Complexity of Simple Nonlogarithmic Loss Functions

Jorma Rissanen, *Fellow, IEEE*

***Abstract*—The *loss complexity* for nonlogarithmic loss functions is defined analogously to the stochastic complexity for logarithmic loss functions such that its mean provides an achievable lower bound for estimation, the mean taken with respect to the worst case data generating distribution. The loss complexity also provides a lower bound for the worst case mean prediction error for all predictors. For the important α-loss functions $|y - \hat{y}|^\alpha$, where $y - \hat{y}$ denotes the prediction or fitting error and α is in the interval $[1, 2]$, an accurate asymptotic formula for the loss complexity is given.**

***Index Terms*—α-loss functions, complexity, maximum entropy, min-max bounds, prediction bound.**

I. Introduction

In [13], Yamanishi defined an *extended stochastic complexity* for a variety of bounded loss functions as follows:

$$-\frac{1}{\lambda}\ln\int \pi(\theta)e^{-\lambda\sum_{t=1}^{n}\delta(y_t,\, h(\boldsymbol{x}_t;\,\theta))}\,d\theta \qquad (1)$$

where $\hat{y} = h(\boldsymbol{x};\, \theta)$ is a parametric estimate or prediction of y, and $\delta(y,\, \hat{y})$ measures the distance between y and its estimate; $\pi(\theta)$ is a prior density function for the parameters and λ is another positive parameter. The main justification for this definition is that its estimation with predictor functions was shown to give effective learning algorithms such as the aggregating strategy for computational learning theory introduced by Vovk [12], who also pioneered the mixture of type (1). With Laplace' method of integration ,Yamanishi further derived an asymptotic expansion, which gives an upper bound for the predictive estimation associated with the extended stochastic complexity and for the batch-mode loss; i.e., one resulting from estimators computed from all the data. In [15] and [14], Yamanishi showed further that the extended stochastic complexity attains a min-max cumulative prediction loss under specific restricted loss functions, where the maximum is taken over sequences.

Inspired by these works, we define an extension of stochastic complexity, which we call *loss complexity*, in a way analogous to that of the stochastic complexity [10] and [11], namely, such that its mean provides a lower bound for the mean accumulated loss. The mean is taken with respect to the worst case data generating distribution in a class that need not coincide with the class of models defined by the loss function. The loss complexity gives also a lower bound for the worst case mean prediction error resulting from any predictor. The analysis of the unbounded loss functions differs drastically from that of the bounded ones, and these bounds appear to be new.

In order to be able to calculate the loss complexity we consider mainly the so-called simple loss functions, [6], for which the normalizing coefficient

$$Z_\lambda = \int_{-\infty}^{\infty} e^{-\lambda\delta(y,\, h(\boldsymbol{x};\,\theta))}\,dy$$

does not depend on $\boldsymbol{x}$ nor θ. This class is seen to include the important α-loss functions $\delta(y,\, \hat{y}) = |y - \hat{y}|^\alpha$ for positive α, [13]. For α in the range $1 \le \alpha \le 2$, we derive an accurate asymptotic formula for the loss complexity. We also determine the optimal parameter λ as a function of the data, in which case the lower bound for the worst case mean loss is given by the loss complexity itself rather than by its mean. The formula for the loss complexity provides a convenient criterion for the selection of model classes, in particular for the absolute value error function, where the lack of everywhere differentiability has been an obstacle in the past. These results allow us to generalize an earlier prediction bound for Gaussian autoregressive moving average (ARMA) processes in [9], which further shows that the lower bound for prediction and estimation, much as the stochastic complexity, is not restricted to a single worst case data generating distribution, but it actually holds in essence for a wide class of distributions.

The loss complexity for simple loss functions turns out to consist of the minimized loss and a term that can be viewed as the ideal code length for the optimal parameters, suitably weighted. The extended stochastic complexity, also, was shown in [13] to admit a similar asymptotic upper bound, where the second term was a weighted real code length for the optimally quantized parameters.

II. Distributions Induced by Simple Loss Functions

Consider a sequence of observed data $(y^n, \boldsymbol{x}^n) = (y_1, \boldsymbol{x}_1)\cdots(y_n, \boldsymbol{x}_n)$, where $\boldsymbol{x}_t = x_{t,1}\, \ldots,\, x_{t,m}$ are vectors of real-valued components and also y_t are real numbers. We are interested in modeling the data generating machinery with a parametric function $\hat{y} = h(\boldsymbol{x};\, \theta)$ to capture the statistical relationship between the two data sequences $\boldsymbol{x}^n$ and y^n, the parameters $\theta = \theta_1,\, \ldots,\, \theta_k$ ranging over a subset Ω of the k-dimensional Euclidean space. As a rule, we take this as a compact set and denote its interior by Ω°, which throughout is assumed to be nonempty. To measure the inevitable deviations between the observed values and their predicted or fitted values a loss function $\delta(y,\, \hat{y})$ is needed, which gives the accumulated loss on the data as

$$L(y^n|\boldsymbol{x}^n;\, \theta) = \sum_{t=1}^{n} \delta(y_t,\, h(\boldsymbol{x}_t;\, \theta)). \qquad (2)$$

Manuscript received April 3, 2001; revised January 9, 2002. This work was supported by the Academy of Finland under Contract 53884.

The author is with the Helsinki Institute for Information Technology, Helsinki, Finland (e-mail: rissanen@hiit.fi).

Communicated by M. Weinberger, Associate Editor for Source Coding.
Digital Object Identifier 10.1109/TIT.2002.807281

The loss function defines a probability model, conditioned on $\boldsymbol{x}$, as follows:

$$p(y|\boldsymbol{x};\,\theta,\,\lambda) = Z^{-1}_{\lambda,\,\theta}(\boldsymbol{x})e^{-\lambda\delta(y,\,h(\boldsymbol{x};\,\theta))} \tag{3}$$

where λ is another positive real-valued parameter and $Z_{\lambda,\,\theta}(\boldsymbol{x})$ the normalizing constant

$$Z_{\lambda,\,\theta}(\boldsymbol{x}) = \int_{-\infty}^{\infty} e^{-\lambda\delta(y,\,h(\boldsymbol{x};\,\theta))}\,dy \tag{4}$$

assumed to exist. Extending this model to sequences by independence, we obtain for each λ a class

$$\mathcal{M}_{\lambda,\,k} = \{p(y^n|\boldsymbol{x}^n;\,\theta,\,\lambda)\colon \theta \in \Omega \subset R^k\}$$

of probability models

$$p(y^n|\boldsymbol{x}^n;\,\theta,\,\lambda) = Z^{-1}_{\lambda,\,\theta}(\boldsymbol{x}^n)e^{-\lambda L(y^n|\boldsymbol{x}^n;\,\theta)} \tag{5}$$

where

$$Z_{\lambda,\,\theta}(\boldsymbol{x}^n) = \prod_t Z_{\lambda,\,\theta}(\boldsymbol{x}_t). \tag{6}$$

Of particular interest to us are the loss functions, called *simple* in [6], where $Z_{\lambda,\,\theta}(\boldsymbol{x}) = Z_\lambda$ does not depend on the parameter θ nor on $\boldsymbol{x}$. For them the estimate $\hat\theta = \hat\theta(y^n,\,\boldsymbol{x}^n)$ that minimizes the loss (2) is the same as the maximum-likelihood estimate that minimizes the other loss function for this class, the ideal code length $\ln 1/p(y^n|\boldsymbol{x}^n;\,\theta,\,\lambda)$. The class of simple loss functions includes the loss for binary data which is 0 for no prediction error and unity otherwise. More importantly, this family also includes all loss functions of the form

$$\delta(y,\,\hat y) = |y-\hat y|^\alpha, \qquad \alpha > 0 \tag{7}$$

called α-loss functions, [13], for which a formula for the normalizing coefficient is given below. Notice that the important quadratic loss function is a special case, giving rise to the normal distribution, and so is the absolute value loss function for $\alpha = 1$, in which case (3) gives Laplace' distribution.

If the integral

$$\int_{\hat\theta(y^n,\,\boldsymbol{x}^n)\in\Omega^\circ} e^{-\lambda L(y^n|\boldsymbol{x}^n;\,\hat\theta(y^n,\,\boldsymbol{x}^n))}\,dy^n = B_{n,\,\lambda}(\boldsymbol{x}^n) \tag{8}$$

is finite, we can define the normalized maximum-likelihood (NML) model, [2], [10]

$$\hat p(y^n|\boldsymbol{x}^n;\,\lambda) = \frac{e^{-\lambda L(y^n|\boldsymbol{x}^n;\,\hat\theta(y^n,\,\boldsymbol{x}^n))}}{B_{n,\,\lambda}(\boldsymbol{x}^n)} \tag{9}$$

$$= \frac{p(y^n|\boldsymbol{x}^n;\hat\theta(y^n,\,\boldsymbol{x}^n),\,\lambda)}{C_{n,\,\lambda}(\boldsymbol{x}^n)} \tag{10}$$

where

$$C_{n,\,\lambda}(\boldsymbol{x}^n) = B_{n,\,\lambda}(\boldsymbol{x}^n)/Z^n_\lambda. \tag{11}$$

We derive next a few important properties of the models in the class $\mathcal{M}_{\lambda,\,k}$ for a simple loss function, which are shared by the exponential family of densities. First, by differentiating the integral (4) with respect to λ we get for all $\hat y$ of type $\hat y = h(\boldsymbol{x};\,\theta)$ and all positive λ

$$E_{\theta,\,\lambda}\delta(Y,\,\hat y) = -\dot Z_\lambda/Z_\lambda \tag{12}$$

where $\dot Z_\lambda = dZ_\lambda/d\lambda$ and $E_{\theta,\,\lambda}$ denotes the expectation with respect to $p(y|\boldsymbol{x};\,\theta,\,\lambda)$. That the order of the differentiation and integration may be switched can be seen by the very definition of the derivative. We also follow the custom to denote random variables by capital letters while using lower case letters for data strings. Further, let $\hat\lambda(y^n,\,\boldsymbol{x}^n,\,\theta)$ minimize the ideal code length

$$-\ln p(y^n|\boldsymbol{x}^n;\,\theta,\,\lambda) = \lambda L(y^n|\boldsymbol{x}^n;\,\theta) + n\ln Z_\lambda \tag{13}$$

and suppose that for all values of y^n and θ, the derivative with respect to λ vanishes at $\hat\lambda = \hat\lambda(y^n,\,\boldsymbol{x}^n,\,\theta)$

$$L(y^n|\boldsymbol{x}^n;\,\theta) + n\dot Z_{\hat\lambda}/Z_{\hat\lambda} = 0. \tag{14}$$

Then, because of (12)

$$L(y^n|\boldsymbol{x}^n;\,\theta) = E_{\theta,\,\hat\lambda}L(Y^n|\boldsymbol{x}^n;\,\theta) \tag{15}$$

where the expectation is taken for the fixed value $\hat\lambda = \hat\lambda(y^n,\,\boldsymbol{x}^n,\,\theta)$.

Next, differentiate the integral (8) with respect to λ. The result is for all λ

$$E_{\hat p(\lambda)}L(Y^n|\boldsymbol{x}^n;\,\hat\theta(Y^n,\,\boldsymbol{x}^n)) = -\dot B_{n,\,\lambda}(\boldsymbol{x}^n)/B_{n,\,\lambda}(\boldsymbol{x}^n) \tag{16}$$

where $\dot B_{n,\,\lambda}(\boldsymbol{x}^n) = dB_{n,\,\lambda}(\boldsymbol{x}^n)/d\lambda$ and the expectation is with respect to $\hat p(\lambda) = \hat p(y^n|\boldsymbol{x}^n;\,\lambda)$. Let $\bar\lambda = \bar\lambda(y^n,\,\boldsymbol{x}^n)$ denote the value of λ that minimizes

$$\lambda L(y^n|\boldsymbol{x}^n;\,\hat\theta(y^n,\,\boldsymbol{x}^n)) + \ln B_{n,\,\lambda}(\boldsymbol{x}^n) \tag{17}$$

and assume that at $\bar\lambda$

$$L(y^n|\boldsymbol{x}^n;\,\hat\theta(y^n,\,\boldsymbol{x}^n)) = -\dot B_{n,\,\bar\lambda}(\boldsymbol{x}^n)/B_{n,\,\bar\lambda}(\boldsymbol{x}^n) \tag{18}$$

for all values of y^n such that $\hat\theta(y^n,\,\boldsymbol{x}^n) \in \Omega^\circ$. Then

$$\begin{aligned} L(y^n|\boldsymbol{x}^n;\,\hat\theta(y^n,\,\boldsymbol{x}^n)) &= E_{\hat p(\bar\lambda)}L(Y^n|\boldsymbol{x}^n;\,\hat\theta(Y^n,\,\boldsymbol{x}^n)) \\ &= -\dot B_{n,\,\bar\lambda}(\boldsymbol{x}^n)/B_{n,\,\bar\lambda}(\boldsymbol{x}^n). \end{aligned} \tag{19}$$

To conclude this section, we show that the distributions $p = p(y|\boldsymbol{x};\,\theta,\,\lambda)$ are maximum-entropy distributions. The simple proof is similar to that given in [3, Ch. 11]. Consider the problem

$$\max_g E_g \ln 1/g(Y) \tag{20}$$

where the maximization is over all g such that

$$E_g\delta(Y,\,h(\boldsymbol{x};\,\theta)) \le E_p\delta(Y,\,h(\boldsymbol{x};\,\theta)).$$

We have first by this restriction on the density functions g

$$E_g \ln 1/p(Y|\boldsymbol{x};\,\theta,\,\lambda) = \lambda E_g\delta(Y,\,h(\boldsymbol{x};\,\theta)) + \ln Z_\lambda \le H(p)$$

where $H(p)$ denotes the entropy of $p(y|\boldsymbol{x};\,\theta,\,\lambda)$. Then, by Shannon's inequality, the entropy $H(g)$ of g satisfies

$$H(g) \le E_g \ln 1/p(Y|\boldsymbol{x};\,\theta,\,\lambda)$$

the right-hand side being upper-bounded by $H(p)$. The equality is reached with $g = p$. This result generalizes the familiar fact that the normal distribution with variance σ^2 has the maximum entropy among all distributions whose variance does not exceed σ^2.

III. Loss Complexity

In [11], we showed that the NML density function solves the min-max problem

$$\min_q \max_g E_g \ln \frac{p(Y^n|\boldsymbol{x}^n;\, \hat{\theta}(Y^n,\, \boldsymbol{x}^n),\, \lambda)}{q(Y^n)} \tag{21}$$

where q and g range over a wide class of distributions, in particular, not restricted to independent and identically distributed (i.i.d.) distributions, and the min-max value $\ln C_{n,\lambda}(\boldsymbol{x}^n)$, (11), is reached for $q = g = \hat{p}(y^n|\boldsymbol{x}^n;\, \lambda)$. Consider the analogous min-max problem

$$\min_f \max_g E_g[L_f(Y^n|\boldsymbol{x}^n) - L(Y^n|\boldsymbol{x}^n;\, \hat{\theta}(Y^n,\, \boldsymbol{x}^n))] \tag{22}$$

where

$$L_f(y^n|\boldsymbol{x}^n) = \sum_t \delta(y_t,\, \hat{y}_t) \tag{23}$$

and $\hat{y}_t$ is obtained either by a predictor $\hat{y}_t = f_t(y^{t-1},\, \boldsymbol{x}^n)$, $f_1(y^0,\, \boldsymbol{x}^n) = 0$, or a more general *estimator* function $\hat{y}_t = f_t(y^n,\, \boldsymbol{x}^n)$. Specifically, we consider estimator functions of the form $\hat{y} = f(\boldsymbol{x};\, \eta)$, where the parameter η, ranging over a subset $\Psi \subseteq R^m$, has m components, $m \le n$. With $\bar{\eta}(y^n,\, \boldsymbol{x}^n)$ denoting any estimate of the parameter we then write

$$\hat{y}_t = f(\boldsymbol{x}_t;\, \bar{\eta}(y^n,\, \boldsymbol{x}^n)). \tag{24}$$

Denote by $\mathcal{F}$ the set of all such estimator functions f.

We have

$$\begin{aligned} &\lambda E_g L_f(Y^n|\boldsymbol{x}^n) + \ln B_{f,n,\lambda}(\boldsymbol{x}^n) \\ &\qquad - (\lambda E_g L(Y^n|\boldsymbol{x}^n;\, \hat{\theta}(Y^n,\, \boldsymbol{x}^n)) + \ln B_{n,\lambda}(\boldsymbol{x}^n)) \qquad (25)\\ &\quad = D(g\|p_f) - D(g\|\hat{p}(\lambda)) \qquad (26)\end{aligned}$$

where $D(\cdot\|\cdot)$ denotes the Kullback–Leibler distance

$$B_{f,n,\lambda}(\boldsymbol{x}^n) = \begin{cases} Z_\lambda^n, & \text{for } f \text{ a predictor} \\ C_{f,n,\lambda}(\boldsymbol{x}^n) Z_\lambda^n, & \text{for } f \text{ an estimator} \end{cases} \tag{27}$$

and

$$C_{f,n,\lambda}(\boldsymbol{x}^n) = Z_\lambda^{-n} \int_{\bar{\eta}(y^n,\, \boldsymbol{x}^n) \in \Psi} e^{-\lambda L_f(y^n|\boldsymbol{x}^n)}\, dy^n \tag{28}$$

assumed to be finite; ψ° denotes the interior of ψ. Further, $\hat{p}(\lambda) = \hat{p}(y^n|\boldsymbol{x}^n;\, \lambda)$, (10), and

$$p_f = \frac{e^{-\lambda L_f(y^n|\boldsymbol{x}^n)}}{B_{f,n,\lambda}(\boldsymbol{x}^n)}. \tag{29}$$

The role of the data generating distributions g is to model the statistical restrictions in the data, all of which may not be captured by the models in the class $\mathcal{M}_{\lambda,k}$. Hence, we should not restrict the distributions g to the set $\mathcal{M}_{\lambda,k}$. However, to obtain stronger inequality bounds we should not let them be just any distributions, and in light of (16), we restrict them to the set

$$\begin{aligned} G(\lambda,\, \boldsymbol{x}^n) = \Big\{ g\colon & E_g L(Y^n|\boldsymbol{x}^n;\, \hat{\theta}(Y^n,\, \boldsymbol{x}^n)) \\ & \le -\dot{B}_{n,\lambda}(\boldsymbol{x}^n)/B_{n,\lambda}(\boldsymbol{x}^n) \Big\} \end{aligned} \tag{30}$$

where the right-hand side of the inequality by (16) equals $E_{\hat{p}(\lambda)} L(Y^n|\boldsymbol{x}^n;\, \hat{\theta}(Y^n,\, \boldsymbol{x}^n))$. This restriction for the data generating distributions is quite natural if we bear in mind the role of the parameter λ, which in case of the normal density functions is inversely proportional to the variance; it is characteristic of the size of the typical loss functions modeled for the data. For a large value of λ, the bulk of the probability mass modeled is for data with small loss and *vice versa*. Hence, in the set $G(\lambda,\, \boldsymbol{x}^n)$ we want to include only the data generating density functions that we consider to be "relevant" for the data we have selected to model. This is also analogous to the restrictions taken in the maximum entropy problem (20), where the data generating distributions are restricted not to exceed the mean loss.

We have the following theorem.

Theorem 1: For all positive λ and for all estimators $f \in \mathcal{F}$ of the form (24),

$$\begin{aligned} &\max_{g \in G(\lambda,\, \boldsymbol{x}^n)} E_g L_f(Y^n|\boldsymbol{x}^n) + \lambda^{-1} \ln C_{f,n,\lambda}(\boldsymbol{x}^n) \\ &\qquad \ge E_{\hat{p}(\lambda)} L(Y^n|\boldsymbol{x}^n;\, \hat{\theta}(Y^n,\, \boldsymbol{x}^n)) + \lambda^{-1} \ln C_{n,\lambda}(\boldsymbol{x}^n). \end{aligned} \tag{31}$$

The equality is reached for

$$f(\boldsymbol{x}_t;\, \bar{\eta}(y^n,\, \boldsymbol{x}^n)) = h(\boldsymbol{x}_t;\, \hat{\theta}(y^n,\, \boldsymbol{x}^n)).$$

For all positive λ and all predictors of the form $\hat{y}_t = f_t(y^{t-1},\, \boldsymbol{x}^n)$, $f_1(y^0,\, \boldsymbol{x}^n) = 0$

$$\begin{aligned} \max_{g \in G(\lambda,\, \boldsymbol{x}^n)} E_g L_f(Y^n|\boldsymbol{x}^n) &\ge E_{\hat{p}(\lambda)} L(Y^n|\boldsymbol{x}^n;\, \hat{\theta}(Y^n,\, \boldsymbol{x}^n)) \\ &\quad + \lambda^{-1} \ln C_{n,\lambda}(\boldsymbol{x}^n). \end{aligned} \tag{32}$$

Proof: By (16) $\hat{p}(\lambda)$ is in $G(\lambda,\, \boldsymbol{x}^n)$, and

$$\max_{g \in G(\lambda,\, \boldsymbol{x}^n)} E_g L_f(Y^n|\boldsymbol{x}^n) \ge E_{\hat{p}(\lambda)} L_f(Y^n|\boldsymbol{x}^n). \tag{33}$$

In view of (26), (11), and (27), the inequality (31) is equivalent with

$$\max_{g \in G(\lambda,\, \boldsymbol{x}^n)} \{D(g\|p_f) - D(g\|\hat{p}(\lambda))\} \ge 0. \tag{34}$$

Since $\hat{p}(\lambda)$ is in $G(\lambda,\, \boldsymbol{x}^n)$, the left-hand side is lower-bounded by $D(\hat{p}(\lambda)\|p_f)$, which is nonnegative. Clearly, the equality is reached for $p_f = \hat{p}(\lambda)$.

Further, for predictors

$$\begin{aligned} D(\hat{p}(\lambda)\|p_f) = \lambda E_{\hat{p}(\lambda)}[&L_f(Y^n|\boldsymbol{x}^n) \\ &- L(Y^n|\boldsymbol{x}^n;\, \hat{\theta}(Y^n,\, \boldsymbol{x}^n))] - \ln C_{n,\lambda}(\boldsymbol{x}^n) \ge 0 \end{aligned} \tag{35}$$

which gives the claim. □

An indication of the loss on the data at hand is $L(y^n|\boldsymbol{x}^n;\, \hat{\theta}(y^n,\, \boldsymbol{x}^n))$, and by picking λ in view of (19) as $\bar{\lambda}$, the theorem gives lower bounds which are relevant for the observed data; the set $G(\lambda, \boldsymbol{x}^n)$ gets replaced by

$$\hat{G}(y^n, \boldsymbol{x}^n) = \Big\{g\colon E_g L(Y^n|\boldsymbol{x}^n; \hat{\theta}(Y^n, \boldsymbol{x}^n)) \le L(y^n|\boldsymbol{x}^n; \hat{\theta}(y^n, \boldsymbol{x}^n))\Big\}. \quad (36)$$

In view of Theorem 1 we define

$$I_\lambda(y^n|\boldsymbol{x}^n) = L(y^n|\boldsymbol{x}^n; \hat{\theta}(y^n, \boldsymbol{x}^n)) + \frac{1}{\lambda}\ln C_{n,\lambda}(\boldsymbol{x}^n) \quad (37)$$

to be the *loss complexity* of the data $(y^n, \boldsymbol{x}^n)$, relative to the model class $\mathcal{M}_{k,\lambda}$. Similarly, relative to the model class

$$\mathcal{M}_k = \{p(y^n|\boldsymbol{x}^n; \theta, \lambda)\colon \theta \in \Omega, \lambda > 0\} \quad (38)$$

the loss complexity is defined to be

$$I(y^n|\boldsymbol{x}^n) = L(y^n|\boldsymbol{x}^n; \hat{\theta}(y^n, \boldsymbol{x}^n)) + \frac{1}{\bar{\lambda}}\ln C_{n,\bar{\lambda}}. \quad (39)$$

Remarks: The term $\ln C_{n,\lambda}(\boldsymbol{x}^n)$ in the loss complexity (37) can be interpreted as the logarithm of the number of "distinguishable" models [1]; i.e., models that can be distinguished from data $\boldsymbol{x}^n$ in such a manner that the probability of error goes to zero as $n \to \infty$; see also the discussion in [11]. Hence, it may be viewed as the code length for the parameter $\hat{\theta}$, in effect optimally quantized, needed to implement the optimal model. The parameter λ provides a weight by which the code length for the model is converted into loss, $\bar{\lambda}$ being the optimal weight. In [13], the extended stochastic complexity, too, was shown to admit an asymptotic expansion as the sum of the minimized accumulated loss and a term which is upper-bounded by an explicitly calculated code length for the parameters, written to an optimal precision and weighted by $1/\lambda$. This makes the extended stochastic complexity and $I_\lambda(y^n|\boldsymbol{x}^n)$ close for long data strings. In [13], no explicitly optimized value for λ as a function of the data string was determined. Rather, it was replaced by an asymptotic expression.

In case of the logarithmic loss, the interpretation of the stochastic complexity as the sum of the negative logarithm of the maximum likelihood and the ideal code length for the parameters is natural, because both are code lengths and hence expressed in the same units. Moreover, since the code length for the parameters must satisfy the Kraft inequality any such two-part code length provides a natural requirement for the estimator functions $\hat{y}_t = f_t(y^n, \boldsymbol{x}^n)$ to be admissible in providing a fair comparison of the losses; the density function they define must integrate to unity. In case of a nonlogarithmic loss function there is no obvious normalization requirement for the estimator function, and one may wonder why not permit the "perfect" estimator defined by $y_t = \hat{y}_t$, which gives zero loss. On intuitive grounds it is reasonable to demand for a fair loss comparison that the estimator function must be described in a decodable manner, but adding the code length of the parameters to a non-code length loss appears arbitrary. The inequality (31) provides the required normalization, and we call any estimator *realizable* and its accumulated mean loss, the left-hand side of (31), *achievable*, if it satisfies this inequality.

As a final comment, we mention that the inequalities in Theorem 1 hold not only for the worst case data generating distribution but for almost all distributions in the family $\mathcal{M}_k$. We prove this later for the α-loss functions, which are analyzed in detail in the following section.

IV. α-Loss Functions

We begin by giving an explicit formula for the normalizing coefficient for α-loss functions, obtained from the integral [5]

$$\int_0^\infty e^{-t}t^{\alpha-1}\,dt = \Gamma(\alpha)$$

with the change of variables $t = \lambda u^\alpha$

$$Z_{\alpha,\lambda} = \int_{-\infty}^{\infty} e^{-\lambda|u|^\alpha}\,du = \int_{-\infty}^{\infty} e^{-\lambda|y-\mu|^\alpha}\,dy = \frac{2}{\alpha\lambda^{1/\alpha}}\Gamma(1/\alpha). \quad (40)$$

Here, $\Gamma(\alpha)$ is the gamma function. With such models we write $\mu = \hat{y} = h(\boldsymbol{x}, \theta)$, where h is a function vanishing at $\theta = 0$. In order to simplify matters we take this function as the inner product $\theta'\boldsymbol{x}_t$. A generalization to other functions is possible, but for our results it would require assumptions that make them behave like the inner product. Extend the so-defined density function

$$p_\alpha(y|\boldsymbol{x}; \theta, \lambda) = Z_{\alpha,\lambda}^{-1}e^{-\lambda|y-\theta'\boldsymbol{x}|^\alpha}$$

to sequences by independence with the result

$$p_\alpha(y^n|\boldsymbol{x}^n; \theta, \lambda) = Z_{\alpha,\lambda}^{-n}e^{-\lambda\sum_t|y_t-\theta'\boldsymbol{x}_t|^\alpha} \quad (41)$$

and denote the family of such models as

$$\mathcal{M}_{\alpha,\lambda,k} = \{p_\alpha(y^n|\boldsymbol{x}^n; \theta, \lambda)\colon \theta \in \Omega \subseteq R^k\} \quad (42)$$

where Ω is a compact subset of R^k with nonempty interior Ω°.

For α-loss functions, the mean loss (12) can be evaluated for all μ as

$$E_{\alpha,\theta,\lambda}|Y-\mu|^\alpha = \frac{1}{\alpha\lambda} \quad (43)$$

where the expectation is with respect to $p_\alpha(y|\boldsymbol{x}; \theta, \lambda)$.

Denoting the normalizing constant (11) by $C_{n,\alpha,\lambda}(\boldsymbol{x}^n)$, we have with the restriction $1 \le \alpha \le 2$ an accurate asymptotic formula for it, which permits calculation of the loss complexity. The result is in the theorem proved in Appendix A.

Theorem 2: For the model class $\mathcal{M}_{\alpha,\lambda,k}$ let Ω be a closed bounded subset of R^k with nonempty interior Ω°. Further let

$$n^{-1}\sum_{t=1}^n \boldsymbol{x}_t\boldsymbol{x}_t' \to \Sigma > 0 \quad (44)$$

as $n \to \infty$. Then for all positive λ and α in the interval $1 \leq \alpha \leq 2$

$$\ln C_{n,\alpha,\lambda}(\boldsymbol{x}^n) = \begin{cases} \frac{k}{2}\ln\frac{n\lambda}{\pi} + \ln(|\Sigma|^{1/2}|\Omega|) + o(1), & \text{for } \alpha = 1 \\ \frac{k}{2}\ln\left[\frac{n\alpha(\alpha-1)}{2\pi}\,\lambda^{2/\alpha} \frac{\Gamma(1-1/\alpha)|\Sigma|^{1/2}|\Omega|}{\Gamma(1/\alpha)}\right] + o(1), & \text{for } \alpha > 1 \end{cases} \tag{45}$$

where $|\Omega|$ denotes the volume of Ω.

Remark: The condition (44) is typical for regression problems, where the rows of the matrix Σ are defined by the inner products of the basis vectors in an infinite-dimensional space such as the polynomials, sinusoidals in Fourier series, or wavelets.

We recall the definition of $\bar{\lambda} = \bar{\lambda}(y^n, \boldsymbol{x}^n)$ as the value of λ that minimizes (17). By Theorem 2, it is clear that it is close to the value, say $\tilde{\lambda} = \tilde{\lambda}(y^n, \boldsymbol{x}^n)$, that minimizes

$$\lambda L(y^n|\boldsymbol{x}^n;\, \hat{\theta}(y^n, \boldsymbol{x}^n)) + \ln B'_{n,\alpha,\lambda}(\boldsymbol{x}^n) \tag{46}$$

where $B'_{n,\alpha,\lambda}(\boldsymbol{x}^n) = Z_\lambda^n C'_{n,\alpha,\lambda}(\boldsymbol{x}^n)$ and $C'_{n,\alpha,\lambda}(\boldsymbol{x}^n)$ denotes the normalizing coefficient $C_{n,\alpha,\lambda}(\boldsymbol{x}^n)$ in Theorem 2 without the remainder term $o(1)$. In fact, we show in Appendix B that

$$\bar{\lambda} = \tilde{\lambda} + \delta_n \tag{47}$$

where $|\delta_n| \leq o(1/\sqrt{n})$.

Our main theorem is as follows.

Theorem 3: Let $L(y^n|\boldsymbol{x}^n;\, \hat{\theta})$ be positive, and let

$$\tilde{\lambda} = (n-k)/(\alpha L(y^n|\boldsymbol{x}^n;\, \hat{\theta}))$$

where $\hat{\theta}$ denotes the fixed parameter $\hat{\theta}(y^n, \boldsymbol{x}^n)$. Then for all α in the interval $1 \leq \alpha \leq 2$ and for all estimators (24)

$$\max_{g \in \hat{G}(y^n, \boldsymbol{x}^n)} E_g L_f(Y^n|\boldsymbol{x}^n)) + \bar{\lambda}^{-1} \ln C_{f,n,\alpha,\bar{\lambda}}(\boldsymbol{x}^n) \geq L(y^n|\boldsymbol{x}^n;\, \hat{\theta})\left[1 + \frac{\alpha}{n-k}\ln C_{n,\alpha,\bar{\lambda}}(\boldsymbol{x}^n)\right] + o\big((\ln n)/\sqrt{n}\,\big) \tag{48}$$

where

$$C_{f,n,\alpha,\bar{\lambda}}(\boldsymbol{x}^n) = Z_{\bar{\lambda}}^{-n} \int_{\bar{\eta}(y^n,\boldsymbol{x}^n) \in \Psi^\circ} e^{-\bar{\lambda} L_f(y^n|\boldsymbol{x}^n)}\, dy^n$$

finite or not, and $\hat{G}(y^n, \boldsymbol{x}^n)$ is defined as

$$\hat{G}(y^n, \boldsymbol{x}^n) = \left\{g\colon E_g L(Y^n|\boldsymbol{x}^n;\, \hat{\theta}(Y^n, \boldsymbol{x}^n)) \leq L(y^n|\boldsymbol{x}^n;\, \hat{\theta})\right\}. \tag{49}$$

The equality is reached for $f(\boldsymbol{x}_t;\, \bar{\eta}(y^n, \boldsymbol{x}^n)) \equiv \boldsymbol{x}'\hat{\theta}(y^n, \boldsymbol{x}^n)$.

Further, for all predictors $\hat{y}_{t+1} = f(y^t, \boldsymbol{x}^n)$ and all y^n

$$\max_{g \in \hat{G}(y^n, \boldsymbol{x}^n)} E_g L_f(Y^n|\boldsymbol{x}^n) \geq L(y^n|\boldsymbol{x}^n;\, \hat{\theta}) \cdot \left[1 + \frac{\alpha}{n-k}\ln C_{n,\alpha,\bar{\lambda}}\right] + o\left((\ln n)/\sqrt{n}\right). \tag{50}$$

We emphasize again that the data generating distributions g are not restricted to i.i.d. distributions.

Proof: From Theorem 1 with (19)

$$\max_{g \in \hat{G}(\bar{\lambda}, \boldsymbol{x}^n)} E_g L_f(Y^n|\boldsymbol{x}^n)) + \bar{\lambda}^{-1} \ln C_{f,n,\alpha,\bar{\lambda}}(\boldsymbol{x}^n) \geq L(y^n|\boldsymbol{x}^n;\, \hat{\theta}) + \bar{\lambda}^{-1} \ln C_{n,\alpha,\bar{\lambda}}(\boldsymbol{x}^n). \tag{51}$$

We need to express $\bar{\lambda}^{-1}$ in terms of $L(y^n|\boldsymbol{x}^n;\, \hat{\theta}(y^n, \boldsymbol{x}^n))$. This is done with (47), where $\tilde{\lambda}$ is written in terms of $L(y^n|\boldsymbol{x}^n;\, \hat{\theta}(y^n, \boldsymbol{x}^n))$. With the behavior of $\ln C_{n,\alpha,\bar{\lambda}}(\boldsymbol{x}^n)$ from Theorem 2 we get the claim (48). The rest follows in a straightforward way. □

By providing the reachable lower bound for estimation the loss complexity with Theorem 2 provides a criterion for selection of model classes

$$\min_k \left\{ L(y^n|\boldsymbol{x}^n;\, \hat{\theta})\left[1 + \frac{\alpha}{n-k}\ln C_{n,\alpha,\bar{\lambda}}(\boldsymbol{x}^n)\right]\right\}. \tag{52}$$

We show next that the worst case bound for predictors in Theorem 1 is not an isolated case, for the same bound in effect holds even when the mean is taken with respect to most of the data generating distributions $p_{\alpha,\theta,\lambda} = p_\alpha(y^n|\boldsymbol{x}^n;\, \theta, \lambda)$, (41). We also have an easy generalization of the inequality for the mean quadratic prediction error for Gaussian processes in [9]; a somewhat different generalization of the quadratic error bound is in [8, eq. (32)].

Theorem 4: Let $\hat{y}_t = f_t(y^{t-1}, \boldsymbol{x}^n)$ be any predictor. Then for all α in the interval $[1, 2]$ and all positive ϵ, the inequality

$$\frac{1}{n} E_{p_{\alpha,\theta,\lambda}} L_f(Y^n|\boldsymbol{x}^n) \geq \frac{1}{\alpha\lambda}\left(1 + (k-\epsilon)\frac{\alpha}{2n}\ln n\right) \tag{53}$$

holds for n large enough and for all $\theta \in \Omega$, except in a set whose volume goes to zero as n grows to infinity.

Proof: Consider

$$E_{p_{\alpha,\theta,\lambda}} \ln \frac{p_{\alpha,\theta,\lambda}}{p_f} = \lambda E_{p_{\alpha,\theta,\lambda}}(L_f(Y^n|\boldsymbol{x}^n) - L(Y^n|\boldsymbol{x}^n;\, \theta)).$$

As stated at the end of Appendix A, the Central Limit Theorem holds for the family $\{p_\alpha(y^n|\boldsymbol{x}^n;\, \theta, \lambda)\}$, which implies the condition required for [9, Theorem 1] to hold. Hence, the right-hand side exceeds $\frac{k-\epsilon}{2}\ln n$ with the quantifications given. With (43) we get (53). □

The question remains of how tight the lower bound in Theorem 3 is for prediction. This is tantamount to the question whether the mean stochastic complexity can be reached predictively in an asymptotic sense. The lower bound can be shown to be reached asymptotically for $\alpha = 2$ in the case where the data generating model is in the class $\mathcal{M}_{2,\lambda}$, because then

$$E_{p_{2,\theta,\lambda}}(\hat{\theta}(Y^n, \boldsymbol{x}^n) - \theta)'(\hat{\theta}(Y^n, \boldsymbol{x}^n) - \theta) = O(1/n).$$

In general, however, the problem appears to be more difficult and we settle here for an example. The reachability of the lower bound in the almost sure sense for Gaussian autoregressive (AR) processes was shown in [7]; a good survey of predictive coding for a number of loss functions other than the α-types is [8].

Example: Let $\alpha = 2$ and take the predictor as the arithmetic mean of the past data

$$\hat{y}_{t+1} = f_t(y^t, \boldsymbol{x}^n) = \frac{1}{t} \sum_{i=1}^{t} y_i = \bar{y}_t.$$

Write

$$L_f(y^n) = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2$$

and

$$L(y^n; \bar{y}_n) = \sum_{t=1}^{n} (y_t - \bar{y}_n)^2.$$

We have the identity

$$L_f(y^n) = L(y^n; \bar{y}_n) + \sum_{t=1}^{n} \left[\sum_{s=1}^{t} \left(y_s - \frac{1}{t-1} \sum_{1}^{t-1} y_i \right)^2 - \sum_{s=1}^{t} \left(y_s - \frac{1}{t} \sum_{1}^{t} y_i \right)^2 \right].$$

Let $E_g Y_i = \mu$ and $E_g (Y_i - \mu)^2 = \sigma^2$. Then

$$\begin{aligned} E_g L_f(Y^n) = {} & E_g L(Y^n; \bar{Y}_n) \\ & + \sum_{t=1}^{n} \left\{ E_g \sum_{s=1}^{t} \left[(Y_s - \mu) - \frac{1}{t-1} \sum_{1}^{t-1} (Y_i - \mu) \right]^2 \right. \\ & \left. - E_g \sum_{s=1}^{t} \left[(Y_s - \mu) - \frac{1}{t} \sum_{1}^{t} (Y_i - \mu) \right]^2 \right\}. \end{aligned}$$

Further

$$\begin{aligned} E_g \sum_{s=1}^{t} & \left[(Y_s - \mu) - \frac{1}{t} \sum_{1}^{t} (Y_i - \mu) \right]^2 \\ & = t\sigma^2 \left[(1 - 1/t)^2 + \frac{t-1}{t^2} \right] = (t-1)\sigma^2 \qquad (54) \end{aligned}$$

and for $t > 1$

$$\begin{aligned} E_g \sum_{s=1}^{t} & \left(Y_s - \mu - \frac{1}{t-1} \sum_{1}^{t-1} (Y_i - \mu) \right)^2 \\ & = E_g \left[(Y_t - \mu) - \frac{1}{t-1} \sum_{1}^{t-1} (Y_i - \mu) \right]^2 + (t-2)\sigma^2 \\ & = \left(\frac{1}{t-1} + t - 1 \right) \sigma^2. \end{aligned}$$

Hence,

$$\begin{aligned} E_g L_f(Y^n) &= E_g L(Y^n; \bar{Y}_n) + \sigma^2 \sum_{t=2}^{n} \frac{1}{t-1} \\ &= E_g L(Y^n; \bar{Y}_n) + \sigma^2 \ln n + O(1). \end{aligned}$$

By (54) $E_g L(Y^n; \bar{Y}_n) = (n-1)\sigma^2$, and by putting $\sigma^2 = (1/n) \sum_1^n (y_t - \bar{y}_t)^2$ we see with the formula for $\ln C_{n,\alpha,\lambda}(\boldsymbol{x}^n)$ in Theorem 2 that the lower bound in Theorem 4 is reached asymptotically to within a constant.

Appendix A
Evaluation of $C_{n,\alpha,\lambda}$

In [10], conditions were given under which the quite accurate asymptotic formula for the normalizing coefficient for the NML density function, in our case $C_{n,\alpha,\lambda}$, holds

$$C_{n,\alpha,\lambda} = \frac{k}{2} \ln \frac{n}{2\pi} + \ln \int_{\Omega} \sqrt{|I(\theta)|}\, d\theta + o(1) \qquad (55)$$

where $|I(\theta)|$ is the Fisher information to be given below. The validity of the formula requires conditions on the density functions of the family considered, all involving at least two times differentiability with respect to the parameters. The α-loss functions are not even once differentiable at the origin for all values of α of interest such as $\alpha = 1$. We overcome this obstacle by considering the modified loss functions as follows:

$$L_{\epsilon,\alpha}(e) = \begin{cases} e^{\alpha}, & \text{for } e \geq \epsilon \\ a'e^2 + b'e^3 + c'e^4 + d'e^5, & \text{for } 0 \leq e \leq \epsilon \end{cases} \qquad (56)$$

where ϵ is a parameter taking small values, in the limit even zero. Because of symmetry it is enough to let e range through nonnegative values only. The coefficients of the fifth degree polynomial required are determined such that at $e = \epsilon$ the two pieces of the function $L_{\epsilon,\alpha}(e)$ have equal values including the first three derivatives. This gives for the scaled coefficients $a = a'\epsilon^{2-\alpha}$, $b = b'\epsilon^{3-\alpha}$, $c = c'\epsilon^{4-\alpha}$, and $d = d'\epsilon^{5-\alpha}$ the equations

$$\begin{aligned} 1 &= a + b + c + d \\ \alpha &= 2a + 3b + 4c + 5d \\ \alpha(\alpha - 1) &= 2a + 6b + 12c + 20d \\ \alpha(\alpha - 1)(\alpha - 2) &= 6b + 24c + 60d. \end{aligned} \qquad (57)$$

The solution is given by

$$\begin{aligned} a &= (3 - \alpha) \left[1 + \frac{1}{6} (2 - \alpha)(7 - \alpha) \right] \\ b &= -(2 - \alpha)[1 + 0.5(3 - \alpha)(6 - \alpha)] \\ c &= 0.5(2 - \alpha)(3 - \alpha)(5 - \alpha) \\ d &= \frac{-1}{6} (2 - \alpha)(3 - \alpha)(4 - \alpha) \end{aligned}$$

and the polynomial part in (56) by

$$f_{\epsilon,\alpha}(e) = a\epsilon^{\alpha} \left(\frac{e}{\epsilon} \right)^2 + b\epsilon^{\alpha} \left(\frac{e}{\epsilon} \right)^3 + c\epsilon^{\alpha} \left(\frac{e}{\epsilon} \right)^4 + d\epsilon^{\alpha} \left(\frac{e}{\epsilon} \right)^5. \qquad (58)$$

By a direct evaluation of $f_{\epsilon,\alpha}(e)$ for a number of different values of α in the interval $0 \leq \alpha \leq 3$ we verified that it is nonnegative and zero only at $y = 0$, which fact, however, is not required for our analysis.

Consider then the class of density functions

$$\mathcal{M}_{\alpha,\lambda,\epsilon} = \{ p_{\epsilon}(y^n | \boldsymbol{x}^n; \theta, \lambda) \colon \theta \in \Omega \}$$

where

$$p_{\epsilon}(y^n | \boldsymbol{x}^n; \theta, \lambda) = Z_{\alpha,\lambda,\epsilon}^{-n} e^{-\lambda \sum_t L_{\epsilon,\alpha}(y_t - \theta' \boldsymbol{x}_t)}$$

and Ω, as before, is a closed and bounded subset of R^k with Ω° as its nonempty interior. Notice that we restrict the predictor function $h(\boldsymbol{x};\,\theta)=\theta'\boldsymbol{x}$ to be linear, which is not unreasonable.

The conditions in [10], under which the expansion (55) holds, are for the model class $\mathcal{M}_{\alpha,\,\lambda,\,\epsilon}$ as follows.

Conditions:

1) The elements $\frac{\lambda}{n}\,\partial^2\sum_t L_{\epsilon,\,\alpha}(y_t-\theta'\boldsymbol{x}_t)/\partial\theta_i\partial\theta_j$ defining the matrix $I(y^n,\,\theta,\,\epsilon)$ are continuous in Ω°, and

$$\begin{aligned} I_n(\theta,\,\epsilon) &= EI(Y^n,\,\theta,\,\epsilon)\\ &= \frac{\lambda}{n}\left\{E\,\frac{\partial^2\sum\limits_t L_{\epsilon,\,\alpha}(Y_t-\theta'\boldsymbol{x}_t)}{\partial\theta_i\partial\theta_j}\right\}\to I(\theta,\,\epsilon)\end{aligned} \tag{59}$$

where the expectation is with respect to $p_\epsilon(y^n|\boldsymbol{x}^n;\,\theta,\,\lambda)$. Moreover, the limit satisfies $0<c_1\le|I(\theta,\,\epsilon)|\le c_2$ for all $\theta\in\Omega^o$, and

$$\int_{\Omega^o}\sqrt{|I(\theta,\,\epsilon)|}\,d\theta<\infty. \tag{60}$$

2) The maximum-likelihood estimator satisfies the Central Limit Theorem: The distribution of $\xi=\sqrt{n}(\hat\theta(\boldsymbol{x}^n)-\theta)$ converges to the normal distribution with mean zero and covariance $I^{-1}(\theta,\,\epsilon)$ for $\theta\in\Omega^o$. (Because of the compactness of Ω the requirement in [10] that the convergence is uniform is not needed nor was used in the proof.)

3) Finally

$$\begin{aligned} &I(y^n,\hat\theta,\epsilon)\\ &\quad=\frac{\lambda}{n}\left\{\frac{\partial^2\sum\limits_t L_{\epsilon,\,\alpha}(y_t-\theta'\boldsymbol{x}_t)}{\partial\theta_i\,\partial\theta_j}\right\}_{\theta=\hat\theta}<C_0<\infty\end{aligned} \tag{61}$$

where C_0 is a positive-definite matrix, and y^n is such that the maximum-likelihood estimate $\hat\theta=\hat\theta(y^n,\,\boldsymbol{x}^n)\in\Omega^o$. In addition, and most stringently, the family of the elements

$$I_{ij}(y^n,\theta(\xi),\epsilon)=\frac{\lambda}{n}\,\partial^2\sum_t L_{\epsilon,\,\alpha}(y_t-\theta'\boldsymbol{x}_t)\Big/\partial\xi_i\,\partial\xi_j \tag{62}$$

for $n\ge1$, $1\le i,\,j\le k$, as a function of the normalized variable ξ, where $\theta(\xi)=\hat\theta(y^n,\,\boldsymbol{x}^n)+\xi/\sqrt{n}$, is equicontinuous at $\xi=0$.

We verify that the members of the model family $\mathcal{M}_{\alpha,\,\lambda,\,\epsilon}$ satisfy these conditions for $1\le\alpha\le2$, and we prove Theorem 2 for the original family $\mathcal{M}_{\alpha,\,\lambda}$. In the proof, we use repeatedly the formulas [5]

$$\int_x^\infty e^{-t}t^{\alpha-1}\,dt=\Gamma(\alpha)-\gamma(\alpha,\,x) \tag{63}$$

$$\begin{aligned}\gamma(\alpha,\,x)&=\int_0^\infty e^{-t}t^{\alpha-1}\,dt\\ &=\frac{x^\alpha}{\alpha}-\frac{x^{\alpha+1}}{\alpha+1}+\frac{x^{\alpha+2}}{2(\alpha+2)}-\cdots.\end{aligned} \tag{64}$$

We evaluate first the normalizing constant

$$Z_{\alpha,\,\lambda,\,\epsilon}=2\int_0^\epsilon e^{-\lambda f(y)}\,dy+2\int_\epsilon^\infty e^{-\lambda y^\alpha}\,dy \tag{65}$$

where the subindexes in the polynomial $f_{\epsilon,\,\alpha}(y)$ are dropped. By expanding

$$e^{-\lambda f(y)}=1-\lambda f(y)+\tfrac{1}{2}\,\lambda^2f^2(y)-\cdots \tag{66}$$

we get with (58)

$$\int_0^\epsilon e^{-\lambda f(y)}\,dy=\epsilon+O(\epsilon^{1+\alpha}). \tag{67}$$

In the second integral in (65) put $t=\lambda y^\alpha$ and apply (64) and (63) to get

$$\begin{aligned}\int_\epsilon^\infty e^{-\lambda y^\alpha}\,dy&=Z_{\alpha,\,\lambda}-\frac{1}{\alpha\lambda^{1/\alpha}}\,\gamma(1/\alpha,\,\lambda\epsilon^\alpha)\\ &=Z_{\alpha,\,\lambda}-\epsilon+O(\epsilon^{1+\alpha}).\end{aligned}$$

Hence, with (40)

$$Z_{\alpha,\,\lambda,\,\epsilon}=Z_{\alpha,\,\lambda}+O(\epsilon^{1+\alpha})=\frac{2}{\alpha\lambda^{1/\alpha}}\,\Gamma(1/\alpha)+O(\epsilon^{1+\alpha}). \tag{68}$$

In order to verify Conditions 1)–3), we need to evaluate the first three derivatives of $L_{\epsilon,\,\alpha}(y-\theta'\boldsymbol{x})$. For $y-\theta'\boldsymbol{x}\ge\epsilon$ they are as follows:

$$\frac{\partial L_{\epsilon,\,\alpha}(y-\theta'\boldsymbol{x})}{\partial\theta}=-\lambda\alpha(y-\theta'\boldsymbol{x})^{\alpha-1}\boldsymbol{x} \tag{69}$$

$$\left\{\frac{\partial^2L_{\epsilon,\,\alpha}(y-\theta'\boldsymbol{x})}{\partial\theta_i\,\partial\theta_j}\right\}=\lambda\alpha(\alpha-1)(y-\theta'\boldsymbol{x})^{\alpha-2}\boldsymbol{x}\boldsymbol{x}' \tag{70}$$

$$\begin{aligned}\frac{\partial^3L_{\epsilon,\,\alpha}(y-\theta'\boldsymbol{x})}{\partial\theta_i\,\partial\theta_j\,\partial\theta_k}&=-\lambda\alpha(\alpha-1)(\alpha-2)\\ &\qquad\cdot(y-\theta'\boldsymbol{x})^{\alpha-3}x_ix_jx_k.\end{aligned} \tag{71}$$

For $0\le e=y-\theta'\boldsymbol{x}\le\epsilon$ they are

$$\frac{\partial L_{\epsilon,\,\alpha}(e)}{\partial\theta}=-\lambda\dot f(e)\boldsymbol{x} \tag{72}$$

$$\left\{\frac{\partial^2L_{\epsilon,\,\alpha}(y-\theta'\boldsymbol{x})}{\partial\theta_i\,\partial\theta_j}\right\}=\lambda\ddot f(e)\boldsymbol{x}\boldsymbol{x}' \tag{73}$$

$$\frac{\partial^3L_{\epsilon,\,\alpha}(y-\theta'\boldsymbol{x})}{\partial\theta_i\,\partial\theta_j\,\partial\theta_k}=-\lambda d^3f(e)/(de)^3x_ix_jx_k \tag{74}$$

where

$$\dot f(e)=2a\epsilon^{\alpha-2}e+3b\epsilon^{\alpha-3}e^2+4c\epsilon^{\alpha-4}e^3+5d\epsilon^{\alpha-5}e^4 \tag{75}$$

$$\ddot f(e)=2a\epsilon^{\alpha-2}+6b\epsilon^{\alpha-3}e+12c\epsilon^{\alpha-4}e^2+20d\epsilon^{\alpha-5}e^3 \tag{76}$$

$$d^3f(e)/de^3=6b\epsilon^{\alpha-3}+24c\epsilon^{\alpha-4}e+60d\epsilon^{\alpha-5}e^2. \tag{77}$$

The elements of the matrix $I(y^n,\,\theta,\,\epsilon)$ in Condition 1) are clearly continuous in Ω°. Denote their expected value needed in (59) by m_ϵ, which with (70) and (76) becomes

$$\begin{aligned}m_\epsilon=2\lambda Z^{-1}_{\alpha,\,\lambda,\,\epsilon}\Bigg[\int_0^\epsilon e^{-\lambda f(y)}\ddot f(y)\,dy&+\alpha(\alpha-1)\\ &\cdot\int_\epsilon^\infty e^{-\lambda y^\alpha}y^{\alpha-2}\,dy\Bigg].\end{aligned} \tag{78}$$

Introduce $f(e)$ from (58) into (66), multiply the result with the second derivative from (76), and integrate term by term to get the first integral

$$\begin{aligned}\int_0^\epsilon \left[1 - \lambda f(y) + \tfrac{1}{2}\,\lambda^2 f^2(y) - \cdots\right] \ddot{f}(y)\,dy \\ = (2a+3b+4c+5d)\epsilon^{\alpha-1} + O(\epsilon^{2\alpha-1}) \\ = \alpha\epsilon^{\alpha-1} + O(\epsilon^{2\alpha-1}).\end{aligned}$$

For $\alpha > 1$ put $t = \lambda y^\alpha$ in the second integral in (78) and apply (64) and (63), which with $\beta = 1 - 1/\alpha$ gives

$$\begin{aligned}&\int_\epsilon^\infty e^{-\lambda y^\alpha} y^{\alpha-2}\,dy \\ &\quad= \frac{1}{\alpha\lambda^\beta}[\Gamma(\beta) - \gamma(\beta,\, \lambda\epsilon^\alpha)] &(79)\\ &\quad= \frac{1}{\alpha\lambda^{1-1/\alpha}}\left[\Gamma(1-1/\alpha) - \frac{\alpha}{\alpha-1}\,\epsilon^{\alpha-1} + O(\epsilon^\alpha)\right]. &(80)\end{aligned}$$

By combining the two integrals and substituting the expression of $Z_{\alpha,\,\lambda,\,\epsilon}$ we get

$$m_\epsilon = \begin{cases} \lambda^2, & \text{for } \alpha = 1 \\ \frac{\alpha\lambda^{2/\alpha}}{\Gamma(1/\alpha)}[(\alpha-1)\Gamma(1-1/\alpha) & \\ \quad -\alpha\lambda^{1-1/\alpha}\epsilon^{\alpha-1} + O(\epsilon^\alpha)], & \text{for } \alpha > 1. \end{cases} \tag{81}$$

The convergence (59) in Condition 1) is satisfied with the assumption (44)

$$I_n(\theta,\,\epsilon) = m_\epsilon\,\frac{1}{n}\sum_t \boldsymbol{x}_t\boldsymbol{x}_t' \to m_\epsilon\Sigma = I(\theta,\,\epsilon) \tag{82}$$

as $n \to \infty$. We also see that

$$I(\theta,\,\epsilon) \to I(\theta) = \frac{\alpha(\alpha-1)\lambda^{2/\alpha}\Gamma(1-1/\alpha)}{\Gamma(1/\alpha)} \tag{83}$$

as $\epsilon \to 0$, so that the limit may be taken as the Fisher information matrix of $p(y|\boldsymbol{x};\,\theta,\,\lambda)$ even though the double derivative does not exist everywhere. Since $I(\theta,\,\epsilon)$ does not depend on θ and since Ω° is bounded the rest of Condition 1), (60), is satisfied.

We next verify Condition 2). The maximum-likelihood estimates of a scalar-valued parameter satisfy the Central Limit Theorem, provided Cramer's conditions, [4], on the differentiability of the likelihood function are satisfied. The proof extends to vector-valued parameters provided the conditions hold componentwise. These conditions require in the present case, first, that $L_{\alpha,\,\epsilon}(e)$ is three times differentiable in the interior of Ω, which it is. Secondly, we need to show that the absolute values of the first two derivatives of the likelihood function are integrable

$$\int_{\Omega^\circ}\left|\frac{\partial p_\epsilon(y|\boldsymbol{x};\,\theta,\,\lambda)}{\partial\theta_i}\right| dy < \infty$$

$$\int_{\Omega^\circ}\left|\frac{\partial^2 p_\epsilon(y|\boldsymbol{x};\,\theta,\,\lambda)}{\partial\theta_i\,\partial\theta_j}\right| dy < \infty.$$

The first inequality is equivalent with

$$\int_{\Omega^\circ} p_\epsilon(y|\boldsymbol{x};\,\theta,\,\lambda)\left|\frac{\partial L_{\alpha,\,\epsilon}(y-\theta'\boldsymbol{x})}{\partial\theta_i}\right| dy < \infty \tag{84}$$

which follows from (69) and (72). To verify the second inequality note that

$$\begin{aligned}&\left|\frac{\partial^2 p_\epsilon(y|\boldsymbol{x};\,\theta,\,\lambda)}{\partial\theta_i\,\partial\theta_j}\right| \\ &\quad= p_\epsilon(y|\boldsymbol{x};\,\theta,\,\lambda)\left\{\left|\frac{\partial^2 \ln p_\epsilon(y|\boldsymbol{x};\,\theta,\,\lambda)}{\partial\theta_i\,\partial\theta_j}\right|\right. \\ &\qquad\left. + \left|\frac{\partial \ln p_\epsilon(y|\boldsymbol{x};\,\theta,\,\lambda)}{\partial\theta_i}\right|\left|\frac{\partial \ln p_\epsilon(y|\boldsymbol{x};\,\theta,\,\lambda)}{\partial\theta_j}\right|\right\}.\end{aligned}$$

The integral of the first term is finite by (70), (73), and (76), while the finiteness of the integral of the second terms follows from (69), (72), and (75).

Further, Cramer's condition

$$\int p_\epsilon(y|\boldsymbol{x};\,\theta,\,\lambda)\left|\frac{\partial^3 L_{\alpha,\,\epsilon}(y-\theta'\boldsymbol{x})}{\partial\theta_i\partial\theta_j\partial\theta_k}\right| dy < \infty \tag{85}$$

needs to be verified. From (71)

$$\int_\epsilon^\infty e^{-\lambda y^\alpha} y^{\alpha-3}\,dy < \epsilon^{\alpha-3} Z_{\alpha,\,\lambda,\,\epsilon}$$

and from (58) and (77) we get

$$\begin{aligned}&\int_0^\epsilon e^{-\lambda f(y)}(6b\epsilon^{\alpha-3} + 24c\epsilon^{\alpha-4}y + \epsilon^{\alpha-5}y^2)\,dy \\ &\qquad\qquad= (6b+12c+20d)\epsilon^{\alpha-2} + O(\epsilon^{2\alpha-2}).\end{aligned}$$

Hence, the inequality (85) holds. Finally, the Fisher information matrix $I(\alpha,\,\epsilon)$ is clearly bounded and positive definite, and by Cramer's conditions [4], the Central Limit Theorem holds for the family $\{p_\epsilon(y|\boldsymbol{x};\,\theta,\,\lambda)\}$.

To verify Condition 3) we get from (73) and (76) for $y_t - \theta'\boldsymbol{x}_t \le \epsilon$

$$\frac{1}{n}\sum_t \frac{\partial^2 L_{\epsilon,\,\alpha}(y_t - \theta'\boldsymbol{x}_t)}{\partial\theta_i\,\partial\theta_j} = O(\epsilon^{\alpha-2})$$

while by (70) for $y_t - \theta'\boldsymbol{x}_t > \epsilon$

$$\begin{aligned}\frac{1}{n}\,\frac{\partial^2 \sum_t L_{\epsilon,\,\alpha}(y_t-\theta'\boldsymbol{x}_t)}{\partial\theta_i\,\partial\theta_j} &= \frac{\lambda\alpha(\alpha-1)}{n}\sum_t (y_t-\theta'\boldsymbol{x}_t)^{\alpha-2}\boldsymbol{x}_i\boldsymbol{x}_j' \\ &\le O(\epsilon^{\alpha-2})\end{aligned}$$

where the last inequality holds for $\alpha \le 2$. Hence, Condition 3) holds for $1 \le \alpha \le 2$.

To complete the proof of Theorem 2 notice first that for every y

$$p_\epsilon(y|\boldsymbol{x};\,\theta,\,\lambda) \to p(y|\boldsymbol{x};\,\theta,\,\lambda)$$

as $\epsilon \to 0$. By [10, Theorem 1], (55) holds for the family $\mathcal{M}_{\alpha,\,\lambda,\,\epsilon}$, where $I(\theta)$ is replaced by $I(\theta,\,\epsilon)$ in (82), and letting $\epsilon \to 0$ we get with (83) the formula in Theorem 2.

We conclude this appendix by showing that the Central Limit Theorem holds for the maximum-likelihood estimates in the family $\mathcal{M}_{\alpha,\,\lambda}$. In fact, let A be an open set in the parameter space, and put $\delta_\epsilon = \sqrt{n}(\hat{\theta}_\epsilon(y^n, \boldsymbol{x}^n) - \theta)$, where $\hat{\theta}_\epsilon(y^n, \boldsymbol{x}^n)$ denotes the maximum-likelihood estimator in the family $\mathcal{M}_{\epsilon, \alpha, \lambda}$. Similarly, put $\delta = \sqrt{n}(\hat{\theta}(y^n, \boldsymbol{x}^n) - \theta)$. Then, by (67) and (68)

$$\int_{\delta_\epsilon \in A} p_\epsilon(y^n | \boldsymbol{x}^n; \theta, \lambda)\, dy^n \to \int_{\delta \in A} p(y^n | \boldsymbol{x}^n; \theta, \lambda)\, dy^n$$

as $\epsilon \to 0$. Since the Central Limit Theorem holds for the family $\mathcal{M}_{\epsilon, \alpha, \lambda}$, the left-hand side for each ϵ converges to the probability of A under the limiting normal density function, as $n \to \infty$, and so does the right-hand side.

Appendix B

With the definitions of $\tilde{\lambda}$, (46), and $\bar{\lambda}$, (18), we get

$$\begin{aligned}
&\bar{\lambda} L(y^n | \boldsymbol{x}^n; \hat{\theta}(y^n, \boldsymbol{x}^n)) + \ln B_{n, \alpha, \bar{\lambda}}(\boldsymbol{x}^n)\\
&\quad = \bar{\lambda} L(y^n | \boldsymbol{x}^n; \hat{\theta}(y^n, \boldsymbol{x}^n)) + \ln B'_{n, \alpha, \bar{\lambda}}(\boldsymbol{x}^n) + \epsilon_{\bar{\lambda}}(n)\\
&\quad \le \tilde{\lambda} L(y^n | \boldsymbol{x}^n; \hat{\theta}(y^n, \boldsymbol{x}^n)) + \ln B_{n, \alpha, \tilde{\lambda}}(\boldsymbol{x}^n)\\
&\quad = \tilde{\lambda} L(y^n | \boldsymbol{x}^n; \hat{\theta}(y^n, \boldsymbol{x}^n)) + \ln B'_{n, \alpha, \bar{\lambda}}(\boldsymbol{x}^n) + \epsilon_{\tilde{\lambda}}(n)\\
&\tilde{\lambda} L(y^n | \boldsymbol{x}^n; \hat{\theta}(y^n, (\boldsymbol{x}^n)) + \ln B'_{n, \alpha, \bar{\lambda}}(\boldsymbol{x}^n)\\
&\quad \le \bar{\lambda} L(y^n | \boldsymbol{x}^n; \hat{\theta}(y^n, (\boldsymbol{x}^n)) + \ln B'_{n, \alpha, \bar{\lambda}}(\boldsymbol{x}^n)
\end{aligned}$$

where by the expression for $C_{n, \alpha, \lambda}(\boldsymbol{x}^n)$ in Theorem 2 the ϵ-terms go to zero as n grows to infinity. These imply

$$\begin{aligned}
&0 \le (\bar{\lambda} - \tilde{\lambda}) L(Y^n | \boldsymbol{x}^n; \hat{\theta}(Y^n, (\boldsymbol{x}^n)) + \ln \frac{B'_{n, \alpha, \bar{\lambda}}(\boldsymbol{x}^n)}{B'_{n, \alpha, \tilde{\lambda}}(\boldsymbol{x}^n)}\\
&\quad \le \epsilon_{\tilde{\lambda}}(n) - \epsilon_{\bar{\lambda}}(n) = o(1).
\end{aligned}$$

From the fact that

$$\ln \frac{B'_{n, \alpha, \bar{\lambda}}(\boldsymbol{x}^n)}{B'_{n, \alpha, \tilde{\lambda}}(\boldsymbol{x}^n)} = \frac{n-k}{\alpha} \ln \frac{\tilde{\lambda}}{\bar{\lambda}}$$

we then deduce

$$0 \le \frac{\bar{\lambda} - \tilde{\lambda}}{\tilde{\lambda}} + \ln \frac{\tilde{\lambda}}{\bar{\lambda}} \le o\left(\frac{1}{n-k}\right).$$

By expanding the logarithm into Tailor series we get

$$0 \le O\left(\left(\frac{\bar{\lambda} - \tilde{\lambda}}{\tilde{\lambda}}\right)^2\right) \le o\left(\frac{1}{n-k}\right)$$

which implies (47).

References

[1] V. Balasubramanian, "Statistical inference, Occam's razor and statistical mechanics on the space of probability distributions," *Neural Comput.*, vol. 9, no. 2, pp. 349–268, 1997.

[2] A. R. Barron, J. Rissanen, and B. Yu, "The MDL principle in modeling and coding," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2743–2760, Oct. 1998. Special Issue to Commemorate 50 Years of Information Theory.

[3] T. M Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 542 pp., 1991.

[4] M. Fisz, *Probability Theory and Mathematical Statistics*. New York: Wiley, 677 pp., 1963.

[5] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*. New York: Academic, 1160 pp., 1980.

[6] P. D. Grünwald, "The minimum description length principle and reasoning under uncertainty," Ph.D. dissertation, Institute for Logic, Language and Computation, Univ. Amsterdam, Amsterdam, The Netherlands, 296 pp., 1998.

[7] E. J. Hannan, A. J. McDougal, and D. S. Poskitt, "Recursive estimation of autoregressions," *J. Roy. Statist. Soc.*, ser. B, vol. 51, no. 2, pp. 217–233, 1989.

[8] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2124–2147, Oct. 1998. Special Issue to Commemorate 50 Years of Information Theory.

[9] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, July 1984.

[10] ——, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory*, vol. 42, pp. 40–47, Jan. 1996.

[11] ——, "Strong optimality of the normalized ML models as universal codes and information in data," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1712–1717, July 2001.

[12] V. G. Vovk, "Aggregating strategies," in *Proc. 3rd Annu. Workshop on Computational Learning Theory*. Los Altos, CA: Morgan Kauffman, 1990, pp. 371–386.

[13] K. Yamanishi, "A decision-theoretic extension of stochastic complexity and its application to learning," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1424–1439, July 1998.

[14] ——, "Minimax relative loss analysis for sequential prediction algorithms using parametric hypotheses," in *Proc. 11th Annu. Conf. Computational Learning Theory*, 1998, pp. 32–43.

[15] ——, "Extended stochastic complexity and minimax relative loss analysis," in *Algorithmic Learning Theory, Proc. 10th Int. Conf., ALT'99 (Lecture Notes in Artificial Intelligence)*. New York: Springer-Verlag, 1999, vol. 1720, pp. 26–38.