Following the Flattened Leader

Wojciech Kotłowski
Centrum Wiskunde & Informatica
kotlowsk@cwi.nl

Peter Grünwald
Centrum Wiskunde & Informatica
pdg@cwi.nl

Steven de Rooij
University of Cambridge
steven@statslab.cam.ac.uk

Abstract

We analyze the regret, measured in terms of log loss, of the maximum likelihood (ML) sequential prediction strategy. This "follow the leader" strategy also defines one of the main versions of Minimum Description Length model selection.

We proved in prior work for single parameter exponential family models that (a) in the misspecified case, the redundancy of follow-the-leader is *not* $\frac{1}{2}\log n + O(1)$, as it is for other universal prediction strategies; as such, the strategy also yields suboptimal individual sequence regret and inferior model selection performance; and (b) that in general it is not possible to achieve the optimal redundancy when predictions are constrained to the distributions in the considered model.

Here we describe a simple "flattening" of the sequential ML and related predictors, that does achieve the optimal worst case *individual sequence* regret of $(k/2)\log n + O(1)$ for k parameter exponential family models for bounded outcome spaces; for unbounded spaces, we provide almost-sure results. Simulations show a major improvement of the resulting model selection criterion.

1 Introduction

Let $x_1, x_2, \ldots \in \mathcal{X}^*$, be a sequence of outcomes revealed one at a time. After observing $x^n = x_1, x_2, \ldots, x_n$, a forecaster assigns a probability distribution on $\mathcal{X}$, denoted $P(\cdot \mid x^n)$. Then, after x_{n+1} is revealed, the forecaster incurs the *log loss* $-\log P(x_{n+1} \mid x^n)$. The performance of the strategy is measured relative to the best in a reference set of strategies, which we call the *model* $\mathcal{M}$. The difference between the accumulated loss of the prediction strategy and the best strategy in the model is called the *regret*. The goal is to minimize the regret in the worst case over all possible data sequences.

Sequential prediction of individual sequences with log loss has been extensively studied in learning theory, in the framework of *prediction with expert advice* (Azoury & Warmuth, 2001; Cesa-Bianchi & Lugosi, 2001; Cesa-Bianchi & Lugosi, 2006). However, it has also been playing an important role in the information theory: a key result based on the Kraft-McMillan inequality (see, e.g., (Cover & Thomas, 1991)) states that, ignoring rounding issues, for every uniquely decodable codelength function L there is a probability distribution P such that $L(x) = -\log P(x)$ and vice versa[1]. Thus, at least when $\mathcal{X}$ is countable, any prediction strategy can also be thought of as a *universal source coding algorithm*; the cumulative logarithmic loss corresponds exactly to the incurred codelength. As Rissanen's theory of Minimum Description Length (MDL) learning (Barron et al., 1998; Grünwald, 2005) is based on universal coding, a sequential prediction strategy with log loss defines an MDL model selection criterion. Similarly, in statistics Dawid's theory of prequential model assessment (Dawid, 1984) is based on sequential prediction. Thus we use the terms "prediction strategy" and "code" interchangeably, as we do for "accumulated log loss" and "codelength".

For parametric models $\mathcal{M} = \{P_\theta \mid \theta \in \Theta\}$, there are three "universal codes" (prediction strategies with low regret) that are particularly well known in the source coding and MDL communities: (1) after putting a prior distribution π on the model parameters, one can predict using the *Bayesian predictive distribution* $P_{\text{BAYES}}(\cdot \mid x^n) = \int_\Theta P_\theta(\cdot \mid x^n)\pi(\theta \mid x^n)\,\mathrm{d}\theta$. (2) If there is a known horizon (maximal number of outcomes), the Shtarkov code (Shtarkov, 1987), also known as the Normalized Maximum Likelihood code (Rissanen, 1996), can be defined. This universal code *minimizes* the worst-case regret. (3) Given an estimator $\bar{\theta} : \mathcal{X}^i \to$

[1] Throughout this text, all logarithms are to the base e and we use nats rather than bits as units of information; however, all results presented here are valid for logarithms of any base.

Θ, one can sequentially select an element of the model using the estimator and use that to predict the next outcome, i.e. $P_{\text{PLUG-IN}}(\cdot \mid x^n) = P_{\bar{\theta}(x^n)}(\cdot \mid x^n)$. Such "plug-in codes" were introduced independently in the context of MDL learning (Rissanen, 1984) and in the context of prequential model validation (Dawid, 1984). If we take $\bar{\theta}(x^n)$ equal to the maximum likelihood (ML) estimator $\hat{\theta}(x^n)$, then the resulting strategy is called the "ML plug-in strategy", which corresponds to the "follow the leader" strategy in learning theory terminology (Kalai & Vempala, 2003; Hutter & Poland, 2005). Strategy (3) always predicts using an element of the model, whereas strategies (1) and (2) do not.

Under weak regularity conditions on the sequence of outcomes, the Bayesian and NML strategies have been shown to achieve asymptotically optimal worst-case regret $(k/2)\log n + O(1)$, where k is the number of parameters of the model (Rissanen, 1989; Rissanen, 1996; Grünwald, 2007). As a consequence, the same $(k/2)\log n + O(1)$-result holds in expectation and almost surely, if the data are sampled from some distribution P^*, as long as P^* satisfies some very weak regularity conditions. In particular, P^* is *not* required to lie in the model $\mathcal{M}$: the results still hold if $P^* \notin \mathcal{M}$, i.e. the "model is wrong", or, as statisticians call it, "the misspecified case". Now if P^* does lie in $\mathcal{M}$, then the same $(k/2)\log n + O(1)$-regret is achieved in expectation under P^* for a large variety of plug-in models including multivariate exponential families, ARMA processes, regression models and so on; examples are (Rissanen, 1986; Hemerly & Davis, 1989; Wei, 1990; Li & Yu, 2000). However, in contrast to the Bayesian and NML results, the plug-in result does *not* hold under misspecification, i.e. if $P^* \notin \mathcal{M}$. We reported earlier (Grünwald & de Rooij, 2005) that under misspecification, already for single parameter exponential family models, the expected regret of the ML plug-in strategy is $\frac{1}{2}c\log n + O(1)$ where c is the variance of an outcome under the true distribution P^* divided by the variance under the element of the model P_θ that minimizes the Kullback-Leibler divergence $D(P^*\|P_\theta)$. Moreover, it is shown by (Grünwald & Kotłowski, 2010) that *no* plug-in estimator can achieve $c = 1$ (thus it does not help to replace maximum likelihood predictions by, say, Bayesian posterior mean or moment-estimator-based predictions). This behavior is especially undesirable when the plug-in ML estimator is used to define an MDL or prequential model selection procedure, because in those circumstances, as we explained in Section 6, it is by definition not safe to assume that $P^* \in \mathcal{M}$. This is quite clearly visible in the results of model selection experiments described by (De Rooij & Grünwald, 2006), where the plug-in based version of MDL is significantly outperformed by MDL based on Bayesian and NML strategies.

While the ML plug-in strategy does not achieve the desired expected regret, (Grünwald & Kotłowski, 2010) describe a simple modification of the plug-in prediction strategy that does do so, in the somewhat specific case where $k = 1$ and the outcomes are generated i.i.d. from some distribution P^*. In this paper, we extend this result to the much more general scenario where k can be larger than 1 and where we consider worst-case individual sequence regret rather than expected regret. Our only assumption is that the outcome space $\mathcal{X}$ is bounded in some sense. Following (Grünwald & Kotłowski, 2010), we propose the *flattened* ML prediction strategy, a modification of the ML strategy that puts it slightly outside $\mathcal{M}$, and in Theorem 11 we show that this strategy achieves optimal asymptotic minimax regret $(k/2)\log n + O(1)$. We also show that when the outcomes are generated i.i.d. from some distribution P^* of which the first four central moments exist, we can remove the assumption of bounded $\mathcal{X}$ and still our prediction strategy achieves the optimal regret $(k/2)\log n$ with probability one.

Our result is important in practice since, in contrast to the Bayesian predictive distribution, the flattened ML strategy is in general just as easy to compute as the ML estimator itself. The flattened ML strategy can be used to define an efficient MDL model selection criterion; we repeated the model selection experiments of (De Rooij & Grünwald, 2006) including this new criterion to find that it displays acceptable performance, unlike the ML plug-in strategy.

Related Work The idea of changing a "follow the leader"-strategy by modifying the leader is not new (Kalai & Vempala, 2003; Hutter & Poland, 2005); however, our "flattened" leader is quite different from the "perturbed" leader described in these earlier papers, and also the setting is quite different: flattened leaders make sense relative to parametric statistical models, which may be regarded as an uncountable set of experts satisfying continuity requirements; perturbed leaders make sense relative to finite or countable sets of otherwise unrelated experts, and the regret bounds obtained in the latter settings are quite different from the $(k/2)\log n$ regret obtained here. The flattened leader is more closely related to the predictive densities considered by (Vidoni, 2008) and (Corcuera & Giummolè, 1999). These authors provide $O(1/n)$-modifications of the ML density that are similar (but nonequivalent) to ours, and they investigate the behavior of these modifications in terms of expected KL-divergence rather than cumulative regret, in a stochastic, rather than an individual sequence setting.

The paper is organized as follows. We introduce the mathematical context for our results in Section 2. We subsequently define the flattened ML strategy 3 and prove that the regret is $(k/2)\log n + O(1)$ in the individual sequence setting with bounded sample space. We give an example of how this estimator can be used in practice in Section 4, where we apply it to the model of Bernoulli distributions and show how its worst case regret develops as a function of the sample size. In Section 5 we return to theory by providing an

"almost sure" analogue of our individual sequence result, where we can relax the boundedness assumption somewhat. The prediction strategy based on the flattened ML estimator can be used to define an MDL model selection criterion; in Section 6 this criterion is evaluated in a series of model selection experiments, showing that it overcomes many of the weaknesses of the ML plug-in prediction strategy without flattening. We end with a conclusion in Section 7.

2 Notation and Definitions

Let $\mathcal{X}$ be a set of outcomes, taking values either in a finite or countable set, or in a subset of Euclidean space. Exponential family models are families of distributions on $\mathcal{X}$ defined relative to a random variable $\phi : \mathcal{X} \to \mathbb{R}^k$ (called "sufficient statistic"), and a function $h : \mathcal{X} \to [0, \infty)$. Let $\mathrm{Z}(\eta) = \int_{x \in \mathcal{X}} e^{\eta^T \phi(x)} h(x) \, \mathrm{d}x$ (the integral to be replaced by a sum for countable $\mathcal{X}$), and $\Theta_{\text{nat}} = \{\eta \in \mathbb{R}^k : \mathrm{Z}(\eta) < \infty\}$.

Definition 1 (Exponential family) *The* exponential family (Barndorff-Nielsen, 1978) *with* sufficient statistic ϕ *and* carrier h *is the family of distributions with densities* $P_\eta(x) = \frac{1}{\mathrm{Z}(\eta)} e^{\eta^T \phi(x)} h(x)$*, where* $\eta \in \Theta_{\text{nat}}$*.* Θ_{nat} *is called the* natural parameter space*. The family is called* regular *if* Θ_{nat} *is an open and convex subset of* $\mathbb{R}^k$*, and if the representation* $P_\eta(x)$ *is* minimal*, i.e. the functions* $\phi_i(x), i = 1, \ldots, k$ *are linearly independent.*

We only consider regular exponential families, but this qualification will henceforth be omitted. Examples include the Poisson, geometric and multinomial families, and the model of multidimensional Gaussian distributions. Moreover, without loss of generality, we will make the simplifying assumption that $\phi(x) \equiv x$, i.e. the exponential family is in the canonical form. All results in this paper are valid for more general ϕ.

The statistic $\phi(X) \equiv X$ is sufficient for η (Barndorff-Nielsen, 1978). This suggests reparameterizing the distribution by the expected value of X, which is called the *mean value parameterization*. The function $\mu(\eta) = E_{P_\eta}[X]$ maps parameters in the natural parameterization to the mean value parameterization. It is a diffeomorphism (Barndorff-Nielsen, 1978), therefore the mean value parameter space Θ_{mean} is also an open set of $\mathbb{R}^k$. We write $\mathcal{M} = \{P_\mu \mid \mu \in \Theta_{\text{mean}}\}$ where P_μ is the distribution with mean value parameter μ.

The sequence of outcomes $x_1, \ldots, x_n$ is abbreviated by x^n (x^0 denotes the empty sequence). At every iteration $n = 0, 1, 2, \ldots$, the prediction $P(\cdot \mid x^n)$ depends on the past outcomes x^n and has the form of a probability distribution on $\mathcal{X}$, therefore it can be considered as a conditional of the joint distribution of outcomes in $\mathcal{X}^n$, which is $P(x^n) = \prod_{i=1}^n P(x_i | x^{i-1})$. Conversely, any probability distribution P on the set $\mathcal{X}^n$ defines a prediction strategy induced by its conditional distributions $P(\cdot \mid x^i)$ for $0 \leq i < n$ (Cesa-Bianchi & Lugosi, 2006; Grünwald, 2007).

We are now ready to define the plug-in prediction strategy.

Definition 2 (Plug-in prediction strategy) *Let* $\mathcal{M} = \{P_\mu \mid \mu \in \Theta_{\text{mean}}\}$ *be an exponential family with mean value parameter domain* Θ_{mean}*. Given* $\mathcal{M}$*, and a function* $\bar{\mu} : \mathcal{X}^* \to \Theta_{\text{mean}}$*, define the* plug-in prediction strategy $P_{\text{PLUG-IN}}$ *by setting, for all* n*, all* x^{n+1}*:*

$$P_{\text{PLUG-IN}}(x_{n+1} \mid x^n) = P_{\bar{\mu}(x^n)}(x_{n+1}).$$

We will be mostly concerned with the maximum likelihood (ML) plug-in prediction strategy:

Definition 3 (ML prediction strategy) *Given* $\mathcal{M}$ *and constants* $x_0 \in \Theta_{\text{mean}}$*,* $n_0 > 0$ *we define the* ML prediction strategy $P_{\text{ML}}(x_{n+1}|x^n)$ *as a plug-in strategy with* $\bar{\mu} = \hat{\mu}_n^\circ$*, where*

$$\hat{\mu}_n^\circ(x^n) = \frac{n_0 x_0 + \sum_{i=1}^n x_i}{n_0 + n}.$$

To understand this definition, note that for exponential families in the mean value parameterization, for any sequence of data, the maximum likelihood parameter $\hat{\mu}_n$ is given by the average $\hat{\mu}_n = n^{-1} \sum x_i$ of the observations (Barndorff-Nielsen, 1978). Here we define our plug-in model in terms of a smoothed maximum likelihood estimator $\hat{\mu}_n^\circ$ that introduces a 'fake initial outcome' x_0 with multiplicity n_0 in order to avoid infinite log loss for the first few outcomes, and to ensure that the plug-in ML code of the first outcome is well-defined. The estimator $\hat{\mu}_n^\circ$ can also be interpreted as "maximum a posteriori" estimator, as it maximizes the posterior distribution with appropriate conjugate prior. In practice we can take $n_0 = 1$ but our result holds for any $n_0 > 0$.

Definition 4 (Regret) *We define* regret *with respect to a sequence* x^n *of a prediction strategy* P *relative to the model* $\mathcal{M}$*, as a difference between the accumulated log loss of* P *and the accumulated log loss of the best*

strategy from $\mathcal{M}$:

$$\begin{aligned}\mathcal{R}(P; x^n) &= \sum_{i=1}^{n} -\log P(x_i|x^{i-1}) - \inf_{\mu\in\Theta_{\text{mean}}} \sum_{i=1}^{n} -\log P_\mu(x_i) \\ &= -\log P(x^n) - \inf_{\mu\in\Theta_{\text{mean}}} -\log P_\mu(x^n).\end{aligned} \tag{1}$$

From the definition, the minimizer:

$$\hat{\mu}_n = \arg\inf_{\mu\in\Theta_{\text{mean}}} -\log P_\mu(x^n) = \arg\max_{\mu\in\Theta_{\text{mean}}} P_\mu(x^n)$$

is the ordinary maximum likelihood estimator, $\hat{\mu}_n = n^{-1}\sum x_i$. Note, however, that $P_{\hat{\mu}_n}$ is *not* the same as the ML plug-in strategy with $n_0 = 0$: since $P_{\hat{\mu}_n}$ uses the ML estimator *based on the whole sequence* to predict all outcomes from the start, its predictions are generally much better than for the ML plug-in criterion.

Under some mild assumptions about the outcomes, two important prediction strategies, NML (*normalized maximum likelihood*) and Bayes, achieve regrets that are (in an appropriate sense) close to optimal. To be more specific, we must introduce the notion of *ineccsi* subsets of Θ_{mean} and the related sequences (Grünwald, 2007). These are formally defined as follows.

Definition 5 (Ineccsi subsets and sequences) *Let $\mathcal{M}$ be a model with a smooth parameterization Θ (e.g., $\mathcal{M}$ may be an exponential family and Θ may represent its mean-value parameterization). The subset $\Theta_0 \subset \Theta$ is* ineccsi *("interior (is) non-empty; closure (is) compact subset of interior") if:*

1. *the interior of Θ_0 is nonempty;*
2. *the closure of Θ_0 is a compact subset of the interior of Θ.*

The sequence $x_1, x_2, \ldots$ is a Θ_0-sequence if there exists m, such that for all $n \geq m$, the ML estimator $\hat{\mu}_n$ exists, is unique and satisfies $\hat{\mu}_n \in \Theta_0$.

Now, the formal definitions of NML and Bayes strategies follow:

Definition 6 (NML prediction strategy) *Given $\mathcal{M}$, an ineccsi subset $\Theta_0 \subset \Theta_{\text{mean}}$, and a finite horizon n, define the* NML prediction strategy with respect to Θ_0 *as:*

$$P_{\text{NML}}(x^n) = \frac{\sup_{\mu\in\Theta_0} P_\mu(x^n)}{\int_{\mathcal{X}^n} \sup_{\mu\in\Theta_0} P_\mu(z^n)\,\mathrm{d}z^n}.$$

Definition 7 (Bayes prediction strategy) *Given $\mathcal{M}$ and a probability distribution $\pi(\mu)$ on Θ_{mean}, define the* Bayes prediction strategy *as:*

$$P_{\text{BAYES}}(x^n) = \int_{\Theta_{\text{mean}}} P_\mu(x^n)\pi(\mu)\,\mathrm{d}\mu.$$

Note that the NML does not define a random process, since its predictions depend on the horizon n, i.e. marginalizing the NML distribution with some horizon larger than n over the first n outcomes does not yield the NML distribution with horizon n. This is not an issue with the Bayesian strategy, which does define a random process.

The following theorem characterizes the regret of the NML and Bayes prediction strategies:

Theorem 8 *Let $\mathcal{M} = \{P_\mu \mid \mu \in \Theta_{\text{mean}}\}$ be a k-dimensional exponential family with mean-value parameter space Θ_{mean}. Let Θ_0 be an ineccsi subset of Θ_{mean} and let $x_1, x_2, \ldots$ be a Θ_0-sequence. Then,*

$$\mathcal{R}(P, x^n) = \frac{k}{2}\log n + O(1), \tag{2}$$

where P is either the NML strategy with respect to Θ_0 with horizon n, or the Bayesian prediction strategy, based on a prior with support Θ_{mean}.

For a proof, see e.g. (Grünwald, 2007). (2) is the famous 'k over 2 log n formula', refinements of which lie at the basis of practical approximations to MDL and Bayesian learning, most notably BIC (Grünwald, 2007). Since the NML strategy in fact *minimizes* the worst-case regret, it follows that a worst-case of $\frac{k}{2}\log n + O(1)$ is optimal. We remark that, if $x_1, x_2, \ldots$ do not form an ineccsi sequence, then the empirical mean of the x_i tends to the boundary of the parameter space. In that case, the behavior of the Bayesian strategy critically depends on the prior, e.g. with the Bernoulli model and the uniform (Laplace) prior, the worst-case regret becomes $\log n$; with Jeffreys' prior, it is still $(1/2)\log n + O(1)$ (Freund, 1996); see also Section 4. In this

paper we concentrate on the ineccsi-case, where the data remain bounded away from the boundary, and the $(k/2)\log n$ regret is achieved for Bayes with *all* priors with support Θ_{mean}.

It is known that when outcomes are generated by one of the distributions in $\mathcal{M}$, the plug-in strategy satisfies (2) as well. However it was shown by (De Rooij & Grünwald, 2005; Grünwald & de Rooij, 2005) that when the outcomes are generated i.i.d. by some distribution P^* outside $\mathcal{M}$, the ML plug-in strategy P_{ML} behaves suboptimally. Specifically, its expected regret satisfies, for all $\mu^* \in \Theta_{\text{mean}}$,

$$E_{P^*}[\mathcal{R}(P_{\text{ML}}, n)] \geq \frac{1}{2}\frac{\text{var}_{P^*}X}{\text{var}_{P_{\mu^*}}X}\log n + O(1), \tag{3}$$

where $\mu^* = E_{P^*}[X]$ is the element in Θ_{mean} minimizing KL divergence $D(P^*\|P_\mu)$ for $\mu \in \Theta_{\text{mean}}$. A similar result in a different context was already proved earlier by (Wei, 1990). The result was later extended to hold (essentially) for all plug-in prediction strategies (not just ML plug-in) by (Grünwald & Kotłowski, 2010). As (3) is satisfied in the average case, the situation can only become worse in the individual sequence case.

3 The Flattened ML Strategy achieves Optimal Regret

While the plug-in strategies behave suboptimally as shown in the previous section, it remains possible that a small modification of the plug-in strategy, which puts the predictions slightly outside $\mathcal{M}$, might lead to the optimal regret (2). As a first example, consider the Bayesian predictive distribution when $\mathcal{M}$ is the normal family with fixed variance σ^2. In this case (see, e.g. (Grünwald, 2007)), the Bayesian code based on prior $\mathcal{N}(\mu_0, \tau_0^2)$ has a simple form $P_{\text{BAYES}}(x_{n+1}|x^n) = f(x_{n+1})$, where f is the density of normal distribution $\mathcal{N}(\mu_n, \tau_n^2)$, with

$$\mu_n = \left(\left(\sum_{i=1}^n x_i\right) + \frac{\sigma^2}{\tau_0^2}\mu_0\right) / \left(n + \frac{\sigma^2}{\tau_0^2}\right), \quad \text{and} \quad \tau_n^2 = \sigma^2 / \left(n + \frac{\sigma^2}{\tau_0}\right).$$

Thus, the Bayesian predictive distribution is itself a Gaussian with mean equal to the smoothed maximum likelihood estimator $\hat{\mu}_n^\circ$ with $n_0 = \sigma^2/\tau_0^2$ and $x_0 = \mu_0$, albeit with a slightly larger variance $\sigma^2 + O(1/n)$. This shows that for the normal family with fixed variance, there exists an "almost" plug-in strategy, which satisfies (2). This led to the conjecture, also in (Grünwald, 2007), that something similar should be possible for exponential families in general. In this section we show that this is indeed the case: we propose a simple modification of the ML strategy, obtained by predicting x_{n+1} using a slightly "flattened" version P_{FML} of the ML strategy P_{ML}, defined as:

Definition 9 (Flattened ML prediction strategy) *Given $\mathcal{M}$ and constants $x_0 \in \Theta_{\text{mean}}$, $n_0 > 0$, we define the* flattened ML prediction strategy P_{FML} *by setting for all n:*

$$P_{\text{FML}}(x_{n+1}|x_n) := P_{\hat{\mu}_n^\circ}(x_{n+1})\frac{n + n_0 + \frac{1}{2}(x_{n+1} - \hat{\mu}_n^\circ)^T I(\hat{\mu}_n^\circ)(x_{n+1} - \hat{\mu}_n^\circ)}{n + n_0 + \frac{k}{2}},$$

where $I(\mu)$ is the Fisher information matrix for model $\mathcal{M}$.

We first check that P_{FML} is properly defined:

Lemma 10 *For every $n = 0, 1, \ldots$, $P_{\text{FML}}(x_{n+1}|x^n)$ represents a valid probability distribution, i.e. it is nonnegative and the sum/integral over $x_{n+1} \in \mathcal{X}$ is equal to* 1.

Proof: For every $x \in \mathcal{X}$, $P_{\text{FML}}(x \mid x^n) \geq 0$ because the information matrix $I(\hat{\mu}_n^\circ)$ is positive definite. To show that $P_{\text{FML}}(\cdot \mid x^n)$ normalizes to 1, let E_μ denote the expectation with respect to P_μ, i.e. $E_{\hat{\mu}_n^\circ}[f(X)] = \int_{\mathcal{X}} f(x)P_{\hat{\mu}_n^\circ}(x)\,\mathrm{d}x$. Then:

$$\begin{aligned}
\int P_{\text{FML}}(x_{n+1}|x^n)\,\mathrm{d}x_{n+1} &= E_{\hat{\mu}_n^\circ}\left[\frac{P_{\text{FML}}(X|x^n)}{P_{\hat{\mu}_n^\circ}(X)}\right] \\
&= \left(n + n_0 + \frac{k}{2}\right)^{-1}\left(n + n_0 + \frac{1}{2}E_{\hat{\mu}_n^\circ}\left[(X - \hat{\mu}_n^\circ)^T I(\hat{\mu}_n^\circ)(X - \hat{\mu}_n^\circ)\right]\right) \\
&= \left(n + n_0 + \frac{k}{2}\right)^{-1}\left(n + n_0 + \frac{1}{2}E_{\hat{\mu}_n^\circ}\left[\text{Tr}\left\{(X - \hat{\mu}_n^\circ)(X - \hat{\mu}_n^\circ)^T I(\hat{\mu}_n^\circ)\right\}\right]\right) \\
&= \left(n + n_0 + \frac{k}{2}\right)^{-1}\left(n + n_0 + \frac{1}{2}\text{Tr}\left\{(\text{Cov}_{P_{\hat{\mu}_n^\circ}}X)I(\hat{\mu}_n^\circ)\right\}\right) = 1,
\end{aligned}$$

where $\mathrm{Cov}_{P_{\hat{\mu}_n^\circ}} X$ is the covariance matrix of $P_{\hat{\mu}_n^\circ}$ and the last equality uses a standard result (Barndorff-Nielsen, 1978) for the mean-value parameterization of exponential families which says that for all $\mu \in \Theta_{\text{mean}}$, $\mathrm{Cov}_\mu X = I^{-1}(\mu)$. ■

The predictions of the corresponding flattened ML strategy are not harder to calculate than those of the ordinary ML strategy, which is often much easier than calculating the predictive distribution of the Bayesian strategy. Moreover, we show below in Theorem 11 that under some mild assumptions about the sequence of outcomes, the flattened ML strategy always achieves the optimal regret, satisfying (2). To this end, we need the following two propositions:

Proposition 1 *Let $X \sim P^*$ with mean μ^*, and let $\mathcal{M}$ be the exponential family with sufficient statistic X and mean-value parameter space Θ_{mean}, such that $\mu^* \in \Theta_{\text{mean}}$. Then for every $\mu \in \Theta_{\text{mean}}$ we have:*

$$E_{P^*}\left[-\log P_\mu(X) + \log P_{\mu^*}(X)\right] = D(\mu^* \| \mu),$$

where $D(\cdot \,\|\, \cdot)$ is the Kullback-Leibler divergence.

Proof: By working out both sides of the equation using Definition 1, we find that they both reduce to $\eta(\mu^*)\mu^* - \log \mathrm{Z}(\eta(\mu^*)) - \eta(\mu)\mu^* + \log \mathrm{Z}(\eta(\mu))$. ■

Proposition 2 *Let $\mathcal{M}$ be the exponential family with sufficient statistic X and mean-value parameter space Θ_{mean}. Then for every $\mu, \mu^* \in \Theta_0$, where Θ_0 is the ineccsi subset of Θ_{mean}, we have:*

$$D(\mu^* \| \mu) = \frac{1}{2}(\mu - \mu^*)^T I(\mu)(\mu - \mu^*) + O(\|\mu - \mu^*\|^3).$$

Proof: We need two standard results regarding the properties of KL divergence (see, e.g. (Barndorff-Nielsen, 1978; Grünwald, 2007)): for any $\mu, \mu^* \in \Theta_{\text{mean}}$, it holds:

1. $D(\mu^* \| \mu) \geq 0$ and the equality only holds for $\mu = \mu^*$,
2. For exponential families, $\partial^2 D(\mu^* \| \mu) / \partial \mu_i \partial \mu_j = I_{ij}(\mu)$.

By Taylor expanding $D(\mu^* \| \mu)$ around μ^* up to the second order, we get:

$$D(\mu^* \| \mu) = D(\mu^* \| \mu^*) + \nabla D(\mu^* \| \mu)^T \big|_{\mu=\mu^*} (\mu - \mu^*) + \frac{1}{2}(\mu - \mu^*)^T I(\bar{\mu})(\mu - \mu^*),$$

for some $\bar{\mu}$ between μ and μ^*. Due to the first property the zeroth order term disappears; the second order term also disappears because the gradient vanishes at the minimum, so we have:

$$\begin{aligned} D(\mu^* \| \mu) &= \frac{1}{2}(\mu - \mu^*)^T I(\bar{\mu})(\mu - \mu^*) = \frac{1}{2}(\mu - \mu^*)^T I(\mu)(\mu - \mu^*) + \frac{1}{2}(\mu - \mu^*)^T \left(I(\bar{\mu}) - I(\mu)\right)(\mu - \mu^*) \\ &\leq \frac{1}{2}(\mu - \mu^*)^T I(\mu)(\mu - \mu^*) + \frac{1}{2}\|I(\bar{\mu}) - I(\mu)\| \|\mu - \mu^*\|^2, \end{aligned} \tag{4}$$

where $\|\cdot\|$ denotes vector or matrix norm, depending on the context. Taylor expanding $I(\bar{\mu})$ around μ up to the first order gives $I(\bar{\mu}) = I(\mu) + \nabla I(\tilde{\mu})^T (\bar{\mu} - \mu)$, for some $\tilde{\mu}$ between $\bar{\mu}$ and μ. From that we get:

$$\|I(\bar{\mu}) - I(\mu)\| \leq \|\nabla I(\tilde{\mu})\| \|\bar{\mu} - \mu\| \leq C \|\bar{\mu} - \mu\|, \tag{5}$$

where $C = \sup_{\mu \in \Theta_0} \|\nabla I(\mu)\|$ is finite since closure of Θ_0 is compact and all derivatives of the information matrix are continuous. It follows from the definition of $\bar{\mu}$ that $\|\bar{\mu} - \mu\| \leq \|\mu - \mu^*\|$; using this in (5) and plugging the result into (4) finishes the proof. ■

Theorem 11 *Let $\mathcal{M}$ be a k-dimensional exponential family with with mean-value parameter space Θ_{mean}. Let Θ_0 be an ineccsi subset of Θ_{mean} and let $x_1, x_2, \ldots$ be a Θ_0-sequence, i.e. for all $n \geq m$, $\hat{\mu}_n \in \Theta_0$. Moreover, assume that the outcomes are bounded, $\|x_i\| \leq B$ for all $i = 1, 2, \ldots$ Then the flattened ML strategy P_{FML} with $x_0 \in \Theta_0$ achieves asymptotically optimal regret, i.e.*

$$\mathcal{R}(P_{\text{FML}}, x^n) = \frac{k}{2} \log n + O(1), \tag{6}$$

where the constant under $O(\cdot)$ depends only on B, Θ_0 and m, while it does not depend on the outcomes x^n.

Proof: Let x_0^n be the sequence of outcomes, composed of n_0 fake outcomes x_0 and the original sequence x^n, i.e. $x_0^n = x_0, \ldots, x_0, x_1, \ldots, x_n$, and we denote $x_{-i} = x_0$, $i = 0, \ldots, n_0 - 1$. We will use it to cope with the fact that we predict with $\hat{\mu}_n^\circ$ using the ML strategy, while we compare to $\hat{\mu}_n$ in the definition of regret (1). Although $\hat{\mu}_n^\circ$ and $\hat{\mu}_n$ are not the same, they are sufficiently similar that if we replace the term $\log P_{\hat{\mu}_n}(x^n)$ with the term $\log P_{\hat{\mu}_n^\circ}(x_0^n)$ in the definition (1); the difference is only small. Let us denote such a modified regret by $\mathcal{R}'(P_{\text{FML}}, x^{n+1})$. We have:

$$\begin{aligned}
\mathcal{R}'(P_{\text{FML}}, x^n) - \mathcal{R}(P_{\text{FML}}, x^n) &= \sum_{i=-n_0-1}^{n} -\log P_{\hat{\mu}_n^\circ}(x_i) - \sum_{i=1}^{n} -\log P_{\hat{\mu}_n}(x_i) \\
&= -n_0 \log P_{\hat{\mu}_n^\circ}(x_0) + \sum_{i=1}^{n} \log \frac{P_{\hat{\mu}_n}(x_i)}{P_{\hat{\mu}_n^\circ}(x_i)} = O(1) + nE_{P_{\text{emp}}}\left[\log \frac{P_{\hat{\mu}_n}(X)}{P_{\hat{\mu}_n^\circ}(X)}\right] \\
&= O(1) + nD(\hat{\mu}_n \| \hat{\mu}_n^\circ) = O(1) - \frac{n}{2}(\hat{\mu}_n - \hat{\mu}_n^\circ)^T I(\hat{\mu}_n^\circ)(\hat{\mu}_n - \hat{\mu}_n^\circ) + nO(\|\hat{\mu}_n - \hat{\mu}_n^\circ\|^3),
\end{aligned}$$

where P_{emp} is the empirical distribution function, which puts mass $1/n$ on every outcome of x^n, $E_{P_{\text{emp}}}[X] = \hat{\mu}_n$, and we used Proposition 1 with $P^* \equiv P_{\text{emp}}$, and then Proposition 2. Using the fact that:

$$\|\hat{\mu}_n - \hat{\mu}_n^\circ\| = \frac{n_0 \|(x_0 - \hat{\mu}_n)\|}{n + n_0} \leq \frac{2n_0 B}{n},$$

and since $\hat{\mu}_n^\circ \in \Theta_0$ for $n \geq m$, we get for all $n \geq m$:

$$\frac{n}{2}(\hat{\mu}_n - \hat{\mu}_n^\circ)^T I(\hat{\mu}_n^\circ)(\hat{\mu}_n - \hat{\mu}_n^\circ) \leq \frac{n}{2}\|I(\hat{\mu}_n^\circ)\|\|\hat{\mu}_n - \hat{\mu}_n^\circ\|^2 \leq \frac{4n_0^2 B^2}{2n} \sup_{\mu \in \Theta_0} \|I(\mu)\| = O(n^{-1}),$$

where $\|I(\mu)\|$ denotes the matrix norm and we used the fact that $\sup_{\mu \in \Theta_0} \|I(\mu)\|$ is finite due to compactness of the closure of Θ_0 and continuity of information matrix.

Thus, we proved that $\mathcal{R}'(P_{\text{FML}}, x^n) - \mathcal{R}(P_{\text{FML}}, x^n) = O(1)$. To show (6), it now suffices to show that $\Delta(n) = \mathcal{R}'(P_{\text{FML}}, x^{n+1}) - \mathcal{R}'(P_{\text{FML}}, x^n) = \frac{k}{2n} + O(n^{-2})$, where the constant under $O(\,\cdot\,)$ does not depend on the outcomes x^n. Then, since $\log n \leq \sum_{i=1}^{n} \frac{1}{i} \leq \log n + 1$, and $\sum_n n^{-2}$ converges, (6) follows. From the definition, we have:

$$\begin{aligned}
\Delta(n) &= -\log P_{\text{FML}}(x_{n+1}|x^n) - \sum_{i=-n_0+1}^{n+1} -\log P_{\hat{\mu}_{n+1}^\circ}(x_i) + \sum_{i=-n_0+1}^{n} -\log P_{\hat{\mu}_n^\circ}(x_i) \\
&= \log\left(1 + \frac{k}{2(n+n_0)}\right) - \log\left(1 + \frac{1}{2(n+n_0)}(x_{n+1} - \hat{\mu}_n^\circ)^T I(\hat{\mu}_n^\circ)(x_{n+1} - \hat{\mu}_n^\circ)\right) + \sum_{i=-n_0+1}^{n+1} \log \frac{P_{\hat{\mu}_{n+1}^\circ}(x_i)}{P_{\hat{\mu}_n^\circ}(x_i)}.
\end{aligned}$$

Let us denote $\xi_n = \frac{1}{2(n+n_0)}(x_{n+1} - \hat{\mu}_n^\circ)^T I(\hat{\mu}_n^\circ)(x_{n+1} - \hat{\mu}_n^\circ)$. We will show that $\log(1 + \xi_n) = \xi_n + O(n^{-2})$. To this end, we use the fact that for every $z > -1$ it holds $-z \leq -\log(1+z) \leq -z + \frac{z^2}{2}$ (this follows e.g. from a Taylor expansion of $\log(1+z)$ around $z = 0$) and show that $\xi_n^2 = O(n^{-2})$:

$$\xi_n^2 \leq \frac{1}{4n^2}\|I(\hat{\mu}_n^\circ)\|^2 \|x_{n+1} - \hat{\mu}_n^\circ\|^4 \leq \frac{1}{4n^2}\left(\sup_{\mu \in \Theta_0} \|I(\mu)\|\right)^2 \left(2n_0 B\right)^4 = O(n^{-2}),$$

for all $n \geq m$. Thus we proved $\log(1 + \xi_n) = \xi_n + O(n^{-2})$. Moreover, $\log(1 + \frac{k}{2n}) = \frac{k}{2n} + O(n^{-2})$, so

$$\Delta(n) = \frac{k}{2n} - \frac{1}{2n}(x_{n+1} - \hat{\mu}_n^\circ)^T I(\hat{\mu}_n^\circ)(x_{n+1} - \hat{\mu}_n^\circ) + \sum_{i=-n_0+1}^{n+1} \log \frac{P_{\hat{\mu}_{n+1}^\circ}(x_i)}{P_{\hat{\mu}_n^\circ}(x_i)} + O(n^{-2}). \tag{7}$$

To bound the sum, we note that it equals $(n + n_0 + 1)D(\hat{\mu}_{n+1}^\circ \| \hat{\mu}_n^\circ)$ where we used Proposition 1 with the empirical distribution again. Then, using Proposition 2, we get:

$$\sum_{i=-n_0+1}^{n+1} \log \frac{P_{\hat{\mu}_{n+1}^\circ}(x_i)}{P_{\hat{\mu}_n^\circ}(x_i)} = \frac{n+n_0+1}{2}(\hat{\mu}_n^\circ - \hat{\mu}_{n+1}^\circ)^T I(\hat{\mu}_n^\circ)(\hat{\mu}_n^\circ - \hat{\mu}_{n+1}^\circ) + (n + n_0 + 1)O(\|\hat{\mu}_n^\circ - \hat{\mu}_{n+1}^\circ\|^3).$$

Since $\hat{\mu}_n^\circ - \hat{\mu}_{n+1}^\circ = (\hat{\mu}_n^\circ - x_{n+1})/(n + n_0 + 1)$, and $\|\hat{\mu}_n^\circ - x_{n+1}\| \leq 2B$, it follows that:

$$\sum_{i=-n_0+1}^{n+1} \log(P_{\hat{\mu}_{n+1}^\circ}(x_i)/P_{\hat{\mu}_n^\circ}(x_i)) = \frac{1}{2n}(x_{n+1} - \hat{\mu}_n^\circ)^T I(\hat{\mu}_n^\circ)(x_{n+1} - \hat{\mu}_n^\circ) + O(n^{-2}).$$

Putting this into (7) gives: $\Delta(n) = \frac{k}{2n} + O(n^{-2})$, as claimed.

The constant in $O(1)$ does not depend on the sequence x^n, because for $n < m$, $\hat{\mu}_n^\circ$ (as a convex combination of x_0 and $\hat{\mu}_n$) is kept away from the boundary of Θ_{mean} and thus $I(\hat{\mu}_n^\circ)$ is bounded from above by a constant independent of the sequence x^n. ■

4 Example: the Bernoulli model

The Bernoulli model is $\{P_\mu \mid \mu \in [0,1]\}$, where $\mathcal{X} = \{0,1\}$ and $P_\mu(x) = \mu^x(1-\mu)^{1-x}$. The Fisher information is $I(\mu) = E_{P_\mu}[(\frac{\mathrm{d}}{\mathrm{d}\mu} \log P_\mu(X))^2] = 1/(\mu(1-\mu))$. After observing x^n, the likelihood is maximized by $\hat{\mu} = o/n$ where $o = x_1 + \ldots + x_n$; we will also use $z = n - o$. It turns out not to be necessary to introduce any fake outcomes in this case (i.e. $n_0 \to 0$). Thus, $\hat{\mu}_n = \hat{\mu}_n^\circ$, and the flattened ML prediction is

$$P_{\text{FML}}(1 \mid x^n) = \hat{\mu}_n \left(\frac{n + \frac{1}{2} I(\hat{\mu}_n)(1-\hat{\mu}_n)^2}{n + \frac{1}{2}} \right) = \frac{n\hat{\mu}_n + \frac{1}{2}(1-\hat{\mu}_n)}{n + \frac{1}{2}} = \frac{o + \frac{z}{2n}}{n + \frac{1}{2}}.$$

The regret for this estimator is maximized for the all-zero or all-one sequence; an easy calculation shows it to be $-\frac{1}{2}\log(16/\pi) + \log(\Gamma(n+\frac{1}{2})/\Gamma(n)) = \frac{1}{2}\log n + O(1)$. Thus, even though the worst case is achieved for non-ineccsi sequences for which technically Theorem 11 does not apply, we find that the flattened ML prediction strategy achieves asymptotically optimal worst-case regret.

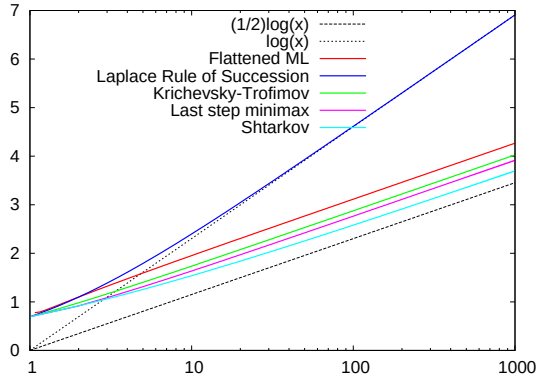


Figure 1: Worst-case regret of Bernoulli estimators.

In Figure 1, we plot this worst-case regret together with the worst-case regret for a number of other estimators: (1) the traditional Laplace estimator $P(1 \mid x^n) = (o+1)/(n+2)$, which is equal to the Bayes predictive distribution using a uniform prior on μ and which does not behave very well on non-ineccsi sequences, (2) the Krichevsky-Trofimov estimator $P(1 \mid x^n) = (o + \frac{1}{2})/(n+1)$ (Krichevsky & Trofimov, 1981), which is equal to the Bayes predictive distribution using Jeffreys' prior, and (3) the "Last Step Minimax" estimator (Takimoto & Warmuth, 2000), also known as "Conditional NML" estimator (Rissanen & Roos, 2007). The regret for this last estimator was shown to be at most $\frac{1}{2}\log(n+1) + \frac{1}{2}$ in (Takimoto & Warmuth, 2000). As baselines, we plot the functions $\frac{1}{2}\log n$ and $\log n$, as well as the regret under the Shtarkov (or NML) distribution. As mentioned in the introduction, the NML distribution is defined only with respect to a known horizon; here the horizon is increased with the sample size, so the Shtarkov results do not reflect a valid prediction strategy but rather provide a tight lower bound on the worst-case regret.

The figure shows that the flattened ML model shows performance comparable to the KT and last step minimax estimators, although the constant term is slightly higher.

5 Almost Sure Convergence in the Stochastic Case

In Section 3, we showed that the flattened ML strategy achieves optimal regret under a mild assumption that the outcomes are bounded and form a Θ_0-sequence. For those cases where this condition is not satisfied, the boundedness requirement can be replaced with the assumption that the data are generated i.i.d. from some distribution of which the first four moments exist. We can then obtain the result (6) with probability one.

The idea of the proof is that when the outcomes are generated i.i.d., they are in some sense bounded with high probability anyway. Specifically, if we allow the bound to increase with n, and if the rate of increase is *faster* than $n^{1/4}$, one can show that when the first four moments of the distribution exist, the outcomes are actually "bounded" (i.e., for all n, the outcome X_n is bounded by the bound for sample size n) with probability one; this is the content of Lemma 12. In Theorem 14, we show that if the bound increases *more slowly* than $n^{1/3}$, most of the analysis done in the proof of Theorem 11 still works. Combining those two facts gives (6) with probability one.

Lemma 12 *Let $X_1, X_2, \ldots$ be i.i.d. random variables and suppose the first m moments of X_n exist. Then, for every $B > 0, \alpha > 0$, with probability 1 there exists an n' such that $\|X_n\| \leq B n^{\frac{1}{m-\alpha}}$ for all $n \geq n'$.*

Proof: Equivalently, we prove that almost surely, the event $A_n = \{\|X_n\| > B n^{\frac{1}{m-\alpha}}\}$ occurs only finitely often. From the Borel-Cantelli lemma we know that this is the case when $\sum_{n=1}^{\infty} P(A_n) < \infty$; we have

$$P(A_n) = P\left(\|X_n\| \geq B n^{\frac{1}{m-\alpha}}\right) = P\left(\|X_n\|^m \geq B^m n^{\frac{m}{m-\alpha}}\right) \leq E\|X_n\|^m B^{-m} n^{-\frac{m}{m-\alpha}},$$

where the last step follows from Markov's inequality. Since $\frac{m}{m-\alpha} > 1$, the sum converges. ■

The next lemma is purely technical and will be needed in further proofs:

Lemma 13 *Let $a_1, a_2, \ldots$ be a positive infinite sequence and let $b_1, b_2, \ldots$ be a positive nonincreasing infinite sequence. Define $A_n = a_1 + \ldots + a_n$, $B_n = b_1 + \ldots + b_n$ and $C_n = a_1 b_1 + \ldots + a_n b_n$. If $A_n = O(n)$ and $B_n = O(1)$, then $C_n = O(1)$.*

Proof: By assumption there are c, n_0 such that for every $n \geq n_0$ we have $A_n \leq c \cdot n$. Fix some $n \geq n_0$ and A_n. Now suppose there is an $a_{n'} > c$ for some $n_0 \leq n' \leq n$. Since $A_{n'} \leq c \cdot n$, there must be an earlier term $a_{n''} < c$ for some $1 \leq n'' < n'$. By increasing $a_{n''}$ to c and decreasing $a_{n'}$ by the same amount, A_n is unchanged while the value of C_n cannot decrease. Thus we may assume w.l.o.g. that $a_i \leq c$ for all $n_0 \leq i \leq n$. But in that case we have $C_n = C_{n_0-1} + \sum_{i=n_0}^{n} a_i b_i \leq C_{n_0-1} + c \cdot \sum_{i=n_0}^{n} b_i = O(1)$. ■

Theorem 14 *Let $X_1, X_2, \ldots$ be i.i.d. generated by a probability distribution P^* of which the first four moments exist and such that $E[X] \in \Theta_{\text{mean}}$. Let $\mathcal{M}$ be a k-dimensional exponential family with mean-value parameter space Θ_{mean}. Then the flattened ML strategy P_{FML} almost surely achieves asymptotically optimal regret, i.e.*

$$\mathcal{R}(P_{\text{FML}}, x^n) = \frac{k}{2} \log n + O(1) \tag{8}$$

holds with probability one.

Proof: Since the first four moments of P^* exists, Lemma 12 states that for large n, the sequence of outcomes $x_1, x_2, \ldots$ is bounded by Bn^q for every $q > 1/4$ with probability one. For simplicity, we take $q = 0.3$, but any $q \in (1/4, 1/3)$ would work. From the strong law of large numbers, we know that the smoothed ML estimator $\hat{\mu}_n^\circ$ converges with probability one. Therefore, for large n, $\hat{\mu}_n^\circ$ is bounded, $\|\hat{\mu}_n^\circ\| \leq C$.

We only give the sketch of the proof, because it closely follows the proof of Theorem 11. The main difference is that in Theorem 11, we had $\|x_n\| \leq B$, while here we have (with probability one) $\|x_n\| \leq Bn^{0.3}$ for large n. A closer look at the proof of Theorem 11 shows that after weakening the bound on $\|x_n\|$ we still get the same rates. The only problem is that now we are not able to prove, that $\Delta(n) = \frac{k}{2n} + O(n^{-2})$. However, to obtain (8), it is enough to show that $\Delta(n) = \frac{k}{2n} + f(n)$, where $f(n)$ is a function such that $\sum_n f(n)$ converges and thus is $O(1)$. To this end, instead of directly bounding ξ_n^2, we will show that $\sum_n \xi_n^2$ converges. Since for large n,

$$\xi_n^2 \leq \frac{1}{4n^2} \sup_{\|\mu\| \leq C} \|I(\mu)\|^2 (\|x_{n+1}\| + C)^4 = C' \frac{\|x_{n+1}\|^4 + O(n^{0.9})}{n^2}$$

for some constant C', we only need to show that the sum $\sum_n \frac{\|x_{n+1}\|^4}{n^2}$ converges. But this follows from Lemma 13 with $a_i = \|x_{i+1}\|^4$ and $b_i = i^{-2}$: we have $A_n = \sum_{i=1}^{n} \|x_{i+1}\|^4 = O(n)$ because A_n/n converges with probability one from the strong law of large numbers (because the fourth moment of P^* exists), and $B_n = \sum_{i=1}^{n} i^{-2} = O(1)$. This means that with probability one, $\sum_n \xi_n^2$ converges. ■

6 Application: Model Selection

The strange behavior of the ML plug-in code first became apparent in a simulation study, where it was found that this code gives rise to much weaker model selection performance than other model selection criteria, such as Bayes factors model selection or even naive maximum likelihood model selection (De Rooij & Grünwald, 2005; De Rooij & Grünwald, 2006); this is especially disturbing since the plug-in based version of MDL has often been advocated for practical use (Rissanen, 1986; Rissanen, 1989; Grünwald, 2007). As mentioned in the introduction, while the expected regret for the ML plug-in estimator is $\frac{k}{2} \log n + O(1)$ when the model contains the data generating distribution, it behaves differently when it does not. This is quite undesirable for model selection: if it is certain that the true distribution is in all considered models, then there is no need to do model selection in the first place!

Since the flattened ML prequential code described in this paper does not suffer from anomalous redundancy under misspecification, we may reasonably hope for better model selection performance. Therefore we have come full circle by turning back to our original (2005) model selection experiments, in order to determine to what extent the flattened ML prequential plug-in code avoids the shortcomings of the unflattened version, and whether or not it yields a useful model selection criterion.

The experimental setup is the same as it was in (De Rooij & Grünwald, 2006), but we provide a brief description here as well to make this paper self-contained. The experiments involve a number of model selection criteria: one based on the flattened ML plug-in code and a number of others, which will be used as a basis for comparison. After defining these model selection criteria, we show the results of the simulation and discuss how the performance of the criterion based on the flattened ML plug-in estimator relates to the results we reported earlier.

6.1 Experiments

All experiments are based on repeatedly sampling a number of outcomes from either the Poisson model $\mathcal{M}_{\text{P}} = \{P_{\text{P}}(X;\mu) \mid \mu \in (0,\infty)\}$ where $P_{\text{P}}(x;\mu) = e^{\mu}\mu^x/x!$ or the geometric model $\mathcal{M}_{\text{G}} = \{P_{\text{G}}(X;\mu) \mid \mu \in (0,\infty)\}$ where $P_{\text{G}}(x;\mu) = \mu^x(\mu+1)^{-(x+1)}$. To make the models easier to compare, both are parameterized by the mean, which is standard for Poisson but not for the geometric model. We first define a number of criteria to select between these models. All criteria can be described in terms of a function L that maps a model $\mathcal{M}$ and a sequence of outcomes x^n to a codelength (negative loglikelihood, accumulated prediction error). Subsequently define the level of evidence in favor of the Poisson model as $\Delta(x^n) = L(\mathcal{M}_{\text{G}}, x^n) - L(\mathcal{M}_{\text{P}}, x^n)$. We select Poisson if $\Delta(x^n) > 0$ and geometric otherwise.

Many common model selection criteria can be defined in terms of a function L. Our experiments involve the following model selection criteria:

The *Known mean* criterion is defined by $L(\mathcal{M}, x^n) = -\log P(x^n)$; here $P \in \mathcal{M}$ is the distribution that satisfies $E_P[X] = \mu$, where μ is the true mean of the data. Although the true mean is not known in practice, this criterion is useful as an ideal baseline. It has the properties that (1) one of the two hypotheses equals the generating distribution and (2) the sample consists of outcomes which are i.i.d. according to this distribution. In (Cover & Thomas, 1991), Sanov's Theorem is used to show that in such a situation, the probability that the criterion prefers the wrong model ("error probability") decreases exponentially in the sample size. If the data are generated using Poisson$[\mu]$ then the error probability decreases exponentially in the sample size, with some error exponent; if the data are generated with Geometric$[\mu]$ then the overall probability is exponentially decreasing with the same exponent (Cover & Thomas, 1991, Theorem 12.9.1 on page 312 and text thereafter). Thus, when the error probability is plotted on a log scale, the slope should be equal whether the generating distribution is Poisson or geometric. This can be observed to be the case in Figures 2a and 2b.

The *Maximum Likelihood (ML)* criterion is defined by $L(\mathcal{M}, x^n) = -\log \sup_{P\in\mathcal{M}} P(x^n)$. This is the same as a (generalized) likelihood ratio test (GLRT) with a threshold of one. The ML criterion is well known to be prone to overfitting: in a complex model, there may be a distribution that provides good fit to the data purely by chance. Two approaches to penalize complex models are known as AIC (Akaike, 1974) and BIC (Schwarz, 1978). However, for both these methods the penalty term depends only on the number of parameters in the models. In this case, both models have only a single parameter, so in $\Delta(x^n)$ the penalty terms cancel: in this case, both AIC and BIC are equivalent to a GLRT with zero threshold!

Bayes factor model selection is obtained if we set $L(\mathcal{M}, x^n) = -\log \int_\mu P_\mu(x^n)\pi(\mu)\,\mathrm{d}\mu$, where the prior π may depend on the model. In this case, $\Delta(x^n)$ is equal to the logarithm of the Bayes factor. We use Jeffreys' prior in our experiments. Because it is improper for the Poisson and geometric models, we use the first observation to normalize the prior. Letting $S = \sum_{i=1}^n x_i$, We obtain the following expressions:

$$\pi_{\text{P}}(\mu|x_1) = \frac{e^{-\mu}\mu^{x_1-\frac{1}{2}}}{\Gamma(\frac{1}{2}+x_1)}; \qquad \pi_{\text{G}}(\mu|x_1) = (x_1+\tfrac{1}{2})\frac{\mu^{x_1-\frac{1}{2}}}{(\mu+1)^{x_1+\frac{3}{2}}};$$

$$L(\mathcal{M}_{\text{P}}, x^n) = -\log\int_0^\infty P_{\text{P}}(x_2^n;\mu)\pi_{\text{P}}(\mu|x_1)\,\mathrm{d}\mu = \log\frac{\Gamma(x_1+\frac{1}{2})}{\Gamma(S+\frac{1}{2})} + (S+\tfrac{1}{2})\log n + \sum_{i=2}^n \log(x_i!);$$

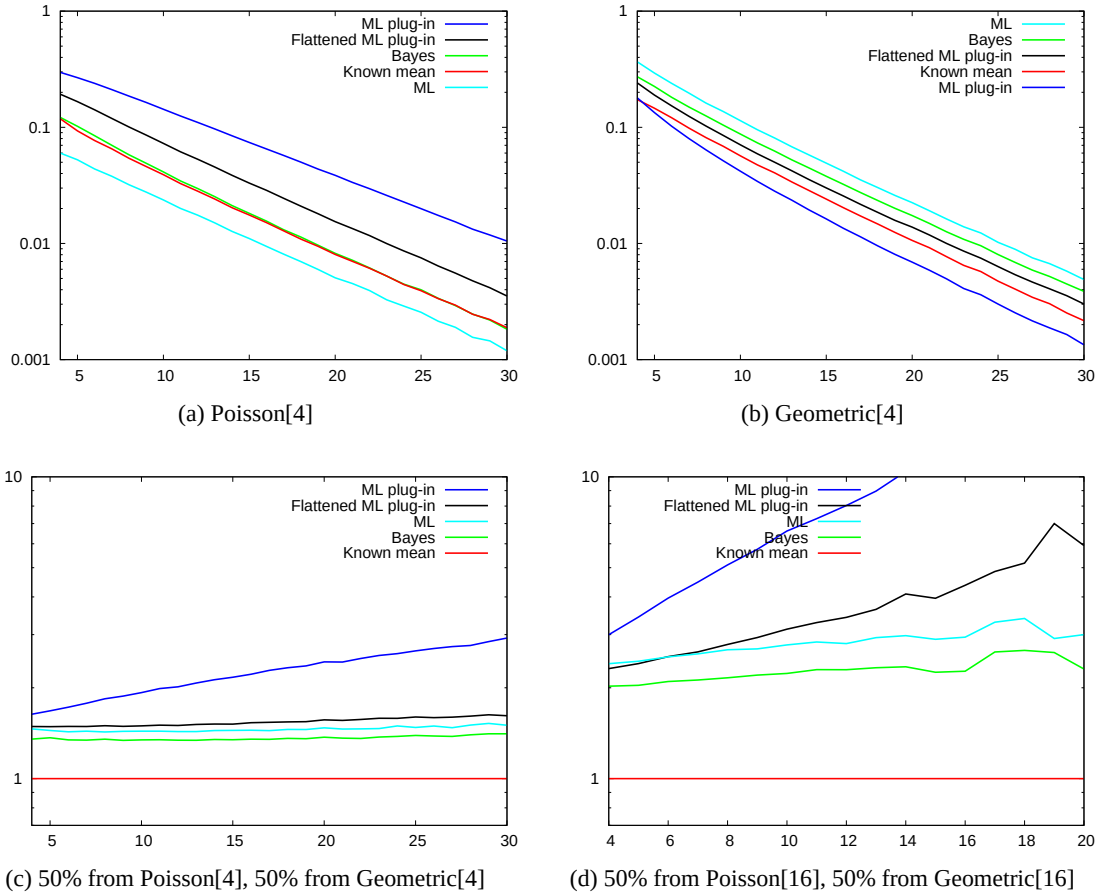
$$L(\mathcal{M}_{\text{G}}, x^n) = -\log\int_0^\infty P_{\text{G}}(x_2^n;\mu)\pi_{\text{G}}(\mu|x_1)\,\mathrm{d}\mu = -\log(x_1+\tfrac{1}{2}) + \log\frac{\Gamma(S+n+\frac{1}{2})}{\Gamma(n)\Gamma(S+\frac{1}{2})}.$$

The *ML plug-in* criterion is defined by setting $L(\mathcal{M}, x^n) = -\log P_{\text{ML}}(x^n)$ where P_{ML} is as in Definition 3. This codelength does not correspond to a Bayesian marginal likelihood, so this criterion does not yield Bayes factor model selection; however P_{ML} is a valid universal code so it does lead to an MDL model selection procedure.

The *flattened ML plug-in* criterion is defined by setting $L(\mathcal{M}, x^n) = -\log P_{\text{FML}}(x^n)$, where U is as in Definition 9.

These five criteria are subjected to two different kinds of tests:

Error probability The error probability for a criterion is the probability that it will select a model that does not contain the distribution from which the data are sampled. We estimate the error probability through repeated sampling: in our experiments, samples are always drawn from a Poisson$[\mu]$ distribution with probability p, or from a Geom$[\mu]$ distribution with probability $1-p$. Figure 2 shows the error probability as a function of the sample size on a log scale, for various values of p and μ. After the first two graphs, we plot the ratio of the error probability of a criterion with the error probability of the baseline "known mean" criterion: this allows for better distinction between the criteria.



(a) Poisson[4]

(b) Geometric[4]

(c) 50% from Poisson[4], 50% from Geometric[4]

(d) 50% from Poisson[16], 50% from Geometric[16]

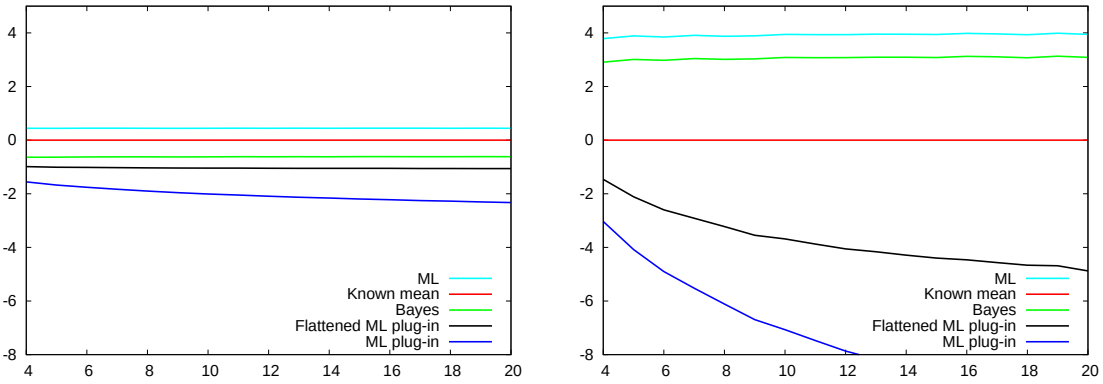
Figure 2: Error probability. For Figures (c) and (d), the error frequency is divided by the baseline, the error frequency of the Known mean criterion. Estimated using 10^6 trials.

Bias Let $\Delta_\mu(x^n)$ be the evidence in favor of the Poisson model according to the known mean criterion. For other criteria C, the quantity Δ_C can be interpreted as an *estimator* for Δ_μ. The bias of such an estimator is $E[\Delta_C(X^n) - \Delta_\mu(X^n)]$, where the expectation is taken under the true distribution. We subsequently estimate this bias for all criteria by calculating the average over many trials. The results are in Figure 3.

6.2 Discussion

In order to establish a context to discuss the behavior of the flattened ML plug-in criterion, we first briefly summarize the conclusions from (De Rooij & Grünwald, 2006), which still apply to the current experiments.

- ML and ML plug-in exhibited worst performance; the Bayesian criterion performed reasonably on all tests.
- We found that the ML criterion consistently displays the largest bias in favor of Poisson. Figure 3 shows how on average, for ML we obtained at least 0.4 nats more evidence in favor of the Poisson model than for known mean. The Poisson model appears to have a greater descriptive power, even though the two models have the same number of parameters: intuitively, the Poisson model allows more information about the data to be stored in the parameter estimate.
- In all graphs in Figure 2 one can observe the unusual slope of the error rate line of the ML plug-in criterion, which clearly favors the geometric distribution. This is very undesirable for model selection, because the error rate when data are sampled from Poisson with probability p and from geometric with probability $1-p$, is dominated by the worst of the two cases, i.e. the case that the data are Poisson distributed. This explains why the error rate is so poor in the case where $p = 0.5$ (Figures 2c and 2d). The bias is visible more explicitly in Figure 3, where ML plug-in can be observed to become more and more favorable to the geometric model as the sample size increases, regardless of whether the data were sampled from a Poisson or geometric distribution.



(a) Poisson[8] (b) Geometric[8]

Figure 3: The classification bias in favor of the Poisson model in nats, estimated using 10^5 trials.

The new experiments also include results for the new flattened ML plug-in criterion. Figure 2 shows that, compared to ML plug-in, the slope of the error probability line for the flattened ML plug-in estimator is much closer to that of known mean. Nevertheless, when the mean is increased in subfigure (d), we see that the error probability seems to go down at a somewhat slower rate than it does for the Bayes and ML criteria.

In Figure 3 we find that, like ML plug-in, flattened ML plug-in is biased in favor of the geometric model. If the data are geometric, then this bias increases with sample size, as it does for ML plug-in, albeit at a slower rate. However, for Poisson data most of this effect appears to have been suppressed. This means that the probability that Poisson data are incorrectly judged to be geometric never becomes much larger than for other criteria, regardless of sample size. So for model selection purposes, the bias is acceptable.

In conclusion, the flattened ML plug-in criterion does indeed seem to provide a substantial improvement in model selection performance over the ML plug-in criterion. That said, the bias in favor of the geometric model has not completely vanished, which may be because of the $O(1)$ terms in the redundancy of the estimator which we did not analyze. The Bayesian criterion is clearly somewhat more reliable, but may be too computationally intensive depending on the considered models.

7 Conclusion

Given a model (set of probability distributions) $\mathcal{M}$, the maximum likelihood estimator $\hat{\theta}(x^n)$ based on past observations $x^n = x_1, \ldots, x_n$ indexes a distribution that is a natural and easy to compute candidate for prediction of the next observation. However, previous work shows that if the data generating distribution P^* is not in the model, then such a "ML plug-in" prediction strategy yields suboptimal expected regret: unlike for other prediction strategies, such as Bayesian prediction, the expected regret is *not* $(k/2)\log n + O(1)$, where k is the number of parameters in the model. This is a serious problem when the "ML plug-in" strategy is used for model selection: there, by its very nature, the possibility that $P^* \notin \mathcal{M}$ deserves serious consideration.

To address this issue, we described a simple "flattening" of the ML distribution and related predictors, using which the optimal worst case *individual sequence* regret of $(k/2)\log n + O(1)$ can be achieved, for exponential family models and bounded outcome spaces (Theorem 11 on page 6). For unbounded spaces, we provided an almost-sure result (Theorem 14 on page 9). In Section 6, we subjected the new prediction strategy to the same model selection experiments that showed the ML plug-in strategy to be suboptimal, obtaining a major improvement in performance.

References

Akaike, H. (1974). A new look at statistical model identification. *IEEE Trans. Autom. Contr., 19*, 716–723.

Azoury, K., & Warmuth, M. (2001). Relative loss bounds for on-line density estimation with the exponential family of distributions. *J. of Machine Learning, 43*, 211–246. Special issue on Theoretical Advances in On-Line Learning, Game Theory and Boosting, edited by Y. Singer.

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory*. Chichester, UK: Wiley.

Barron, A., Rissanen, J., & Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Trans. Inf. Th., 44*, 2743–2760. Special Commemorative Issue: Inf. Theory: 1948-1998.

Cesa-Bianchi, N., & Lugosi, G. (2001). Worst-case bounds for the logarithmic loss of predictors. *J. of Machine Learning, 43*, 247–264.

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning and games*. Cambridge University Press.

Corcuera, J., & Giummolè, F. (1999). A generalized Bayes rule for prediction. *Scandinavian J. of Statistics, 26*, 265–279.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. John Wiley.

Dawid, A. (1984). Present position and potential developments: Some personal views, statistical theory, the prequential approach. *J. Royal Stat.Soc., Ser. A, 147*, 278–292.

De Rooij, S., & Grünwald, P. D. (2005). MDL model selection using the ML plug-in code. *Proc. 2005 International Symposium on Inf. Th. (ISIT 2005)*. Adelaide, Australia.

De Rooij, S., & Grünwald, P. D. (2006). An empirical study of MDL model selection with infinite parametric complexity. *J. of Mathematical Psychology, 50*, 180–192.

Freund, Y. (1996). Predicting a binary sequence almost as well as the optimal biased coin. *Proc. 9th Conf. on Computational Learning Theory (COLT 1996)* (pp. 89–98).

Grünwald, P. (2007). *The minimum description length principle*. Cambridge, MA: MIT Press.

Grünwald, P., & Kotłowski, W. (2010). Prequential plug-in codes that achieve optimal redundancy rates even if the model is wrong. arXiv:1002.0757.

Grünwald, P. D. (2005). MDL tutorial. *Advances in Minimum Description Length: Theory and Applications*. MIT Press.

Grünwald, P. D., & de Rooij, S. (2005). Asymptotic log-loss of prequential maximum likelihood codes. *Proc. 18th Conf. on Computational Learning Theory (COLT 2005)* (pp. 652–667).

Hemerly, E., & Davis, M. (1989). Strong consistency of the PLS criterion for order determination of autoregressive processes. *The Annals of Statistics, 17*, 941–946.

Hutter, M., & Poland, J. (2005). Adaptive online prediction by following the perturbed leader. *J. of Machine Learning Research, 6*, 639–660.

Kalai, A., & Vempala, S. (2003). Efficient algorithms for online decision. *Proc. 16th Conf. on Computational Learning Theory (COLT 2003)* (pp. 506–521). Berlin: Springer.

Krichevsky, R., & Trofimov, V. (1981). The performance of universal encoding. *IEEE Trans. Inf. Th., IT-27*, 199–207.

Li, L., & Yu, B. (2000). Iterated logarithmic expansions of the pathwise code lengths for exponential families. *IEEE Trans. Inf. Th., 46*, 2683–2689.

Rissanen, J. (1984). Universal coding, information, prediction and estimation. *IEEE Trans. Inf. Th., 30*, 629–636.

Rissanen, J. (1986). A predictive least squares principle. *IMA J. of Math. Contr. and Inf., 3*, 211–222.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. World Scientific Publishing Company.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Trans. Inf. Th., IT-42*, 40–47.

Rissanen, J., & Roos, T. (2007). Conditional NML universal models. *Proc. Inf. Th. and Applications Workshop (ITA-07)* (pp. 337–341). IEEE Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*.

Shtarkov, Y. (1987). Universal sequential coding of single messages. *Problems of Inf. Trans., 23*, 175–186.

Takimoto, E., & Warmuth, M. (2000). The last-step minimax algorithm. *Proc. 11th Conf. on Algorithmic Learning Theory (ALT 2000)*.

Vidoni, P. (2008). Improved predictive model selection. *J. of Stat. Planning and Inference, 138*, 3713–3721.

Wei, C. (1990). On predictive least squares principles. *The Annals of Statistics, 20*, 1–42.