Asymptotic Log-loss of Prequential Maximum Likelihood Codes

Peter Grünwald[1] and Steven de Rooij[2]

[1] CWI Amsterdam, `Peter.Grunwald@cwi.nl`, `www.grunwald.nl`.
[2] CWI Amsterdam, `S.de.Rooij@cwi.nl`, `www.cwi.nl/~rooij`.

Abstract. We analyze the Dawid-Rissanen *prequential maximum likelihood codes* relative to one-parameter exponential family models $\mathcal{M}$. If data are i.i.d. according to an (essentially) *arbitrary* P, then the redundancy grows at rate $\frac{1}{2}c\ln n$. We show that $c = \sigma_1^2/\sigma_2^2$, where σ_1^2 is the variance of P, and σ_2^2 is the variance of the distribution $M^* \in \mathcal{M}$ that is closest to P in KL divergence. This shows that prequential codes behave quite differently from other important universal codes such as the 2-part MDL, Shtarkov and Bayes codes, for which $c = 1$. This behavior is undesirable in an MDL model selection setting.

1 Introduction

Universal coding lies at the basis of on-line prediction algorithms for data compression and gambling purposes. It has been extensively studied in the COLT community, typically under the name of 'sequential prediction with log loss', see, for example [8, 1, 3]. It also underlies Rissanen's theory of MDL (minimum description length) learning [2, 9] and Dawid's theory of prequential model assessment [6]. Roughly, a code is *universal* with respect to a set of candidate codes $\mathcal{M}$ if it achieves small *redundancy*: it allows one to encode data using not many more bits than the optimal code in $\mathcal{M}$. The redundancy is very closely related to the *expected regret*, which is perhaps more widely known within the COLT community – see Section 4. The main types of universal codes are the *Shtarkov* or *NML* code, the *Bayesian mixture* code, the *2-part MDL code* and the *prequential maximum likelihood (ML) code*, also known as the 'ML plug-in code' or the 'predictive MDL code' [2, 9]. This code was introduced independently by Rissanen [14] and by Dawid [6], who proposed it as a probability forecasting strategy rather than directly as a code. The main ideas are explained in Section 2. Here we study the case where no code in $\mathcal{M}$ corresponds to the data-generating distribution P, for the wide class of 1-parameter exponential families $\mathcal{M}$. We find that then the redundancy of the prequential code can be quite different from that of the other methods: the redundancies are $\frac{1}{2}c\ln n + O(1)$. Under regularity conditions on P and $\mathcal{M}$, we find $c = 1$ for Bayes, NML and 2-part codes, whereas for the prequential ML codes, we can get any $c > 0$, depending on P and $\mathcal{M}$.

Relevance: As discussed in Section 5, the result has interesting consequences for parameter estimation and practical data compression. The most important and surprising consequence is for MDL learning and model selection, where our result implies that the prequential ML code may behave suboptimally *even if one of the models under consideration is correct after all!*

Contents In Section 2 we informally state and explain our result. Section 3 contains the formal statement of our main result (Theorem 1), as well as a proof. Section 4 provides Theorem 2, which implies that a version of our result still holds if 'redundancy' is replaced by 'expected regret'. We discuss further issues, including relevance of the result and related work, in Section 5. Section 6 states and proves various lemma's needed in the proofs of Theorem 1 and 2. We provide a lot more discussion, and discuss the proofs in much greater detail, in the technical report [10].

2 Main Result, Informally

Suppose $\mathcal{M} = \{M_\theta : \theta \in \Theta\}$ is a k-dimensional parametric family of distributions, which we use to model some distribution P from which we sample i.i.d. outcomes $Z_1, Z_2, \ldots$ A code is *universal* for $\mathcal{M}$ if it is almost as efficient at coding the outcomes from P as the best element of $\mathcal{M}$. (We use codes and distributions interchangeably.) The *redundancy* of a universal code U with respect to P is defined as

$$\mathcal{R}_U(n) := E_P[L_U(Z_1, \ldots, Z_n)] - \inf_{\theta \in \Theta} E_P[-\ln M_\theta(Z_1, \ldots, Z_n)], \tag{1}$$

where L_U is the length function of U and $M_\theta(Z_1, \ldots, Z_n)$ denotes the probability mass or density of $Z_1, \ldots, Z_n$ under distribution M_θ; these and other notational conventions are detailed in Section 3. Often, P is assumed to be an element of the model. If such is the case, then by the information inequality [5] the second term is minimized for $M_\theta = P$, so that

$$\mathcal{R}_U(n) = E_P[L_U(Z_1, \ldots, Z_n)] - E_P[-\ln P(Z_1, \ldots, Z_n)], \tag{2}$$

We use nats rather than bits as units of information to simplify equations; a nat can be thought of as the amount of information needed to distinguish between e rather than 2 alternatives. Thus, (2) can be interpreted as the expected number of additional nats one needs to encode n outcomes if one uses the code U instead of the optimal (Shannon-Fano) code with lengths $-\ln P(Z_1, \ldots, Z_n)$. A good universal code achieves small redundancy (or regret, see Section 4) for all or 'most' $P \in \mathcal{M}$.

The four major types of universal codes, Bayes, NML, 2-part and prequential ML, all achieve redundancies that are (in an appropriate sense) close to optimal. Specifically, under regularity conditions on $\mathcal{M}$ and its parameterization, these four types of universal codes all satisfy, for all $P \in \mathcal{M}$,

$$\mathcal{R}(n) = \frac{k}{2} \ln n + O(1), \tag{3}$$

where the $O(1)$ may depend on θ and the universal code used. (3) is the famous 'k over 2 log n formula', refinements of which lie at the basis of most practical approximations to MDL learning, see [9].

In this paper we consider the case where the data are i.i.d. according to an *arbitrary* P not necessarily in the model $\mathcal{M}$. It is now appropriate to rename the redundancy to *relative* redundancy, since we measure the number of nats we lose relative to the best element in the model, rather than relative to the generating distribution P. The definition (1) remains unchanged. It can no longer be rewritten as (2) however: Assuming it exists and is unique, let M_{θ^*} be the element of $\mathcal{M}$ that minimizes KL divergence to P:

$$\theta^* := \arg\min_{\theta\in\Theta} D(P\|M_\theta) = \arg\min_{\theta\in\Theta} E_P[-\ln M_\theta(Z)],$$

where the equality follows from the definition of the KL divergence [5]. Then the relative redundancy satisfies

$$\mathcal{R}_U(n) = E_P[L_U(Z_1,\ldots,Z_n)] - E_P[-\ln M_{\theta^*}(Z_1,\ldots,Z_n)]. \tag{4}$$

It turns out that for the NML, 2-part MDL and Bayes codes, the relative redundancy (4) with $P \notin \mathcal{M}$, still satisfies (3), at least under conditions on $\mathcal{M}$ and P (Section 4). In this paper, we show for the first time that (3) does *not* hold for the prequential ML code. The prequential ML code U works by sequentially predicting Z_{i+1} using a (slightly modified) ML or Bayesian MAP estimator $\hat\theta_i = \hat\theta(z^i)$ based on the past data, that is, the first i outcomes $z^i = z_1,\ldots,z_i$. The total codelength $L_U(z^n)$ on a sequence z^n is given by the sum of the individual 'predictive' codelengths (log losses): $L_U(z^n) = \sum_{i=0}^{n-1}[-\ln M_{\hat\theta_i}(z_{i+1})]$. In our main theorem, we show that if L_U denotes the prequential ML code length, and $\mathcal{M}$ is a regular one-parameter exponential family ($k = 1$), then

$$\mathcal{R}_U(n) = \frac{1}{2}\frac{\mathrm{var}_P X}{\mathrm{var}_{M_{\theta^*}} X}\ln n + O(1), \tag{5}$$

where X is the sufficient statistic of the family. Example 1 below illustrates the phenomenon. Note that if $P \in \mathcal{M}$, then $M_{\theta^*} = P$ and (5) becomes the familiar expression. The result holds as long as $\mathcal{M}$ and P satisfy a mild condition that is stated and discussed in the next section. Section 5 discusses the consequences of this result for compression, estimation and model selection, as well as its relation to the large body of earlier results on prequential ML coding.

Example 1. Let $\mathcal{M}$ be the family of Poisson distributions, parameterized by their mean μ. Since neither the NML universal code nor Jeffreys' prior are defined for this model it is attractive to use the prequential ML code as a universal code for this model. The ML estimator $\hat\mu_i$ is the empirical mean of $z_1,\ldots,z_i$.

Suppose Z, Z_1, Z_2, $\ldots$ are i.i.d. according to a degenerate P with $P(Z = 4) = 1$. Since the sample average is a sufficient statistic for the Poisson family, $\hat\mu_i$ will be equal to 4 for all $i \geq 1$. On the other hand, μ^*, the parameter (mean) of the distribution in $\mathcal{M}$ closest to P in KL-divergence, will be equal to 4 as well. Thus the redundancy (4) of the prequential ML code is given by

$$\mathcal{R}_U(n) = -\ln M_{\hat\mu_0}(4) + \ln M_4(4) + \sum_{i=1}^{n-1}[-\ln M_4(4) + \ln M_4(4)] = O(1),$$

assuming an appropriate definition of $\hat{\mu}_0$. In the case of the Poisson family, we have $Z = X$ in (5). Thus, since $\mathrm{var}_P Z = 0$, this example agrees with (5).

Now suppose data are i.i.d. according to some P_τ, with $P_\tau(Z = z) \propto (z+1)^{-3}$ for all z smaller than τ, and $P_\tau(Z = z) = 0$ for $z \geq \tau$. It is easy to check that, for $\tau \to \infty$, the entropy of P_τ converges to a finite constant, but the variance of P_τ tends to infinity. Thus, by choosing τ large enough, the regret obtained by the Poisson prequential ML code can be made to grow as $c \log n$ for arbitrarily large c.

Example 2. The Hardy-Weinberg model deals with genotypes of which the alleles are assumed independently Bernoulli distributed according to some parameter p. There are four combinations of alleles, usually denoted 'aa', 'AA', 'aA', 'Aa'; but since 'aA' and 'Aa' result in the same genotype, the Hardy-Weinberg model is defined as a probability distribution on three outcomes. We model this by letting X be a random variable on the underlying space, that maps 'aA' and 'Aa' to the same value: $X(aa) = 0$, $X(aA) = X(Aa) = \frac{1}{2}$ and $X(AA) = 1$. Then $P(X = 0) = (1-p)^2$, $P(X = \frac{1}{2}) = 2p(1-p)$ and $P(X = 1) = p^2$. The Hardy-Weinberg model is an exponential family with sufficient statistic X. To see this, note that for any parameter $p \in [0, 1]$, we have $EX = \mu = P(A) = p$, so we can parameterize the model by the mean of X. The variance of the distribution with parameter μ is $\frac{1}{2}\mu(1-\mu)$. Now suppose that we code data in a situation where the Hardy-Weinberg model is *wrong* and the genotypes are in fact distributed according to $P(X = \frac{1}{2}) = P(X = 1) = \frac{1}{2}$ and $P(X = 0) = 0$, such that mean and variance of X are $\frac{3}{4}$ and $\frac{2}{32}$ respectively. The closest distribution in the model has the same mean (since the mean is a sufficient statistic), and variance $\frac{3}{32}$. Thus the prequential ML code will achieve an asymptotic regret of $\frac{1}{3} \ln n$ rather than $\frac{1}{2} \ln n$ (up to $O(1)$).

3 Main Result, Formally

In this section, we define our quantities of interest and we state and prove our main result. Throughout this text we use nats rather than bits as units of information. Outcomes are capitalized if they are to be interpreted as random variables instead of instantiated values. A sequence of outcomes $z_1, \ldots, z_n$ is abbreviated to z^n. We write E_P as a shorthand for $E_{Z \sim P}$, the expectation of Z under distribution P. When we consider a sequence of n outcomes independently distributed $\sim P$, we use E_P even as a shorthand for the expectation of $(Z_1, \ldots, Z_n)$ under the n-fold product distribution of P. Finally, $P(Z)$ denotes the probability mass function of P in case Z is discrete-valued, and it denotes the density of P, in case Z takes its value in a continuum. When we write 'density function of Z', then, if Z is discrete-valued, this should be read as 'probability mass function of Z'. Note however that in our main result, Theorem 1 below, we do not assume that the data-generating distribution P admits a density.

Let $\mathcal{Z}$ be a set of outcomes, taking values either in a finite or countable set, or in a subset of k-dimensional Euclidean space for some $k \geq 1$. Let $X : \mathcal{Z} \to \mathbb{R}$

be a random variable on $\mathcal{Z}$, and let $\mathcal{X} = \{x \in \mathbb{R} : \exists z \in \mathcal{Z} : X(z) = x\}$ be the range of X. Exponential family models are families of distributions on $\mathcal{Z}$ defined relative to a random variable X (called 'sufficient statistic') as defined above, and a function $h : \mathcal{Z} \to [0, \infty)$. Let $\mathrm{Z}(\eta) := \int_{z \in \mathcal{Z}} e^{-\eta X(z)} h(z) dz$ (the integral to be replaced by a sum for countable $\mathcal{Z}$), and $\Theta_\eta := \{\eta \in \mathbb{R} : \mathrm{Z}(\eta) < \infty\}$.

Definition 1 (Exponential family). *The* single parameter exponential family [12] *with* sufficient statistic X *and* carrier h *is the family of distributions with densities* $M_\eta(z) := \frac{1}{\mathrm{Z}(\eta)} e^{-\eta X(z)} h(z)$*, where* $\eta \in \Theta_\eta$*.* Θ_η *is called the* natural parameter space*. The family is called* regular *if* Θ_η *is an open interval of* $\mathbb{R}$*.*

In the remainder of this text we only consider single parameter, regular exponential families with a 1-to-1 parameterization, but this qualification will henceforth be omitted. Examples include the Poisson, geometric and multinomial families, and the model of all Gaussian distributions with a fixed variance or mean.

The statistic $X(z)$ is sufficient for η [12]. This suggests reparameterizing the distribution by the expected value of X, which is called the *mean value parameterization*. The function $\mu(\eta) = E_{M_\eta}[X]$ maps parameters in the natural parameterization to the mean value parameterization. It is a diffeomorphism (it is one-to-one, onto, infinitely often differentiable and has an infinitely often differentiable inverse) [12]. Therefore the mean value parameter space Θ_μ is also an open interval of $\mathbb{R}$. We note that for some models (such as Bernoulli and Poisson), the parameter space is usually given in terms of the a non-open set of mean-values (e.g., $[0, 1]$ in the Bernoulli case). To make the model a regular exponential family, we have to restrict the set of parameters to its own interior. Henceforth, whenever we refer to a standard statistical model such as Bernoulli or Poisson, we assume that the parameter set has been restricted in this sense.

We are now ready to define the prequential ML model. This is a distribution on infinite sequences $z_1, z_2, \ldots \in \mathcal{Z}^\infty$, recursively defined in terms of the distributions of Z_{n+1} conditioned on $Z^n = z^n$, for all $n = 1, 2, \ldots$, all $z^n = (z_1, \ldots, z_n) \in \mathcal{Z}^n$. In the definition, we use the notation $x_i := X(z_i)$.

Definition 2 (Prequential ML model). *Let* Θ_μ *be the mean value parameter domain of an exponential family* $\mathcal{M} = \{M_\mu \mid \mu \in \Theta_\mu\}$*. Given* $\mathcal{M}$ *and constants* $x_0 \in \Theta_\mu$ *and* $n_0 > 0$*, we define the* prequential ML model U *by setting, for all* n*, all* $z^{n+1} \in \mathcal{Z}^{n+1}$*:*

$$U(z_{n+1} \mid z^n) = M_{\hat{\mu}(z^n)}(z_{n+1}),$$

where $U(z_{n+1} \mid z^n)$ *is the density/mass function of* z_{n+1} *conditional on* $Z^n = z^n$*,*

$$\hat{\mu}(z^n) := \frac{x_0 \cdot n_0 + \sum_{i=1}^n x_i}{n + n_0},$$

and $M_{\hat{\mu}(z^n)}(\cdot)$ *is the density of the distribution in* $\mathcal{M}$ *with mean* $\hat{\mu}(z^n)$*.*

We henceforth abbreviate $\hat{\mu}(z^n)$ to $\hat{\mu}_n$. We usually refer to the prequential ML model in terms of the corresponding codelength function

$$L_U(z^n) = \sum_{i=0}^{n-1} L_U(z_{i+1} \mid z_i) = \sum_{i=0}^{n-1} - \ln M_{\hat{\mu}_i}(z_{i+1}).$$

To understand this definition, note that for exponential families, for any sequence of data, the ordinary maximum likelihood parameter is given by the average $n^{-1}\sum x_i$ of the observed values of X [12]. Here we define our prequential model in terms of a slightly modified maximum likelihood estimator that introduces a 'fake initial outcome' x_0 with multiplicity n_0 in order to avoid infinite code lengths for the first few outcomes (a well-known problem called by Rissanen the "inherent singularity" of predictive coding [16, 10]) and to ensure that the prequential ML code of the first outcome is well-defined. In practice we can take $n_0 = 1$ but our result holds for any $n_0 > 0$. The justification of our modification to the ML estimator is discussed further in Section 5 and in [10].

Theorem 1 (Main result). *Let $X, X_1, X_2, \ldots$ be i.i.d. $\sim P$, with $E_P[X] = \mu^*$. Let $\mathcal{M}$ be a single parameter exponential family with sufficient statistic X and μ^* an element of the mean value parameter space. Let U denote the prequential ML model with respect to $\mathcal{M}$. If $\mathcal{M}$ and P satisfy Condition 1 below, then*

$$\mathcal{R}_U(n) = \frac{\mathrm{var}_P X}{\mathrm{var}_{M_{\mu^*}} X}\frac{1}{2}\ln n + O(1). \tag{6}$$

To reconcile this with (5), note that M_{μ^*} is the element of $\mathcal{M}$ achieving smallest expected codelength, i.e. it achieves $\inf_{\mu\in\Theta_\mu} D(P\|M_\mu)$ [12].

Condition 1 *We require that the following holds both for $T := X$ and $T := -X$:*

- *If T is unbounded from above then there is a $k \in \{4, 6, \ldots\}$ such that the first k moments of T exist under P and that $\frac{d^4}{d\mu^4}D(M_{\mu^*}\|M_\mu) = O\left(\mu^{k-6}\right)$.*
- *If T is bounded from above by a constant g then $\frac{d^4}{d\mu^4}D(M_{\mu^*}\|M_\mu)$ is polynomial in $1/(g-\mu)$.*

The condition implies that Theorem 1 can be applied to most single-parameter exponential families that are relevant in practice. To illustrate, in [10] we have explicitly computed the fourth derivative of the divergence for a number of exponential families: the Poisson, geometric, Normal, Bernoulli, exponential and Pareto families; all parameters beside the mean are treated as fixed values. As can be seen from [10, Figure 1], for these exponential families, our condition applies whenever the fourth moment of P exists. We explain the need for Condition 1 directly after the proof of Theorem 1:

Proof. (of Theorem 1) Here we only give the proof for the special case that $X = Z$, i.e. the sufficient statistic of $\mathcal{M}$ is equal to the observed outcome. Extension to the general case is trivial, see [10]. Thus, in the proof of Theorem 1 as well as all the Lemmas and Propositions it makes use of, it is implicitly understood that $X = Z$. The proof relies on Lemma 2 (Section 6), which has a complicated and lengthy proof. But as we shall see, once one takes Lemma 2 for granted, the proof becomes quite straightforward:

We abbreviate $\delta_i = \hat{\mu}_i - \mu^*$ and $\frac{d^k}{d\mu^k} D(M_{\mu^*} \| M_\mu) = D^{(k)}(\mu)$. That is, $D^{(k)}(\mu)$ is the k-th derivative of the function $f(\mu) := D(M_{\mu^*} \| M_\mu)$. The proof is based on a Taylor expansion of the KL divergence $D(M_{\mu^*} \| M_{\hat{\mu}_i})$ around μ^*:

$$D(M_{\mu^*} \| M_{\hat{\mu}_i}) = 0 + \delta_i D^{(1)}(\mu^*) + \frac{{\delta_i}^2}{2} D^{(2)}(\mu^*) + \frac{{\delta_i}^3}{6} D^{(3)}(\mu^*) + \frac{{\delta_i}^4}{24} D^{(4)}(\mu)$$

The last term is the remainder term of the Taylor expansion, in which $\mu \in [\mu^*, \hat{\mu}_i]$. The second term $D^{(1)}(\mu^*)$ is zero, since $D(\mu^* \| \mu)$ has its minimum at $\mu = \mu^*$ [12]. As is well-known [12], for exponential families the term $D^{(2)}(\mu)$ coincides precisely with the Fisher information $I(\mu)$ evaluated at μ. Another standard result [12] for the mean-value parameterization says that for all μ,

$$I(\mu) = \frac{1}{\mathrm{var}_{M_\mu} X}. \tag{7}$$

Therefore:

$$D(M_{\mu^*} \| M_{\hat{\mu}_i}) = \frac{1}{2} {\delta_i}^2 / \mathrm{var}_{M_{\mu^*}}(X) + \frac{1}{6} {\delta_i}^3 D^{(3)}(\mu^*) + \frac{1}{24} {\delta_i}^4 D^{(4)}(\mu) \tag{8}$$

We now compute the expected sum of (8), where $\hat{\mu}_i$ (and therefore δ_i) is a random variable that takes on values according to P, while μ^* is fixed. We get:

$$\sum_{i=0}^{n-1} \mathop{E}_{\hat{\mu}_i \sim P} [D(M_{\mu^*} \| M_{\hat{\mu}_i})] = \frac{1}{2 \mathrm{var}_{M_{\mu^*}}(X)} \sum_{i=0}^{n-1} E_P \left[{\delta_i}^2 \right] + R(n), \tag{9}$$

where the remainder term $R(n)$ is given by

$$R(n) = \sum_{i=0}^{n-1} \mathop{E}_{\hat{\mu}_i \sim M^*} \left[\frac{1}{6} {\delta_i}^3 D^{(3)}(\mu^*) + \frac{1}{24} {\delta_i}^4 D^{(4)}(\mu) \right] \tag{10}$$

where μ and δ_i are random variables depending on $\hat{\mu}_i$ and i. Theorem 1 now follows by rewriting the term on the left and the two terms on the right in (9):

1. Lemma 1 (Section 6) shows that $\mathcal{R}_U(n) = \sum_{i=0}^{n-1} E_{\hat{\mu}_i \sim P} [D(M_{\mu^*} \| M_{\hat{\mu}_i})]$, so that the left-hand side of (9) is equal to the left-hand side of (6).
2. Lemma 2 (Section 6) shows that $R(n) = O(1)$.
3. To evaluate the remaining term $\frac{1}{2 \mathrm{var}_{M_{\mu^*}}(X)} \sum_{i=0}^{n-1} E_P \left[{\delta_i}^2 \right]$ in (9), note that $\hat{\mu}_i$ is almost the ML estimator. This suggests that each term in the sum $\sum_{i=0}^{n-1} E_P \left[{\delta_i}^2 \right]$ should be almost equal to the variance of the ML estimator, which is $\mathrm{var} X / i$. Because of the slight modification that we made to the estimator, the variance becomes $\mathrm{var} X/(i+1) + O((i+1)^{-2})$ (this is easy to verify; an explicit proof is given as Theorem 5 in [10]). We then get:

$$\begin{aligned} \sum_{i=0}^{n-1} E_P \left[(\hat{\mu}_i - \mu^*)^2 \right] &= \sum_{i=0}^{n-1} O\left((i+1)^{-2} \right) + \mathrm{var}_P(X) \sum_{i=0}^{n-1} (i+1)^{-1} \\ &= O(1) + \mathrm{var}_P(X) \ln n, \end{aligned} \tag{11}$$

so that Theorem 1 follows with items (1) and (2) above.

Discussion Lemma 1 follows relatively easily by rewriting the sum using the chain rule for relative entropy and the fact that X is a sufficient statistic. The truly difficult part of the proof is Lemma 2. It involves infinite sums of expectations over unbounded fourth-order derivatives of the KL divergence. To make this work, we (1) slightly modify the ML estimator by introducing the initial fake outcome x_0. And (2), we need to impose Condition 1. To understand it, consider the case $T = X$, X unbounded from above. The condition essentially expresses that, as $\hat{\mu}$ increases to infinity, the fourth order Taylor-term does not grow too fast. Similarly, if X is bounded from above by g, the condition ensures that the fourth-order term grows slowly enough as $\hat{\mu} \uparrow g$. The same requirements are imposed for decreasing $\hat{\mu}$.

4 Redundancy vs. Regret

The 'goodness' of a universal code relative to a model $\mathcal{M}$ can be measured in several ways: rather than using redundancy (as we did here), one can also choose to measure codelength differences in terms of *regret*, where one has a further choice between *expected regret* and *worst-case regret* [2]. Here we only discuss the implications of our result for the expected regret measure.

Let $\mathcal{M} = \{M_\theta \mid \theta \in \Theta\}$ be a family of distributions parameterized by Θ. Given a sequence $z^n = z_1, \ldots, z_n$ and a universal code U for $\mathcal{M}$ with lengths L_U, the *regret* of U on sequence z^n is defined as

$$L_U(z^n) - \inf_{\theta \in \Theta}[-\ln M_\theta(z^n)]. \tag{12}$$

Note that if the (unmodified) ML estimator $\hat{\theta}(z^n)$ exists, then this is equal to $L_U(z^n) + \ln M_{\hat{\theta}(z^n)}(z^n)$. Thus, one compares the codelength achieved on z^n by U to the best possible that could have been achieved on that particular z^n, using any of the distributions in $\mathcal{M}$. Assuming $Z_1, Z_2, \ldots$ are i.i.d. according to some (arbitrary) P, one may now consider the expected regret

$$\widehat{\mathcal{R}}_U(n) := E_P[L_U(Z^n) - \inf_{\theta \in \Theta}[-\ln M_\theta(Z^n)]],$$

To quantify the difference with redundancy, consider the function

$$d(n) := \inf_{\theta \in \Theta} E_P[-\ln M_\theta(Z^n)] - E_P\,[\inf_{\theta \in \Theta}[-\ln M_\theta(Z^n)]],$$

and note that for any universal code, $\mathcal{R}_U(n) - \widehat{\mathcal{R}}_U(n) = d(n)$. In case $P \in \mathcal{M}$, then under regularity conditions on $\mathcal{M}$ and its parameterization, it can be shown [4] that $\lim_{n\to\infty} d(n) = \frac{k}{2}$ where k is the dimension of $\mathcal{M}$. In our case, where P is not necessarily in $\mathcal{M}$, we have the following :

Theorem 2. *Let $\mathcal{X}$ be finite. Let P, M_μ and μ^* be as in Theorem 1. Then*

$$\lim_{n\to\infty} d(n) = \frac{1}{2}\frac{\mathrm{var}_P X}{\mathrm{var}_{M_{\mu^*}} X}. \tag{13}$$

For lack of space we have to omit the proof. It can be found in [10]. Note that the previous result for one-dimensional models $\mathcal{M}$ with $P \in \mathcal{M}$, $d(n) \to 1/2$, is a special case of (13). We conjecture that, under a condition similar to Condition 1, the same result still holds for general, not necessarily finite or countable or bounded $\mathcal{X}$, but at the time of writing this submission we did not yet find the time to sort out the details. In any case, our result is sufficient to show that in some cases (namely, if $\mathcal{X}$ is finite), we have

$$\widehat{\mathcal{R}}_U(n) = \frac{1}{2} \frac{\mathrm{var}_P X}{\mathrm{var}_{M_{\mu^*}} X} \ln n + O(1),$$

so that, up to $O(1)$-terms, the redundancy and the regret of the prequential ML code behave in the same way.

Incidentally, Theorem 2 can be used to substantiate the claim we made in Section 2, which stated that the Bayes (equipped with a strictly positive differentiable prior), NML and 2-part codes still achieve relative redundancy of $\frac{1}{2} \ln n$ if $P \neq \mathcal{M}$, at least if $\mathcal{X}$ is finite. Let us informally explain why this is the case. It is easy to show that Bayes, NML and (suitably defined) 2-part codes achieve regret $\frac{1}{2} \ln n + O(1)$ for *all* sequences $z_1, z_2, \ldots$ such that $\hat{\theta}(z^n)$ is bounded away from the boundary of the parameter space M, for all large n [2, 9]. It then follows using, for example, the Chernoff bound that these codes must also achieve expected regret $\frac{1}{2} \ln n + O(1)$ for *all* distributions P on $\mathcal{X}$ that satisfy $E_P[X] = \mu^* \in \Theta_\mu$. Theorem 2 then shows that they also achieve relative redundancy $\frac{1}{2} \ln n + O(1)$ for *all* distributions P on $\mathcal{X}$ that satisfy $E_P[X] = \mu^* \in \Theta_\mu$. We omit further details.

5 Discussion

1. Variations of Prequential Coding As discussed in [10], in order to apply the prequential ML code, either the ML estimator has to be slightly modified, or, [6], the first few outcomes have to be ignored. While our modification of the ML estimator can be re-interpreted in terms of the second, more common solution to the 'startup problem', it has an additional advantage: our modified ML estimators are equivalent to (a) Bayesian MAP estimators with a conjugate prior, and (b) Bayesian mean estimators resulting from a variation of conjugate priors. Thus, for a large class of priors, our result continues to hold if one uses 'prequential Bayes MAP' or 'prequential Bayes mean' rather than prequential ML coding [10]. Rissanen's *predictive MDL* code behaves differently for unordered data, but our result implies that Rissanen's scheme cannot achieve redundancy $\frac{1}{2} \ln n + O(1)$ for arbitrary P either [10]. In fact, we conjecture:

Conjecture 1. Let $\mathcal{M}$ be a regular exponential family with sufficient statistic X and let $\mathcal{P}$ be the set of distributions on $\mathcal{Z}$ such that $E_P[X^4]$ exists. There exists *no "in-model" estimator* such that the corresponding prequential code achieves redundancy $\frac{1}{2} \ln n + O(1)$ for all $P \in \mathcal{P}$.

Here, by an *in-model estimator* we mean an algorithm that takes as input any sample of arbitrary length and outputs an $M \in \mathcal{M}$. Let us contrast this with "out-model estimators": fix some prior on the parameter set Θ_μ and let $P(\mu \mid z_1, \ldots, z_{n-1})$ be the Bayesian posterior with respect to this prior and data $z_1, \ldots, z_{n-1}$. One can think of the Bayesian predictive distribution $P(z_n \mid z_1, \ldots, z_{n-1}) := \int_{\mu \in \Theta_\mu} M_\mu(z_n) P(\mu \mid z_1, \ldots, z_{n-1}) d\mu$ as an *estimate* of the distribution of Z, based on data $z_1, \ldots, z_{n-1}$. But unlike estimators as defined in the conjecture above, the resulting *Bayesian predictive estimator* will in general not be a member of $\mathcal{M}$, but rather of its convex closure: we call it an *out-model estimator.* The redundancy of the Bayesian universal model is equal to the *accumulated Kullback-Leibler (KL) risk* of the Bayesian predictive estimator [10]. Thus, the accumulated KL risk of the Bayesian predictive estimator is $\frac{1}{2} \ln n + O(1)$ even under misspecification. Thus, if our conjecture above holds true, then in-model estimators behave in a fundamentally different way from out-model estimators in terms of their asymptotic risk.

Example 3. The well-known Laplace and Krichevsky-Trofimov estimators for the Bernoulli model [9] define prequential ML models according to Definition 2: they correspond to $x_0 = 1/2, n_0 = 2$, and $x_0 = 1/2, n_0 = 1$ respectively. Yet, they also correspond to Bayesian predictive distributions with uniform prior or Jeffreys' prior respectively. This implies, for example, that the codelength achieved by the Bayesian universal model with Jeffreys' prior and the prequential ML model with $x_0 = 1/2, n_0 = 1$ must *coincide.* We claimed before that the expected regret for a Bayesian universal model is $\frac{1}{2} \log n + O(1)$ if data are i.i.d. $\sim P$, for essentially all distributions P. This may seem to contradict our result which says that the expected regret of the prequential ML model can be $0.5c \log n + O(1)$ with $c \neq 1$ if $P \notin \mathcal{M}$. But there really is no contradiction: since the Bernoulli model happens to contain *all* distributions P on $\{0, 1\}$, we cannot have $P \notin \mathcal{M}$ so Theorem 1 indeed says that $c = 1$ no matter what P we choose. But with more complicated models such as the Poisson or Hardy-Weinberg model, it is quite possible that $P \notin \mathcal{M}$. Then the Bayesian predictive distribution will *not* coincide with any prequential ML model and we can have $c \neq 1$.

2. Practical Relevance We just discussed the theoretical relevance of Theorem 1. But, as discussed in [10], it also has important practical consequences for data compression and MDL model selection. We first consider data compression. Our result indicates that the redundancy can be both larger and smaller than $\frac{1}{2} \ln n$, depending on the variance of the 'true' P. In practical data compression tasks, it is often the case that $P \notin \mathcal{M}$. Then, the prequential ML can behave either better or worse than the Bayesian code, depending on the situation. Let us now consider the interesting case of *MDL model selection* (which is not the same as ¿data compression!) between two non-overlapping parametric models $\mathcal{M}_1$ and $\mathcal{M}_2$ where one of the two models is *correct*: we observe some data z^n generated by a distribution from either $\mathcal{M}_1$ or $\mathcal{M}_2$. We pick the model such that the corresponding prequential codelength of z^n is minimized. Without loss of generality, assume $P \in \mathcal{M}_1$. Then P cannot be in $\mathcal{M}_2$. By Theorem 1, the prequential codelength relative to $\mathcal{M}_1$ is approximately equal to the Bayesian codelength, whereas relative to $\mathcal{M}_2$ the codelengths are quite different. It follows that prequential ML model selection will behave differently from Bayesian model selection. In [7] we performed some model selection experiments between the Poisson and geometric model. The results strongly suggest that the prequential ML codes typically behave worse (and never better!) than the Bayesian or NML codes; we provide a theoretical explanation for this phenomenon (but no formal proof) in [7]. These experiments thus show that *under misspecification our result is relevant even in a well-specified context!*

3. Related Work There are a plethora of results concerning the redundancy and/or the regret for the prequential ML code, for a large variety of models including multivariate exponential families, ARMA processes, regression models and so on. Examples are [15, 11, 17, 13]. In all these papers it is shown that either the regret or the redundancy grows as $\frac{k}{2}\ln n + o(\ln n)$, either in expectation or almost surely. [13] even evaluates the remainder term explicitly. The reason that these results do not contradict ours, is that they invariably concern the case where the generating P is in $\mathcal{M}$, so that automatically $\mathrm{var}_{M^*}(X) = \mathrm{var}_P(X)$.

6 Proofs

Lemma's 1 and 2, used in the proof of Theorem 1, are stated and proved, respectively, in Section 6.1 and 6.2. But we first provide some general results about deviations between average and mean, on which the proofs of Lemma 2 and Theorem 2 are based.

Theorem 3. *Suppose $X, X_1, X_2, \ldots$ are i.i.d. with mean 0. If the first $k \in \mathbb{N}$ moments of X exist, then we have Then $E\left[(\sum_{i=1}^n X_i)^k\right] = O\left(n^{\lfloor \frac{k}{2} \rfloor}\right)$.*

Remark It follows as a special case of Theorem 2 of [18] that $E\left[|\sum_{i=1}^n X_i|^k\right] = O(n^{\frac{k}{2}})$ which almost proves this theorem and which is in fact sufficient for our purposes, as can be seen from the proof of Lemma 2. The advantage of Theorem 3 over Theorem 2 of [18] is that Theorem 3 has an elementary proof. Unfortunately, we had to omit this proof for lack of space. It can be found in [10].

Theorem 4. *Let $X, X_1, \ldots$ be i.i.d., let $\hat{\mu}_n := (n_0 \cdot x_0 + \sum_{i=1}^n X_i)/(n + n_0)$ and $\mu^* = E[X]$. If the first k moments of X exist, then $E[(\hat{\mu}_n - \mu^*)^k] = O(n^{-\lceil \frac{k}{2} \rceil})$.*

Proof. We define $Y_i := X_i - \mu^*$; this can be seen as a new sequence of i.i.d. random variables with mean 0, and $y_0 := x_0 - \mu^*$. Rewrite $E\left[(\hat{\mu}_n - \mu^*)^k\right]$ as

$$E\left[\left(n_0 y_0 + \sum_{i=1}^n Y_i\right)^k\right](n+n_0)^{-k} = O\left(n^{-k}\right)\sum_{p=0}^k \binom{k}{p}(n_0 y_0)^p E\left[\left(\sum_{i=1}^n Y_i\right)^{k-p}\right]$$

$$= O\left(n^{-k}\right)\sum_{p=0}^k \binom{k}{p}(n_0 y_0)^p \cdot O\left(n^{\lfloor \frac{k-p}{2} \rfloor}\right),$$

where in the last step we used Theorem 3. We sum $k+1$ terms of which the term for $p=0$ grows fastest in n, so the expression is $O(n^{-\lceil \frac{k}{2} \rceil})$ as required.

Theorem 5. *Let $X, X_1, \ldots$ be i.i.d. random variables, define $\hat{\mu}_n := (n_0 \cdot x_0 + \sum_{i=1}^{n} X_i)/(n+n_0)$ and $\mu^* = E[X]$. Let $k \in \{0, 2, 4, \ldots\}$. If the first k moments exists then $P(|\hat{\mu}_n - \mu^*| \geq \delta) = O\left(n^{-\lceil \frac{k}{2} \rceil} \delta^{-k}\right)$.*

Proof. Using Markov's inequality and Theorem 4 we get:
$P(|\hat{\mu}_n - \mu^*| \geq \delta) = P\left((\hat{\mu}_n - \mu^*)^k \geq \delta^k\right) \leq E\left[(\hat{\mu}_n - \mu^*)^k\right] \delta^{-k} = O\left(n^{-\frac{k}{2}} \delta^{-k}\right)$.

6.1 Lemma 1: Redundancy for Exponential Families

Lemma 1. *Let U be a prequential ML model and $\mathcal{M}$ be an exponential family as in Theorem 1. We have $\mathcal{R}_U(n) = \sum_{i=0}^{n-1} E_{\hat{\mu}_i \sim P}\left[D(M_{\mu^*} \parallel M_{\hat{\mu}_i})\right]$.*

Proof. We have:

$$\arg\inf_{\mu} E_P\left[-\ln M_\mu(X^n)\right] = \arg\inf_{\mu} E_P\left[\ln \frac{M_{\mu^*}(X^n)}{M_\mu(X^n)}\right] = \arg\inf_{\mu} D(M_{\mu^*} \parallel M_\mu).$$

In the last step we used Proposition 1 below. The divergence is minimized when $\mu = \mu^*$ [12], so we find that (the last step following again from Proposition 1):

$$\mathcal{R}_U(n) = E_P[-\ln U(X^n)] - E_P[-\ln M_{\mu^*}(X^n)] = E_P\left[\ln \frac{M_{\mu^*}(X^n)}{U(X^n)}\right]$$
$$= E_P\left[\sum_{i=0}^{n-1} \ln \frac{M_{\mu^*}(X_i)}{M_{\hat{\mu}_i}(X_i)}\right] = \sum_{i=0}^{n-1} E_P\left[\ln \frac{M_{\mu^*}(X_i)}{M_{\hat{\mu}_i}(X_i)}\right] = \sum_{i=0}^{n-1} \underset{\hat{\mu}_i \sim P}{E}\left[D(M_{\mu^*} \parallel M_{\hat{\mu}_i})\right].$$

Proposition 1. *Let $X \sim P$ with mean μ^*, and let M_μ index an exponential family with sufficient statistic X, so that M_{μ^*} exists. We have:*
$E_P\left[-\ln \frac{M_{\mu^*}(X)}{M_\theta(X)}\right] = D(M_{\mu^*} \parallel M_\theta)$.

Proof. Let $\eta(\cdot)$ denote the function mapping parameters in the mean value parameterization to the natural parameterization. (It is the inverse of the function $\mu(\cdot)$ which was introduced in the discussion of exponential families.) By working out both sides of the equation we find that they both reduce to $\eta(\mu^*)\mu^* + \ln \mathrm{Z}(\eta(\mu^*)) - \eta(\theta)\mu^* - \ln \mathrm{Z}(\eta(\theta))$.

6.2 Lemma 2: Convergence of the sum of the remainder terms

Lemma 2. *Let $R(n)$ be defined as in (10). Then $R(n) = O(1)$.*

Proof. We omit irrelevant constants and the term for the first outcome, which is well-defined because of our modification of the ML estimator. We abbreviate $\frac{d^k}{d\mu^k} D(M_{\mu^*} \| M_\mu) = D^{(k)}(\mu)$ as in the proof of Theorem 1. First we consider the third order term. We write $E_{\delta_i \sim P}$ to indicate that we take the expectation over data which is distributed according to P, of which δ_i is a function. We use Theorem 4 to bound the expectation of ${\delta_i}^3$; under the condition that the first three moments exist, which is assumed to be the case, we obtain,

$$\sum_{i=1}^{n-1} \underset{\delta_i \sim P}{E} \left[{\delta_i}^3 D^{(3)}(\mu^*) \right] = D^{(3)}(\mu^*) \sum_{i=1}^{n-1} E[\delta_i^3] = D^{(3)}(\mu^*) \sum_{i=1}^{n-1} O(i^{-2}),$$

which is $O(1)$, since the constants implicit in $O(\cdot)$ are the same across terms.

The fourth order term is more involved, because $D^{(4)}(\mu)$ is not necessarily constant across terms. To compute it we first distinguish a number of regions in the value space of δ_i: let $\Delta_- = (-\infty, 0)$ and let $\Delta_0 = [0, a)$ for some constant value $a > 0$. If the individual outcomes X are bounded on the right hand side by a value g then we require that $a < g$ and we define $\Delta_1 = [a, g)$; otherwise we define $\Delta_j = [a + j - 1, a + j)$ for $j \geq 1$. Now we must establish convergence of:

$$\sum_{i=1}^{n-1} \underset{\delta_i \sim P}{E} \left[{\delta_i}^4 D^{(4)}(\mu) \right] \quad = \quad \sum_{i=1}^{n-1} \sum_{j} P(\delta_i \in \Delta_j) \underset{\delta_i \sim P}{E} \left[{\delta_i}^4 D^{(4)}(\mu) \mid \delta_i \in \Delta_j \right]$$

If we can establish that the sum converges for all regions Δ_j for $j \geq 0$, then we can use a symmetrical argument to establish convergence for Δ_- as well, so it suffices if we restrict ourselves to $j \geq 0$. First we show convergence for Δ_0. In this case, the basic idea is that since the remainder $D^{(4)}(\mu)$ is well-defined over the interval $\mu^* \leq \mu < \mu^* + a$, we can bound it by its extremum on that interval, namely $m := \sup_{\mu \in [\mu^*, \mu^* + a)} \left| D^{(4)}(\mu) \right|$. Now we get:

$$\left| \sum_{i=1}^{n-1} P(\delta_i \in \Delta_0) E \left[{\delta_i}^4 D^{(4)}(\mu) \mid \delta_i \in \Delta_0 \right] \right| \quad \leq \quad \left| \sum_{i=1}^{n-1} 1 \cdot E \left[{\delta_i}^4 \left| D^{(4)}(\mu) \right| \right] \right|,$$

which is less or equal than $\left| m \sum_i E \left[{\delta_i}^4 \right] \right|$. Using Theorem 4 we find that $E[{\delta_i}^4]$ is $O(i^{-2})$, the sum of which converges. Theorem 4 requires that the first four moments of P exist, but this is guaranteed to be the case: either the outcomes are bounded from both sides, in which case all moments necessarily exist, or the existence of the required moments is part of the condition on the main theorem.

Now we distinguish between the unbounded and bounded cases. First we assume X is unbounded from above. In this case, we must show convergence of:

$$\left| \sum_{i=1}^{n-1} \sum_{j=1}^{\infty} P(\delta_i \in \Delta_j) E \left[{\delta_i}^4 D^{(4)}(\mu) \mid \delta_i \in \Delta_j \right] \right|$$

We bound this expression from above. The δ_i in the expectation is at most $a + j$. Furthermore $D^{(4)}(\mu) = O(\mu^{k-6})$ by assumption on the main theorem, where $\mu \in [a + j - 1, a + j)$. Depending on k, both boundaries could maximize this function, but it is easy to check that in both cases the resulting function is

$O(j^{k-6})$. So we get:

$$\ldots \leq \sum_{i=1}^{n-1} \sum_{j=1}^{\infty} P(|\delta_i| \geq a + j - 1)(a+j)^4 O(j^{k-6})$$

Since we know from the condition on the main theorem that the first $k \geq 4$ moments exist, we can apply Theorem 5 to find that $P(|\delta_i| \geq a + j - 1) = O(i^{-\lceil \frac{k}{2} \rceil}(a+j-1)^{-k}) = O(i^{-\frac{k}{2}})O(j^{-k})$ (since k has to be even); plugging this into the equation and simplifying we obtain $\sum_i O(i^{-\frac{k}{2}}) \sum_j O(j^{-2})$. For $k \geq 4$ this expression converges.

Now we consider the case where the outcomes are bounded from above by g. This case is more complicated, since now we have made no extra assumptions as to existence of the moments of P. Of course, if the outcomes are bounded from both sides, then all moments necessarily exist, but if the outcomes are unbounded from below this may not be true. To remedy this, we map all outcomes into a new domain in such a way that all moments of the transformed variables are guaranteed to exist. Any constant x^- defines a mapping $g(x) := \max\{x^-, x\}$. We define the random variables $Y_i := g(X_i)$, the initial outcome $y_0 := g(x_0)$ and the mapped analogues of μ^* and $\hat{\mu}_i$, respectively: $\mu^\dagger$ is defined as the mean of Y under P and $\tilde{\mu}_i := (y_0 \cdot n_0 + \sum_{j=1}^{i} Y_j)/(i + n_0)$. Since $\tilde{\mu}_i \geq \hat{\mu}_i$, we can bound:

$$\begin{aligned}\left|\sum_i P(\delta_i \in \Delta_1) E\left[{\delta_i}^4 D^{(4)}(\mu) \mid \delta_i \in \Delta_1\right]\right| &\leq \sum_i P(\hat{\mu}_i - \mu^* \geq a) \sup_{\delta_i \in \Delta_1} \left|{\delta_i}^4 D^{(4)}(\mu)\right| \\ &\leq \sum_i P(|\tilde{\mu}_i - \mu^\dagger| \geq a + \mu^* - \mu^\dagger) g^4 \sup_{\delta_i \in \Delta_1} \left|D^{(4)}(\mu)\right| \quad (14)\end{aligned}$$

By choosing x^- small enough, we can bring $\mu^\dagger$ and μ^* arbitrarily close together; in particular we can choose x^- such that $a + \mu^* - \mu^\dagger > 0$ so that application of Theorem 5 is safe. It reveals that the summed probability is $O(i^{-\frac{k}{2}})$ for any even $k \in \mathbb{N}$. Now we bound $D^{(4)}(\mu)$ which is $O((g-\mu)^{-m})$ for some $m \in \mathbb{N}$ by the condition on the main theorem. Here we use that $\mu \leq \hat{\mu}_i$; the latter is maximized if all outcomes equal the bound g, in which case the estimator equals $g - n_0(g - x_0)/(i + n_0) = g - O(i^{-1})$. Putting all of this together, we get $\sup \left|D^{(4)}(\mu)\right| = O((g-\mu)^{-m}) = O(i^m)$; if we plug this into the equation we obtain:

$$\ldots \;\leq\; \sum_i O(i^{-\frac{k}{2}}) g^4 O(i^m) = g^4 \sum_i O(i^{m-\frac{k}{2}})$$

This converges if we choose $k \geq 6m$. We can do this because the construction of $g(\cdot)$ ensures that all moments exist, and therefore certainly the first $6m$.

7 Conclusion and Future Work

We established two theorems about the relative coding redundancy. The two theorems combined state, essentially, that the expected regret of exponential family models behaves differently for the prequential plug-in universal code than for the Bayes, NML/Shtarkov and 2-part codes. This has consequences for coding, for MDL model selection, and for the behavior of the Kullback-Leibler risk. We conjecture that a similar result holds if the prequential code is not based on the ML but any other asymptotically consistent estimator, as long as each estimate is required to lie in the model $\mathcal{M}$. In future work, we hope to extend Theorem 2 to general 1-parameter exponential families with arbitrary sample spaces.

Acknowledgment This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

1. K. Azoury and M. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3), 2001.
2. A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inf. Theory*, 44(6):2743–2760, 1998.
3. N. Cesa-Bianchi and G. Lugosi. Worst-case bounds for the logarithmic loss of predictors. *Journal of Machine Learning*, 43(3):247–264, 2001.
4. B.S. Clarke and A.R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inf. Theory*, IT-36(3):453–471, 1990.
5. T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 1991.
6. A.P. Dawid. Present position and potential developments: Some personal views, statistical theory, the prequential approach. *Journal of the Royal Statistical Society, Series A*, 147(2):278–292, 1984.
7. S. de Rooij and P. Grünwald. An empirical study of MDL model selection with infinite parametric complexity. Available at the CoRR arXiv at `http://xxx.lanl.gov/abs/cs.LG/0501028abs.cs.LG/0501028`, 2005.
8. Y. Freund. Predicting a binary sequence almost as well as the optimal biased coin. In *Proc. Ninth Annual Conf. on Comp. Learning Theory (COLT' 96)*, 1996.
9. P. Grünwald. MDL tutorial. In P. Grünwald, J. Myung, and M. Pitt, editors, *Advances in Minimum Description Length*. MIT Press, 2005.
10. P. Grünwald and S. de Rooij. Asymptotic log–loss of prequential maximum likelihood codes. Available at the CoRR arXiv at `http://xxx.lanl.gov/`, 2005.
11. E.M. Hemerly and M.H.A. Davis. Strong consistency of the PLS criterion for order determination of autoregressive processes. *Ann. Statist.*, 17(2):941–946, 1989.
12. R. Kass and P. Vos. *Geometric Foundations of Asymptotic Inference*. Wiley, 1997.
13. L. Li and B. Yu. Iterated logarithmic expansions of the pathwise code lengths for exponential families. *IEEE Trans. Inf. Theory*, 46(7):2683–2689, 2000.
14. J. Rissanen. Universal coding, information, prediction and estimation. *IEEE Trans. Inf. Theory*, 30:629–636, 1984.
15. J. Rissanen. A predictive least squares principle. *IMA Journal of Mathematical Control and Information*, 3:211–222, 1986.
16. J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, 1989.
17. C.Z. Wei. On predictive least squares principles. *Ann. Statist.*, 20(1):1–42, 1990.
18. P. Whittle. Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability and its Applications*, V(3), 1960.