JMLR: Workshop and Conference Proceedings vol 23 (2012) 7.14–7.17 25th Annual Conference on Learning Theory

# Commentary on "The Optimality of Jeffreys Prior for Online Density Estimation and the Asymptotic Normality of Maximum Likelihood Estimators"

**Peter Grünwald** PDG@CWI.NL
*CWI, Amsterdam and Leiden University*

**Editor:** Shie Mannor, Nathan Srebro, Robert C. Williamson

## 1. Background

In the field of *prediction with expert advice*, a standard goal is to sequentially predict data as well as the best expert in some reference set of 'expert predictors'. *Universal data compression*, a subfield of information theory, can be thought of as a special case. Here, the set of expert predictors is a statistical model, i.e. a family of probability distributions, and the predictions are scored using the logarithmic loss function, which, via the Kraft inequality, gives the procedure an interpretation in terms of data compression. A prediction strategy is a function that, for each $n$, given data $x^n \equiv x_1, \ldots, x_n$, outputs a "predictive" probability distribution $p(\cdot \mid x^n)$ for $X_{i+1}$. For a given model $\mathcal{M}$, the *Shtarkov* or *Normalized Maximum Likelihood (NML)* strategy relative to $\mathcal{M}$, is the prediction strategy that achieves the *minimax optimal individual-sequence regret* relative to $\mathcal{M}$. NML has a number of drawbacks, detailed below, and is therefore often approximated by more convenient strategies such as *Sequential Normalized Maximum Likelihood (SNML)* or the *Bayesian* strategy. The latter predicts using the Bayesian predictive distribution for the model $\mathcal{M}$, defined relative to some prior $\pi$, which is often taken to be *Jeffreys' prior* — in that case we abbreviate it to J.B. The text below has been written so as to be (hopefully) understandable for readers who do not know too many details of these concepts; for such details, see e.g. Grünwald (2007) and/or Kotlowski and Grünwald (2011) (KG from now on).

## 2. The HB Result and Why it is Interesting

Hedayati and Bartlett (2012) (HB from now on) elegantly characterize the parametric statistical families for which NML, SNML, and J.B. coincide: under a mild regularity condition on the model, which is standard in the statistical literature, they coincide iff the SNML strategy, viewed as a random process, is *exchangeable*. This is indeed the case for several models, such as the Gaussian location family, the full Gaussian family in which both mean and variance are parameters, and the exponential distributions. Yet it is not the case for other simple models such as, e.g., the Bernoulli distributions. The HB result is important for several reasons:

1. *Worst-case optimality of Jeffreys' prior.* Jeffreys' prior was introduced into Bayesian statistics by H. Jeffreys as early as 1946, as a prior to be used for parametric models

when no clear prior knowledge about the parameters is available. Jeffreys' motivated the prior from differential-geometric considerations. In the early 1990s it became clear that it also has a completely different information-theoretic interpretation as the prior which is minimax optimal for data-compression purposes, both when optimality is measured in an individual sequence sense (called 'minimax log-loss regret' in learning theory) and when optimality is measured in an expected sense (called 'minimax redundancy' in information theory) — see Grünwald (2007), in particular Chapter 6 and 8, for appropriate references and general background. However, the optimality of Jeffreys' prior was consistently shown to hold *only (a) asymptotically and (b) only if the parameter space is truncated* (formally the parameter space has to be *ineccsi*, see Grünwald (2007)). HB's result shows that, under some conditions, J.B. is minimax regret-optimal also nonasymptotically, for all sample sizes, and also for full, nontruncated parameter spaces. They also show that if a Bayesian strategy (code) is optimal at all sample sizes, it *has* to be based on Jeffreys' prior, thus further clarifying the special status of this prior. If we consider nontruncated parameter spaces, then Jeffreys' prior is often *improper*: its density integrates to $\infty$ rather than 1. The HB result shows that J.B. can be minimax-regret optimal nevertheless. Interestingly however, if one measures optimality by the more traditional (expected) minimax redundancy, then the use of improper Jeffreys' prior in nontruncated parameter space is not necessarily optimal Liang and Barron (2004); see (Grünwald, 2007, Section 11.4.3.) for more discussion.

2. *Avoiding infinite regret: a better basis for MDL model selection.* When given a finite number of models and some data, the *Minimum Description Length (MDL) Principle for model selection* (Barron et al., 1998) tells us to associate each model with its respective NML strategy, and pick the model for which the corresponding NML strategy gives the smallest codelength (cumulative log loss). Unfortunately, for most interesting parametric models (including, e.g., the Gaussian model), the minimax coding regret is infinite, and hence no NML strategy exists. The only obvious way to avoid this problem is, once again, to truncate the parameter space, but this has an arbitrary flavor to it. While this undefined NML problem convinced some researchers that MDL was no good at all, others started looking for generalizations of NML that are always well-defined. The first hint that such a generalization might be possible came, in my view, with Kakade et al. (2006), who showed that, in the abstract setting of Gaussian processes, where the standard minimax regret is hopelessly infinite, one can define a notion of 'conditional' regret and a 'conditional' version of NML, which achieves the minimax conditional regret. In my 2007 MDL book Grünwald (2007), I worked out their proposal in more detail and came to the conclusion that it led to a viable generalization of the NML concept that was applicable to arbitrary parametric models. In fact, Chapter 11 of my book considers more than seven proposals for extending NML, the proposal of Kakade et al. being called 'conditional NML-II'. By now, it is becoming increasingly clear that this is the only 'natural' generalization, and indeed, justifiably, HB simply adopt the name 'NML' for 'conditional NML-II'. To avoid any confusion, I henceforth call it 'generalized NML'.

Generalized NML still has two problems associated with it: first, it is often hard to compute, and second, unlike Bayesian strategies, for many models it depends on the *horizon*, i.e. the amount of data that will eventually be seen, and which in practice might be unknown. For these reasons, Rissanen and Roos (2007); Roos and Rissanen (2008) developed an approximation to NML which they called *sequential NML*, or SNML for short. It is horizon-independent and often more easily computable than the full NML. KG showed that SNML provides a reasonably good approximation of generalized NML, and also found that for some special models, NML, SNML and the Bayesian strategy with Jeffreys' prior are all equivalent. The importance of the HB result is that tells us exactly *when* this is the case.

3. *Philosophical Considerations.* At each point in time $i$, the SNML strategy for predicting(coding) the next outcome $x_{i+1}$ may be viewed as the strategy that would lead to minimax optimal regret if the *horizon* where known to be $i+1$, i.e. as if one would stop the sequential prediction right after the currently performed step. Thus, it is essentially a *last-step minimax* strategy in the sense of Takimoto and Warmuth (2000). A priori, it is not at all clear why, at each day, predicting as if that day were the last of a sequence of days on which one had to predict, could ever be optimal if that day is in fact *not* the last of those days. Yet the HB result shows that in many cases, it is. (see Section 1 of KG for more on this issue).

The HB result solves the first open problem stated in Section 6 of KG. Yet, like many a good solution to an open problem, it immediately gives rise to a new question:

## 3. Open Problem raised by HB result

Exchangeability of the SNML random process is an elegant characterization of equivalence between NML, SNML and Jeffreys-Bayes, but it is tedious to check and no list can be found in the literature of families satisfying it. *Is there some other, equivalent characterization which can be immediately read off of the definition of the family? Or if not, perhaps there is some other notion equivalent to SNML-exchangeability, which is more standard in the literature, so that a list of families satisfying it is readily available?*[1]

Below I shall give some first ideas on possible ways to investigate these questions. I hope HB (or somebody else) will follow up on it!

**Towards a Characterization Easier to Check than SNML Exchangeability** The examples of parametric models with exchangeable SNML given by BH and KG are: the exponential model, the Gaussian location family and the full Gaussian family. In addition, P. Harremoës (personal communication) established that for the inverse Gaussian distributions, (a) for every finite sample size, the integrand in Jeffreys prior is proportional to the integrand in the NML strategy; and (b) Jeffreys prior is proper and the NML strategy is well-defined. Taken together this strongly suggests that for these models, NML and J.B. coincide, which, with HB's result, would imply that SNML is exchangeable. Assuming this is the case, we find that those families for which SNML exchangeability has been established all fall into the following groups: they are

[1] Many thanks to W. Kotlowski who first brought up this question.

1. infinitely divisible exponential families;
2. exponential families that are closed under convolution, i.e. if $X_1, \ldots X_n \sim$ i.i.d. $p_\theta$, where $p_\theta$ is a member of the family, then the sum $Y := \sum_{i=1}^{n} X_i$ has a distribution $p_{\theta'}$ for another member of the family
3. *exponential dispersion models* with continuously-valued dispersion (Jørgensen, 1997);
4. Most intriguingly, they are all 1- and 2-parameter subfamilies of the 3-parameter family of *Tweedie*[2] *distributions* (Jørgensen, 1997).

It would be of substantial mathematical interest to sort out whether SNML exchangeability is equivalent to, or implied by, or implies, any of the four classes defined above. A first step in sorting this out would be to check whether the geometric family is SNML-exchangeable: while the geometric family is infinitely divisible, hence in class (1), it is not in class (2), (3) or (4). A second step would be to check whether the Gamma family and the Poisson family are SNML-exchangeable — both families are in all four classes... in any case, these considerations suggest that SNML-exchangeability might hold for a wide variety of models, and that it is in fact a deep notion whose importance extends far beyond the log-loss prediction setting.

## References

A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.

P. Grünwald. *The Minimum Description Length Principle.* MIT Press, 2007.

F. Hedayati and P. Bartlett. The optimality of Jeffreys prior for online density estimation and the asymptotic normality of maximum likelihood estimators. In *Proceedings COLT '12*, 2012.

B. Jørgensen. *The Theory of Dispersion Models.* Chapman & Hall, 1997.

S. Kakade, M. Seeger, and D. Foster. Worst-case bounds for Gaussian process models. In *Proceedings NIPS 2005*, 2006.

W. Kotlowski and P. Grünwald. Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. In *Proceedings COLT '11*, pages 761–779, Budapest, 2011.

F. Liang and A. R. Barron. Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory*, 50: 2708–2726, 2004.

J. Rissanen and T. Roos. Conditional NML universal models. In *Proceedings ITA-07*, pages 337–341, 2007.

T. Roos and J. Rissanen. On sequentially normalized maximum likelihood models. In *Proceedings WITMSE-08*, 2008.

E. Takimoto and M. Warmuth. The minimax strategy for Gaussian density estimation. In *Proceedings COLT '00*, 2000.

[2] Many thanks to P. Harremoës who first realized that Tweedie distributions have "something" to do with NML and Jeffreys' prior.