MDL Model Selection using the ML Plug-in Code

Steven de Rooij rooij@cwi.nl Peter Grünwald pdg@cwi.nl
National Research Institute for Mathematics and Computer Science (CWI)
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

***Abstract*— We analyse the behaviour of the ML Plug-in code, also known as the Rissanen-Dawid prequential ML code, relative to single parameter exponential families $\mathcal{M}$. If the data are i.i.d. according to an (essentially) arbitrary P, then the redundancy grows at $\frac{1}{2}c\log n$. We find that, in contrast to other important universal codes such as the 2-part MDL, Shtarkov and Bayesian codes where $c = 1$, here c equals the ratio between the variance of P and the variance of the element of $\mathcal{M}$ that is closest to P in KL-divergence. We show how this behaviour can impair model selection performance in a simple setting in which we select between the Poisson and geometric models.**

I. Introduction

The ML plug-in code, also known as the Rissanen-Dawid prequential ML code or the predictive MDL code [1], [2] was introduced independently by Rissanen [3] in the context of MDL learning and by Dawid [4], who proposed it as a probability forecasting strategy rather than directly as a code. As a universal coding scheme it is attractive because it is often easier to compute than the Shtarkov or Bayesian codes, while it also avoids the problem of discretisation that turns up in the design of two-part codes based on models with real-valued parameters. Roughly, the plug-in code works by sequentially coding each outcome x_i based on the ML estimator for all previous outcomes $x^{i-1} = x_1, \ldots, x_{i-1}$. There are several results that show that under the right circumstances, the plug-in code is efficient in the sense that its expected regret is to within $O(1)$ of the regret of the minimax optimal code. To explain what this means, let U be a universal code relative to a finite-dimensional parametric model $\mathcal{M}$. The *regret* that U achieves on a data sequence is the difference between the number of bits needed by U and by the element of $\mathcal{M}$ that achieves the shortest codelength. A closely related concept is *redundancy*, which in our context is within $O(1)$ (a constant not depending on the sample size) of the regret, see Section IV. There are many papers about the redundancy and/or expected regret for the plug-in code, for a large variety of models including multivariate exponential families, ARMA processes, regression models and so on. Examples are [5], [6], [7], [8], [9]. In all these papers the plug-in code is shown to achieve an asymptotic expected regret or redundancy of $\frac{k}{2}\lg n$, where k is the number of parameters of the model and n is the sample size. This matches the behaviour of the Shtarkov, Bayesian and two-part universal codes and is optimal in several ways, see [1]. These findings notwithstanding our main result (Theorem 1 below) states that, at least for single-parameter exponential family models, the redundancy in fact grows with $\frac{1}{2}\lg n \cdot \frac{\mathrm{var}_P X}{\mathrm{var}_M X}$, where M is the distribution in $\mathcal{M}$ that is closest to P in Kullback-Leibler divergence, i.e. it minimises $D(P\|M)$. Note that in contrast to all the cited works, we do not assume that $P \in \mathcal{M}$: the model may be *misspecified*. Yet *if* $P \in \mathcal{M}$, then we have $M = P$ so that the redundancy grows like it does in the other universal models. But when doing model selection between non-overlapping models, of necessity (at least) one of the two models does not contain the generating process. In this case, the Shtarkov, Bayes and universal codes typically still achieve asymptotic expected regret $\frac{1}{2}\lg n$, whereas our result states that plugin codes behave differently. We illustrate how this leads to suboptimal inference results of the ML plug-in code using computer experiments on a toy model selection problem in which we compare the Poisson and geometric models, and data are generated according to a distribution that is in one of the two. In fact, the results of these experiments suggested the anomalous behaviour of the plug-in code to us in the first place. Section II describes the experimental setup, of which results appear in Section III. Then we formally state our theorem on the redundancy with the necessary conditions and a proof sketch in Section IV. We end with a conclusion and areas of further research.

II. Models and Selection Criteria

In [10], which elaborates on [11], we empirically compare several different model selection criteria for selection between the Poisson and geometric models. We test these criteria in computer experiments in which samples of i.i.d. outcomes are repeatedly drawn from either a Poisson or a geometric distribution with some fixed mean. In other words, one of the models under consideration contains the true distribution. We then use one of the criteria under consideration to select a model for the empirical data. We repeat this scenario many times, and for each criterion observe a frequency of misclassification: the number of times it selected the wrong model divided by the number of times it was used. This frequency is interpreted as an approximation of the probability of error; more on this in Section III. While in [10] all selection criteria have the same role, here we focus on the criterion that is based on the ML plug-in code, considering other criteria only to provide a context for evaluation. We will describe these criteria in more detail below, but first we provide some general information about the two models under consideration.

We parameterise the Poisson and geometric families of distributions by the mean $\mu \in (0, \infty)$, to allow for easy comparison. This is possible because for both models, the empirical mean (average) of the observed data is a sufficient statistic. The probability mass functions with mean μ of the Poisson and geometric distributions are, respectively:

$$P_{\mathrm{P}}(x \mid \mu) = \frac{e^{-\mu}\mu^x}{x!} \qquad P_{\mathrm{G}}(x \mid \mu) = \frac{\mu^x}{(\mu+1)^{x+1}} \quad (1)$$

As is easy to see, the ML estimator $\hat{\mu}$ for both distributions is the average of the data.

We will now develop the four selection criteria that are discussed and compared in Section III.

a) Known μ criterion: The known μ criterion is not based on a code that is universal for the Poisson and geometric models, but rather 'cheats', i.e. it uses knowledge of the mean of the generating distribution. It is simply defined to select the Poisson model for a data sequence $x^n = x_1, \ldots, x_n$ iff $P_{\mathrm{P}}(x^n \mid \mu) > P_{\mathrm{G}}(x^n \mid \mu)$, where the value of μ is generously provided free of charge. Since this criterion is thus based on an expected optimal code, its performance can be interpreted as a baseline of sorts.

b) ML/BIC criterion: The ML criterion selects the Poisson model iff $P_{\mathrm{P}}(x^n \mid \hat{\mu}) > P_{\mathrm{G}}(x^n \mid \hat{\mu})$, where the ML estimator $\hat{\mu}$ is the average of the sample x^n. The Bayesian Information Criterion (BIC, [2]) reduces to ML because the number of parameters are the same for both models. This criterion is not based on a valid universal code since it does not encode the parameter estimate. It thus invites overfitting: the model that can represent the most information about the sample in its parameter estimate has unfair advantage. In our setting this turns out to be the Poisson model: we find that the ML criterion misclassifies geometric data as Poisson more often than the other way around. Nevertheless both ML and BIC are often used in practise because of their simplicity, and we will find that, in our context, even this naive approach may be preferable over the ML plug-in criterion.

c) ML Plug-in criterion: In the ML plug-in universal code outcomes are coded sequentially using the probability distribution indexed by the ML estimator for the previous outcomes [4], [3]; for a general introduction see [2]. The codelength for the entire sequence $-\lg P^{\text{plug}}(x^n)$ thus becomes the sum over $i \in \{1, \ldots, n\}$ of the number of bits $-\lg P(x_i \mid \hat{\mu}(x^{i-1}))$ needed to encode outcome x_i using the code based on the ML estimator on x^{i-1}.

For both the Poisson model and the geometric model, the maximum likelihood estimator is not well-defined until after a nonzero outcome has been observed (since 0 is not inside the allowed parameter range). This startup problem is an inherent feature of the predictive plug-in approach [4], [12]. It means that we need to use another code for the first few outcomes. It is not really clear how we can use the model assumption (Poisson or geometric) here, so we pick a simple code for the nonnegative integers that does not depend on the model. This will result in the same codelength for both models; therefore it does not influence which model is selected. Since there are usually only very few initial zero outcomes, we may reasonably hope that the results are not too distorted by this approach which is in line with the suggestions in [4].

Using this universal code we define the ML plug-in criterion such that it selects the Poisson model iff $-\lg P_{\mathrm{P}}^{\text{plug}}(x^n) < -\lg P_{\mathrm{G}}^{\text{plug}}(x^n)$.

d) Bayesian criterion: In the Bayesian framework we select a prior $w(\theta)$ on the unknown parameter and compute the marginal likelihood $P^{\text{Bayes}}(x^n) = \int_{\Theta} P(x^n \mid \theta) w(\theta) d\theta$, which corresponds to a universal code with lengths $-\lg P^{\text{Bayes}}(x^n)$. Objective Bayesian reasoning suggests we use Jeffreys' prior [13] for several reasons; one reason is that it is uniform over all 'distinguishable' elements of the model, which implies that the obtained results are independent of the parametrisation of the model [14]. Unfortunately, the normalisation factor in Jeffreys' prior diverges for both the Poisson model and the geometric model. But if one is willing to accept a so-called *improper* prior, which is not normalised, then it is possible to compute a perfectly proper Bayesian posterior, after observing the first outcome x_1, and use that as a prior to compute the marginal likelihood of the rest of the data $x_2^n = x_2, \ldots, x_n$. The resulting criterion selects the Poisson model iff $-\lg P_{\mathrm{P}}^{\text{Bayes}}(x_2^n \mid x_1) < -\lg P_{\mathrm{G}}^{\text{Bayes}}(x_2^n \mid x_1)$, where, writing $s = x_1 + \ldots + x_n$, we use:

$$\begin{aligned} P_{\mathrm{P}}^{\text{Bayes}}(x_2^n \mid x_1) &= \frac{\Gamma(s+\frac{1}{2})}{\Gamma(x_1)} \Big/ \left(n^{s+\frac{1}{2}} \prod_{i=2}^{n} x_i! \right) \\ P_{\mathrm{G}}^{\text{Bayes}}(x_2^n \mid x_1) &= (x_1 + \tfrac{1}{2}) \frac{\Gamma(s+\frac{1}{2})\Gamma(n)}{\Gamma(s+n+\frac{1}{2})} \end{aligned}$$

For lack of space we must refer the reader to [10] for more details. We include this criterion because it seems to be a reasonably well motivated approach from a theoretical point of view; it also was the criterion that we experimentally found to perform best in [10].

III. Experimental Results

As described in Section II, we test model selection criteria based on samples which are drawn from a Poisson or geometric distribution. It follows from theory in Cover and Thomas [15, Theorem 12.9.1 on page 312 and text thereafter] that, if the data are generated using Poisson[μ] then the probability of error of the known μ criterion decreases exponentially in the sample size, with some error exponent, and if the data are generated with Geom[μ] then the overall probability is exponentially decreasing with the same exponent. This suggests that our approach is only feasible for very limited values of mean and sample size, because if either one of them is very large, it becomes exceedingly unlikely that a sample is misclassified by any method. We have chosen to use mean 8 and a sample size that varies from 4 to 20 because we found that for these numbers, one million trials suffice to get an accurate result. With the theoretical behaviour of the error rate for the known μ criterion still in mind, we chose to plot our results on a logarithmic scale in Figure 1, expecting a straight line for the known μ criterion with equal slopes on the left and right hand graphs. This proves to be the case.

We list a number of additional striking features of Figure 1 below:

- The ML/BIC criterion is biased in favour of Poisson, with an even lower error frequency than the known μ criterion

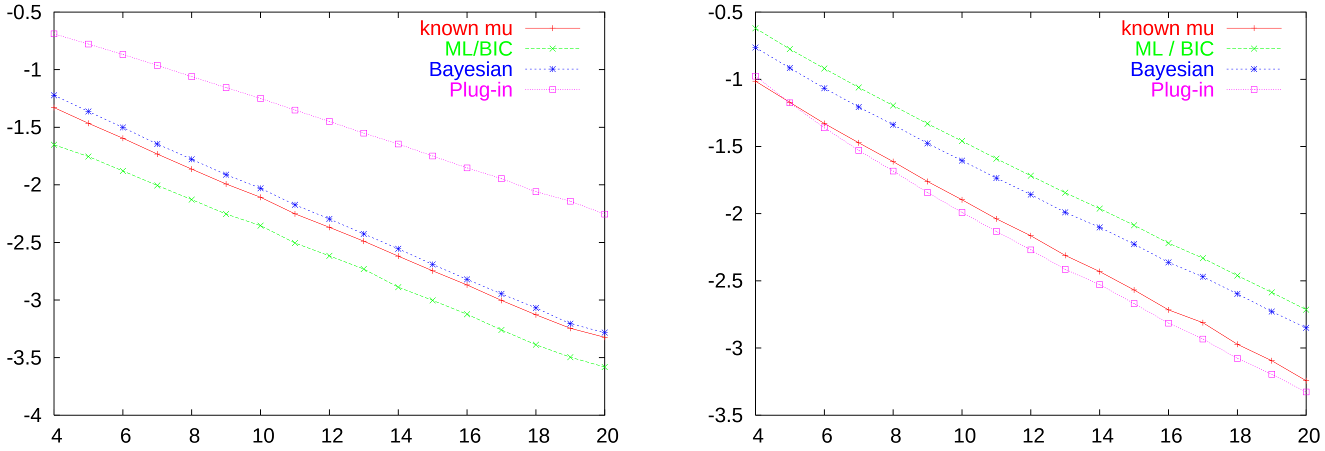

Fig. 1. $\log_{10}$ of the error frequency for different sample sizes when the data are generated with mean 8 by Poisson (left) or geometric (right).

when the generating distribution is Poisson. (The term "biased" is used loosely here.)

- The Bayesian criterion is always closer to known μ than ML/BIC. It compensates for the bias inherent in the ML criterion, even though it is not totally unbiased itself.
- The lines for the ML plug-in criterion show an unusual slope. ML plug-in clearly makes fewer errors when the sample is geometric than when it is Poisson, an effect that increases with the sample size. At $n = 20$ the point is reached where a sample is about $10^{3.33-2.25} \approx 12$ times more likely to be misclassified as geometric than as Poisson.

While the behaviour of the ML plug-in criterion may seem appropriate if the data are subjectively judged more likely to come from a geometric distribution, there is actually a strong argument that *even under those circumstances it is not*. Suppose that we put a fixed prior on the generating distribution, with nonzero probability for both distributions. The marginal probability of error is a linear combination of the probabilities of error for the two generating distributions; as such it is dominated by the probability of error with the *worst* exponent. So if minimising the probability of error is our goal, then we must conclude that the behaviour of the plug-in criterion is suboptimal. (On a side note, minimising the probability of error with respect to a fixed prior is *not* the goal of classical hypothesis testing, since in that setting the two hypotheses do not play a symmetrical role.) Figure 2 illustrates this effect. To visually separate the behaviour of the four criteria more clearly, we have subtracted the $\log_{10}$ of the error frequency of the known μ criterion from all results, the known μ criterion thus literally functions as a baseline here. The graph shows how, even if there is a 75% chance that the data are geometrically distributed, then the ML plug-in criterion still has a worse (marginal) probability of error than any of the other criteria from $n \geq 16$ onward. This includes ML/BIC which is biased in the direction that conflicts with the prior on the generating distribution! Additional unreported experiments suggest that this effect only becomes stronger if the mean is higher, as one would expect on the basis of the theory in the next section.

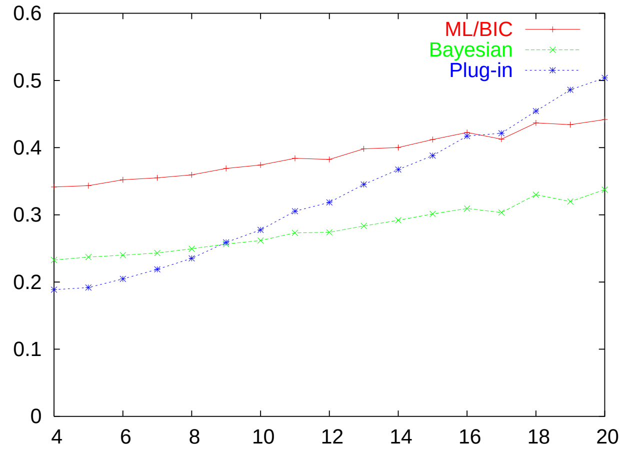

Fig. 2. The difference in the $\log_{10}$ of the error frequency between the known μ criterion and the other criteria. Here 75% of the data with mean 8 are generated using the Poisson model but even so the plug-in criterion behaves worse than ML for $n \geq 16$.

These results suggest that the ML plug-in codelength exhibits unusual asymptotics, rendering it a (somewhat) suboptimal candidate universal model at least in the context of model selection. In the next section we will confirm that this is indeed the case in the form of our main result.

IV. Main Result

Suppose $Z, Z_1, Z_2, \ldots$ are i.i.d. according to an arbitrary distribution P, and are modelled by some exponential family of distributions $\mathcal{M}$. Let $\mathcal{Z}$ be a set of outcomes, taking values 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. Single parameter exponential family models are families of distributions on $\mathcal{Z}$ defined relative to a random variable X (called 'sufficient statistic') as defined above. The model can be reparameterised by the expected value of X, which is called the *mean value parametrisation*.

Our main result covers single parameter, regular exponential family models, for which the mean value parameter space Θ_μ is an open interval of $\mathbb{R}$. We write $\mathcal{M} = \{M_\mu \mid \mu \in \Theta_\mu\}$ where M_μ is the distribution with mean value parameter μ. Examples of this wide family of models include the Poisson, geometric and multinomial families, and the model of all Gaussian distributions with a fixed mean or fixed variance.

We write E_P as a shorthand for $E_{Z\sim P}$, the expectation of Z under distribution 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. Note that in our main result, Theorem 1 below, we do not assume that the data generating distribution P admits a density. Finally, we compute codelengths in nats rather than bits, in order to simplify equations.

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 $n = 1, 2, \ldots$. In the definition, we use the notation $x_i := X(z_i)$.

Definition 1 (ML plug-in model): Let $\mathcal{M} = \{M_\mu \mid \mu \in \Theta_\mu\}$ be an exponential family with mean value parameter domain Θ_μ. Given $\mathcal{M}$ and constants $x_0 \in \Theta_\mu$ and $n_0 > 0$, we define the *ML plug-in 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 $Z^n = z^n$, and

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

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 [16]. Here we define our plug-in 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 circumvent the "startup problem"; note that one may choose this initial outcome to be the average of the first few outcomes that were actually observed, with their actual multiplicity, if a workaround such as the one described in Section II is used. Furthermore it also allows us to adapt our results to other point estimators (see Section V).

Definition 2 (Relative redundancy): We define *relative redundancy* with respect to P of a code U that is universal on a model $\mathcal{M}$, as:

$$\mathcal{R}_U(n) := E_P[L_U(Z^n)] - \inf_{\mu\in\Theta_\mu} E_P[-\ln M_\mu(Z^n)], \quad (2)$$

where L_U is the length function of U.

We use the term *relative redundancy* rather than just *redundancy* to emphasise that it measures redundancy relative to the element of the model that minimises the codelength rather than to P, which is not necessarily an element of the model. A definition of *expected regret* is obtained if the infimum over μ is moved inside the expectation, which yields a slightly larger quantity. We have proven for finite sample spaces that the difference converges to a constant and conjecture that this continues to hold for infinite sample spaces [17]. This remains a matter for further research.

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. Finally let U denote the ML plug-in model with respect to $\mathcal{M}$. If $\mathcal{M}$ and P satisfy Condition 1 below, then

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

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 – we checked that it holds for Poisson, geometric, Gaussian, exponential and Pareto models, as long as the fourth moment of the generating P exists. The reason why the condition is needed will become clear in the proof sketch:

Proof Sketch

The precise proof Theorem 1, given in [17], is very technical. Here we merely describe the underlying ideas, which are relatively simple. Consider the case $n_0 = 0$, so that for $n \geq 1$, $\hat\mu_n$ is just the standard ML estimator. Let z^i be the initial i outcomes of an arbitrary sequence z^n. As is well-known, a straightforward second-order Taylor expansion of $D(M_{\mu^*}\|M_{\hat\mu(z^i)})$ around μ^* gives

$$D(M_{\mu^*}\|M_{\hat\mu_i}) = \frac{1}{2}I(\mu^*)(\hat\mu_i - \mu^*)^2 + \text{Remainder}, \quad (3)$$

where $I(\mu^*)$ is the Fisher information in one observation, evaluated at μ^*. For exponential families in their mean-value parameterisation, another standard result [16] says that for all μ, we have that $I(\mu)\mathrm{var}_{M_\mu} X = 1$. Therefore, ignoring the remainder term and the term for $i = 0$, we get

$$\sum_{i=0}^{n-1} \underset{\hat\mu_i\sim P}{E}[D(M_{\mu^*}\|M_{\hat\mu_i})] \approx \frac{1}{2}\sum_{i=1}^{n-1}\frac{E_P(\hat\mu_i - \mu^*)^2}{\mathrm{var}_{M_{\mu^*}} X}$$

$$= \frac{1}{2}\frac{\mathrm{var}_P X}{\mathrm{var}_{M_{\mu^*}} X}\sum_{i=1}^{n-1}\frac{1}{i} = \frac{1}{2}\frac{\mathrm{var}_P X}{\mathrm{var}_{M_{\mu^*}} X}\ln n + O(1), \quad (4)$$

Here the second equality follows because for exponential families, the ML estimator $\hat\mu_i$ is just the empirical average $i^{-1}\sum x_i$. Thus, Theorem 1 follows if we can show (a) that the left-hand side of (4) is equal to the relative redundancy $\mathcal{R}_U(n)$ and (b) that, as $n \to \infty$, the remainder terms in (3), summed over n as in (4), form a convergent series (i.e. sum to something finite). Result (a) follows relatively easily by rewriting the sum using the chain rule for relative entropy and using the fact that X is a sufficient statistic. The truly difficult part of the proof is (b), which 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 $T = X$, where 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 $\hat{\mu}$.

V. Plug-in Models with Other Estimators

The Bayesian MAP estimator: The Bayesian maximum a posteriori estimator can, if a conjugate prior is used, always be interpreted as an ML estimator based on the sample and some additional 'fake data' ([18]; see also the notion of *ESS (Equivalent Sample Size) Priors* discussed in, for example, [19]). Therefore, the ML plug-in model as defined above can also be interpreted as a MAP plug-in model for that class of priors, and the whole analysis carries over to that setting.

The Bayesian mean estimator: It follows by the work of Hartigan [20, Chapter 7] on the so-called 'maximum likelihood prior', that by slightly modifying conjugate priors, we can construct priors such that the Bayesian *mean* rather than MAP estimator is of the form of our modified ML estimator.

VI. Conclusion and Future Work

We have described computer experiments on model selection between the Poisson and geometric models using four different selection criteria, which are detailed in Section II. The criterion based on the ML plug-in code exhibited behaviour that we found unexpected. Subsequent analysis showed that, while other important universal codes such as the NML/Shtarkov and Bayesian codes, achieve a regret of $\frac{1}{2}\ln n$, this code achieves a relative redundancy of $\frac{1}{2}\frac{\mathrm{var}_P X}{\mathrm{var}_{M_{\mu^*}} X}\ln n$, where n is the sample size (Section IV).

The notions of expected regret and redundancy are formally defined in Section IV. The notions are closely related: in [17] we prove that the difference converges to a constant if the sample space is finite (this is known for the case that $P \in \mathcal{M}$; we show it still holds, but with a different constant, if $P \notin \mathcal{M}$). We conjecture that this continues to be true when the sample space is infinite. However, the precise relationship between these two notions remains a matter for future research.

In the introduction we cited a substantial amount of literature in which the regret for the ML plug-in code is proven to grow with $\frac{1}{2}\ln n$. While this may seem to contradict our results, in fact it does not: In those articles, settings are considered where $P \in \mathcal{M}$, and under such circumstances our own findings predict precisely that behaviour.

Finally, our result seems to be reasonably robust regarding slight variations in the definition of the ML plug-in code: in our framework the so-called "start-up problem" (the unavailability of an ML estimate for the first few outcomes) is resolved by introducing fake initial outcomes; in Section V we show how, by choosing appropriate fake initial data, our framework also covers plug-in codes that use other point estimators such as the Bayesian MAP and mean estimators.

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

References

[1] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2743–2760, 1998, special Commemorative Issue: Information Theory: 1948-1998.

[2] P. D. Grünwald, "MDL tutorial," in *Advances in Minimum Description Length: Theory and Applications*, P. D. Grünwald, I. J. Myung, and M. A. Pitt, Eds. MIT Press, 2005.

[3] J. Rissanen, "Universal coding, information, prediction and estimation," *IEEE Transactions on Information Theory*, vol. 30, pp. 629–636, 1984.

[4] A. Dawid, "Present position and potential developments: Some personal views, statistical theory, the prequential approach," *J. Royal Stat.Soc., Ser. A*, vol. 147, no. 2, pp. 278–292, 1984.

[5] J. Rissanen, "A predictive least squares principle," *IMA Journal of Mathematical Control and Information*, vol. 3, pp. 211–222, 1986.

[6] L. Gerensce'r, "Order estimation of stationary gaussian ARMA processes using Rissanen's complexity," Computer and Automation Institute of the Hungaian Academy of Sciences, Tech. Rep., 1987.

[7] E. Hemerly and M. Davis, "Strong consistency of the PLS criterion for order determination of autoregressive processes," *The Annals of Statistics*, vol. 17, no. 2, pp. 941–946, 1989.

[8] C. Wei, "On predictive least squares principles," *The Annals of Statistics*, vol. 20, no. 1, pp. 1–42, 1990.

[9] L. Li and B. Yu, "Iterated logarithmic expansions of the pathwise code lengths for exponential families," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2683–2689, 2000.

[10] S. de Rooij and P. Grünwald, "An empirical study of MDL model selection with infinite parametric complexity," Submitted for publication. Available at CORR Computer Science arXiv, http://arxiv.org/abs/cs.LG/0501028, 2005.

[11] A. D. Lanterman, "Hypothesis testing for Poisson versus geometric distributions using stochastic complexity," in *Advances in Minimum Description Length: Theory and Applications*, P. D. Grünwald, I. J. Myung, and M. A. Pitt, Eds. MIT Press, 2005.

[12] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Company, 1989.

[13] J. M. Bernardo and A. F. Smith, *Bayesian Theory*. John Wiley, 1994.

[14] V. Balasubramanian, "Statistical inference, Occam's Razor, and statistical mechanics on the space of probability distributions," *Neural Computation*, vol. 9, pp. 349–368, 1997.

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley, 1991.

[16] R. E. Kass and P. W. Vos, *Geometrical Foundations of Asymptotic Inference*. Wiley, 1997.

[17] P. Grünwald and S. de Rooij, "Asymptotic log-loss of prequential maximum likelihood codes," Submitted for publication. Available at CORR Computer Science arXiv, 2005.

[18] J. Berger, *Statistical Decision Theory and Bayesian Analysis*, revised and expanded second ed., ser. Springer Series in Statistics. New York: Springer-Verlag, 1985.

[19] P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and P. Grünwald, "On predictive distributions and Bayesian networks," *Journal of Statistics and Computing*, vol. 10, pp. 39–54, 2000.

[20] J. Hartigan, *Bayes Theory*. New York: Springer-Verlag, 1983.