLE, and  $R^*(Y, \mathcal{B}_\delta)$  is but one or two bits for any  $Y$ . No coding strategy can maintain bounded competitive regret for all  $Y$  versus such experts, for any finite radius  $\delta$ . To see that this is so, suppose that  $\beta$  were such a code. Since the sets in  $\mathcal{B}_\delta$  are of fixed size,  $R(Y, \beta)$  would be bounded by some constant, say  $R(Y, \beta) < B$ . Now choose a set of the form  $\Theta_{[-c, c]}$  where  $c$  is sufficiently large so that the parametric complexity  $\log C_n(\Theta_{[-c, c]}) > B$ . Since the parametric complexity is the minimax regret, it cannot everywhere be greater than the regret obtained by the code

$\beta$ , and we have a contradiction. Thus in the Gaussian location problem, no coding procedure has finite competitive complexity ratio versus the experts  $\mathcal{B}_\delta$ . In addition, (11) implies that we cannot obtain bounded competitive regret for any class of experts containing  $\mathcal{B}_\delta$ . Since we cannot compete against such experts, we need to consider a less informative collection that is not uniformly well-informed for all  $\mu$ .

In keeping with our interest in model selection, a more realistic class of experts consists of all intervals that contain the origin. The previous experts  $\mathcal{B}_\delta$  are equally precise for all  $Y$  and have constant regret. A less informative collection are more accurate for certain sequences, in particular sequences with mean near zero. Let  $\mathcal{E}_0$  denote the set of intervals of *positive* length that include zero,

$$\mathcal{E}_0 = \cup \Theta_{[a, b]}, \quad a \leq 0 \leq b, \quad a \neq b.$$

For this class of ‘origin-covering experts’, the best expert is the interval that minimizes the regret (10). If  $\bar{Y} \geq 0$ , the left endpoint of this interval is zero. Expressed on the standard error scale, the right endpoint of the best interval is  $\hat{z}\sigma/\sqrt{n} > 0$ , where  $\hat{z}$  is defined by

$$\hat{z} = \arg \min_{z > 0} \log \left( 1 + \frac{z}{\sqrt{2\pi}} \right) + \frac{\log e}{2} (z - z_{\bar{Y}})^2, \quad (12)$$

with

$$z_{\bar{Y}} = \sqrt{n} \bar{Y} / \sigma. \quad (13)$$

The expression for the optimal endpoint has ‘kinks’ at  $|z_{\bar{Y}}| = 1/\sqrt{2\pi}$ . For  $|z_{\bar{Y}}| \leq 1/\sqrt{2\pi}$ ,  $\hat{z} = 0$ . For larger  $z_{\bar{Y}} > 1/\sqrt{2\pi}$ ,

$$\hat{z} = \frac{1}{2} \left( z_{\bar{Y}} - \sqrt{2\pi} + \sqrt{(z_{\bar{Y}} + \sqrt{2\pi})^2 - 4} \right), \quad (14)$$

whereas for  $z_{\bar{Y}} < -1/\sqrt{2\pi}$ ,

$$\hat{z} = \frac{1}{2} \left( z_{\bar{Y}} + \sqrt{2\pi} - \sqrt{(z_{\bar{Y}} - \sqrt{2\pi})^2 - 4} \right). \quad (15)$$

One obtains a slightly shorter message length by picking an expert whose parameter region does *not* contain  $\bar{Y}$ . The endpoint of the interval is shrunken toward zero. For  $\bar{Y} > 0$ , the shrinkage toward zero is about

$$\hat{z} - z_{\bar{Y}} \approx \frac{-2}{z_{\bar{Y}} + \sqrt{2\pi}}, \quad \text{for } z_{\bar{Y}} \gg 0.$$

It is a straightforward task to find a coding procedure which obtains the minimax competitive complexity ratio. Our direct approach is to find a Kraft-tight prefix code  $\alpha_0^*$  for which the competitive complexity ratio is constant,

$$\sup_Y \frac{R(Y, \alpha_0^*)}{R^*(Y, \mathcal{E}_0)} = \rho_0^*,$$

say. Given such a code, the fact that it is Kraft tight implies that it is the minimax code since any other code which is shorter for coding some  $Y$  will also be longer for some other  $Y'$ . To construct  $\alpha_0^*$ , we observe that no prefix code can obtain the expert regret  $R^*(Y, \mathcal{E}_0)$  for all  $Y$  because the ‘density’ for  $z_{\bar{Y}}$  implied by the regret  $R^*(Y, \mathcal{E}_0)$ ,

$$f(z) = (1/\sqrt{2\pi})2^{-R^*(z\sigma/\sqrt{n}, \mathcal{E}_0)},$$

is not integrable. (The constant  $1/\sqrt{2\pi}$  arises from the maximum likelihood density  $h_{\bar{Y}}$  given in (4).) It is, however, a fairly simple numerical problem to find the smallest constant  $\rho_0^*$  for which

$$f^*(z) = (1/\sqrt{2\pi})2^{-\rho_0^* R^*(z\sigma/\sqrt{n}, \mathcal{E}_0)} \quad (16)$$

is a density. In this problem, the multiplier is approximately  $\rho_0^* \approx 3.26863$ . Because of the segmented form of the optimal endpoint  $\hat{z}$  given in (14) and (15), we integrated  $f^*(z)$  over the region  $|z| < 1/\sqrt{2\pi}$  analytically and added to this a numerical estimate of the integral over the rest of the parameter space. The code  $\alpha_0^*$  could then be implemented using an arithmetic coder for the mixture density

$$g_0(Y) = \int P(Y|\bar{y})f^*(\sqrt{n}\bar{y}/\sigma)(\sqrt{n}/\sigma)d\bar{y}.$$

We summarize this result as

**Theorem 1** *The minimax competitive complexity ratio in comparison to codes based upon the experts  $\mathcal{E}_0$  is*

$$\inf_{\alpha} \sup_Y \frac{R(Y, \alpha)}{R^*(Y, \mathcal{E}_0)} \approx 3.26863,$$

which is attained by the code  $\alpha_0^*$  implied by the density  $f^*(z) = (1/\sqrt{2\pi})2^{-\rho_0^* R^*(z\sigma/\sqrt{n}, \mathcal{E}_0)}$ .

For those who find this numerically generated code unappealing, we construct an explicit two-part code which obtains similar performance in the next section in a simplified, approximate version of this problem.

Before closing this section, we recognize one may view the experts  $\mathcal{E}_0$  as too well informed in the sense that they ‘know’ the sign of  $\mu$ . In this case, one can consider the class of experts based on less informative, symmetric intervals  $\mathcal{E}_0^s$  around zero. For these, the best expert for coding  $Y$  with standardized mean  $z_{\bar{Y}} = \sqrt{n}\bar{Y}/\sigma$  has symmetric endpoints  $[-\hat{z}_s\sigma/\sqrt{n}, \hat{z}_s\sigma/\sqrt{n}]$  where (compare to (12))

$$\hat{z}_s = \arg \min_{z>0} \log \left( 1 + \frac{2z}{\sqrt{2\pi}} \right) + \frac{\log e}{2} (z - z_{\bar{Y}})^2.$$

The only difference from the regret obtained by the asymmetric experts  $\mathcal{E}_0$  is the doubling of the  $z$  score in the leading complexity term of (12).

The resulting optimal symmetric endpoint is 0 for  $|z_{\bar{Y}}| \leq 2/\sqrt{2\pi}$ . For  $z_{\bar{Y}} > 2/\sqrt{2\pi}$ ,

$$\hat{z}_s = \frac{1}{4} \left( 2z_{\bar{Y}} - \sqrt{2\pi} + \sqrt{(2z_{\bar{Y}} + \sqrt{2\pi})^2 - 16} \right),$$

and for  $z_{\bar{Y}} < -2/\sqrt{2\pi}$ ,

$$\hat{z}_s = \frac{1}{4} \left( 2z_{\bar{Y}} + \sqrt{2\pi} - \sqrt{(2z_{\bar{Y}} - \sqrt{2\pi})^2 - 16} \right).$$

This endpoint is zero over twice the region as with the asymmetric experts, and is also shrunken toward zero. We again find the minimax complexity ratio by determining the smallest multiple  $c$  of the regret for which

$$(1/\sqrt{2\pi})2^{-c R^*(\hat{z}_s\sigma/\sqrt{n}, \mathcal{E}_0^s)}$$

is a density. The same combination of analytic and numerical integration shows that the competitive complexity ratio vis-a-vis symmetric experts is about two-thirds that for asymmetric experts,

$$\min_{\alpha} \rho_n(\alpha, \mathcal{E}_0^s) \approx 2.2398.$$

#### IV. RESULTS FOR CODES WITH INTEGER REGRET

The regret of the minimax codes in the previous section approaches zero as the standardized mean  $\sqrt{n}\bar{Y}/\sigma$  goes to zero. Such performance is only possible when using a one-part code like  $\alpha_0^*$  in a context in which the gain of a fractional bit can be realized. These gains are real when coding an ensemble of many sequences, each with its own distinct mean value; here, the fractional bits can be accumulated and the savings realized. When coding a single series, however, gains of a fractional bit offer no advantage. In such cases, it becomes interesting to study the competitive complexity ratio when the regret takes on integer values. The results in this section are also more in the spirit of two-part codes and lead to methods that are familiar in that context.

In order to work with two-part codes, we define the regret as

$$\tilde{R}(Y, A) = \left\lceil L(Y, A) - \log \frac{1}{P_{\bar{Y}}(Y)} \right\rceil. \quad (17)$$

This regret is the least upper bound on the actual difference in integer code lengths under arbitrary quantization,  $R(Y, A) \leq \tilde{R}(Y, A)$ . This definition also gives a regret as an integer so that we can think of it as the explicit length in bits of a prefix.

With this definition and a naive selection of experts, we can construct a two-part code that obtains the minimax competitive complexity ratio, which in this case is 2. For the rest of this section, we consider the following competitive complexity ratio

$$\tilde{\rho}_n(\alpha, \mathcal{A}) = \sup_Y \frac{\tilde{R}(Y, \alpha)}{\tilde{R}^*(Y, \mathcal{A})}.$$

In addition, we define the minimum expert regret  $\tilde{R}^*$  to capture the notion of naive selection of experts by forcing the chosen expert to contain the sufficient statistic  $\bar{Y}$ ,

$$\tilde{R}^*(Y, \mathcal{A}) = \min_{A \in \mathcal{A}: \bar{Y} \in A} \tilde{R}(Y, A).$$

Under this definition with  $\bar{Y} > 0$ , the interval of the best expert is  $[0, \bar{Y}]$ , and the minimum regret is

$$\tilde{R}^*(Y, \mathcal{E}_0) = \tilde{R}(Y, \Theta_{[0, \bar{Y}]}) = \left\lceil \log \left( 1 + \frac{z_{\bar{Y}}}{\sqrt{2\pi}} \right) \right\rceil, \quad (18)$$

where  $z_{\bar{Y}} = \sqrt{n} \bar{Y} / \sigma$ . Because of rounding,  $\tilde{R}^*$  is a step function with increments where  $|z_{\bar{Y}}| = \sqrt{2\pi}(2^j - 1)$ ,  $j = 1, 2, \dots$

A variety of two-part coding procedures  $\alpha$  have bounded competitive regret  $\tilde{\rho}_n(\alpha, \mathcal{E}_0)$  under this definition. Their construction takes the following general approach: form a countable partition of the parameter space and construct a two-part code by attaching a prefix with a universal code for the index of the chosen subset to the message. The second part of the message encodes  $Y$  given  $\bar{Y}$  lies in the region indicated by the prefix. To be competitive versus  $\mathcal{E}_0$ , such a procedure must use short codes when competing against accurate experts, the small sets in  $\mathcal{E}_0$  near the origin. To accomplish this, we enumerate a partition of the parameter space by counting out from the origin and encoding the index using a prefix code for integers. One such prefix code is the so-called universal prior of Rissanen [6]. This code represents the positive integer  $j > 0$  using about  $2.9 + \log^* j$  bits, where  $\log^* x = \log x + \log \log x + \dots$ , and the summands are added so long as the prior term is positive. A simpler code for analysis is the so-called doubly-compound code of Elias [3]. This code concatenates a simple prefix code for  $\log j$  with the binary representation of  $j$ . The length of the doubly compound code is about  $\ell_d(j) \approx \log j + 2 \log \log j$  bits. Both of these codes are asymptotically optimal as defined in [3]. The length of each grows at a rate  $\log j + o(\log j)$ . It may come as some surprise, but we will find the so-called unary code more useful. The unary code represents the integer  $j$  as a sequence of  $j$  bits:  $j - 1$  zeros followed by a single 1. The unary code is

Table 1: Two prefix codes for integers.

$j$	Doubly-Compound Code		Unary Code	
	Bits	$\ell_d(j)$	Bits	$\ell_u(j)$
1	0	1	1	1
2	10 1	3	01	2
3	1100 10	6	001	3
4	1100 11	6	0001	4
8	1110 111	7	0000001	8
32	-	11	-	32

not optimal in the sense of [3] or [6], but codes small integers particularly well. Table 1 shows the doubly compound and unary codes for several small integers.

A coarse partition of the parameter space indexed with a unary prefix produces a code  $\gamma$  which is minimax with respect to  $\tilde{\rho}_n$ . The partitioning of the optimal procedure divides the parameter space into sets of increasing size as we move from the origin. In particular, the optimal partition is a set of intervals whose boundaries are located at points of increase of the naive expert regret (18),  $z_{\bar{Y}} = \pm(2^j - 1)\sqrt{2\pi}$ . These points define a partition of the positive half of the parameter space into intervals which we denote as

$$I_j = [(2^{j-1} - 1)\sqrt{2\pi}\sigma/\sqrt{n}, (2^j - 1)\sqrt{2\pi}\sigma/\sqrt{n}].$$

The prefix of  $\gamma$  is formed as follows. The first bit denotes the sign of  $\bar{Y}$ , so we can hence restrict attention to the positive real axis. The next bits of the prefix give the unary code for the smallest  $j$  such that  $\bar{Y} \in I_j$ . Since the enumerated intervals are growing geometrically in length, this enumeration is in effect on a log scale. The rest of the prefix accounts for the conditional parametric complexity of  $I_j$ . Each partition of the parameter space identifies the location of  $\bar{Y}$  rather than  $\mu$ , and thus the conditional parametric complexity measures the associated regret. From (8), the conditional parametric complexity of the  $j$ th interval satisfies  $\log C_n(\bar{Y}|I_j) = j - 1$  and so consumes  $j - 1$  bits. To summarize, the code  $\gamma$  requires a sign bit, the unary code for the index  $j$  of the interval  $I_j$  containing  $\bar{Y}$ , and the log of the conditional parametric complexity of  $I_j$ . The regret of this code is thus

$$\tilde{R}(Y, \gamma) = 1 + j + (j - 1) = 2j, \quad |\bar{Y}| \in I_j, \quad j = 1, 2, \dots$$

By construction, the regret of the naive expert code is piecewise constant with value  $j$  when  $\bar{Y}$  lies in  $I_j$ ,  $\tilde{R}^*(Y, I_j) = j$ .

Consequently, the regret of  $\gamma$  is precisely twice this so that

$$\tilde{\rho}_n(\gamma, \mathcal{E}_0) = 2,$$

which can be compared to the competitive complexity ratio  $\rho_0^* \approx 3.3$  obtained with continuous regret and optimized experts. Since the complexity ratio for  $\gamma$  is fixed for all  $Y$ ,  $\gamma$  obtains the minimax regret given the naive selection of experts. Any code which has less regret than  $\gamma$  for some  $\bar{Y}$  will do worse than  $\gamma$  for some  $\bar{Y}'$  since  $\gamma$  is Kraft tight. A similar procedure produces the minimax regret versus symmetric experts.

## V. MULTIVARIATE MODELS

The results obtained for  $\rho_n$  in the scalar location model extend immediately to normal models with  $p$  parameters and orthogonal estimates. An important illustration of such models are wavelet regression models used in function estimation and denoising ([2], [10]). In an orthogonal regression, the expert code has access to a collection of coordinate experts that supply an interval for each of the model parameters  $\theta_j$ ,  $j = 1, \dots, p$ . In essence, such side information tells the expert code which parameters to include in the model. The combination of orthogonality with normality implies that the maximum likelihood estimates of the  $p$  model parameters  $\hat{\theta}_p$  are independent. Thus, an arithmetic coder for the density  $f^*(z)$  defined in (16) can efficiently represent  $p$  parameters with the same competitive ratio,  $\rho_0^* \approx 3.3$ , as obtained in the scalar problem.

## VI. DISCUSSION

Our results have several implications for the use of stochastic complexity in model selection.

First, the regret of the minimax code is about two or three times that of the competing code which is given the best intervals for the model parameters. Choosing a model on the basis of smallest competitive ratio will produce a different selection criterion from those often advocated for use in *MDL*[1]. The latter criteria in effect use a spherical prior for encoding the parameters, and one can construct pathological examples where the competitive complexity ratio of such codes is at least  $p/2$ .

Second, since the minimax code favors certain sequences because of the structure of the experts, our results imply that stochastic complexity is not invariant of the coded sequence. Rather than being a fixed model property as can be obtained

in the Bernoulli setting, the regret of the minimax code described in Theorem 1 depends upon the mean of the coded sequence. Although dependent upon the data and choice of experts, the notion of the competitive complexity ratio does lead to a minimax solution which is free of the ambiguity of various prefix schemes that can be used to define a range for the parameter space and so bound the integral defining the parametric complexity (2).

Finally, these results qualitatively differ from those obtained in a traditional minimax analysis in statistics. In this setting, one compares the risk attained by a regression model that can select from any of  $p$  predictors to that of a model that benefits from using the the right variables. The best ratio of expected squared error is ([2], [4])

$$\min_{\hat{Y}} \sup_{\theta} \frac{E \|\hat{Y} - EY\|^2}{(1 + \dim(\theta))\sigma^2} \leq 2 \log p,$$

and this bound is essentially tight. That is, the minimax ratio of the squared error risk of an estimator  $\hat{Y}$  to that obtained by a model using the best subspace is on the order of the log of the number of predictors,  $\log p$ . Whereas the ratio of regrets using a worst-case analysis is bounded, this ratio grows with the number of model parameters. An explanation of this difference appears to lie in the use of the maximum likelihood fit to define the worst-case regret and is the subject of our current research.

## REFERENCES

- [1] Barron, A., J. Rissanen, and B. Yu (1998). The minimum description length principle in coding and modeling. *IEEE Trans. on Information Theory*, **44**, 2743-2760.
- [2] Donoho, D. and I. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425-455.
- [3] Elias, P. (1975). Universal codeword sets and representations of the integers. *IEEE Trans. on Information Theory*, **21**, 194-203.
- [4] Foster, D. P. and E. I. George (1994). The risk inflation criterion for multiple regression. *Annals of Statistics*, **22**, 1947-1975.
- [5] Foster, D. P. and R. A. Stine (1999). Local asymptotic coding and the minimum description length. *IEEE Trans. on Information Theory*, **45**, 1289-1293.
- [6] Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, **11**, 416-431.
- [7] Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific Press, Singapore.
- [8] Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Trans. on Information Theory*, **42**, 40-47.

- [9] Rissanen, J. (1999). Hypothesis selection and testing by the MDL principle. *The Computer Journal*, **42**, 260–269.
- [10] Rissanen, J. (1999). MDL denoising. Unpublished manuscript.
- [11] Shtarkov, Y. M. (1987). Universal coding of single messages. *Problems of Information Transmission*, **23**, 3–17.