Minimum Description Length Model Selection

de Rooij, S.

Publication date
2008

Chapter 3

Behaviour of Prequential Codes under Misspecification

Universal coding lies at the basis Rissanen's theory of MDL (minimum description length) learning [4, 39] and Dawid's theory of prequential model assessment [26]. It also underlies on-line prediction algorithms for data compression and gambling purposes. In the introductory chapter (Section 1.2.3), we defined universality of a code in terms of its worst-case regret. Roughly, a code is universal with respect to a model $\mathcal{M}$ if it achieves small worst-case regret: it allows one to encode data using not many more bits than the optimal code in $\mathcal{M}$. We also described the four main universal codes: the *Shtarkov* or *NML* code, the *Bayesian mixture* code, the *2-part MDL code* and the *prequential maximum likelihood code* (PML), also known as the "ML plug-in code" or the "predictive MDL code" [4, 38]. This code was introduced independently by Rissanen [69] and by Dawid [26], who proposed it as a forecasting strategy rather than as a code. In this chapter we study the behaviour of the PML code if the considered model $\mathcal{M}$ does not contain the data-generating distribution P^*. We require that the model is a 1-parameter exponential family, but our results can possibly be extended to models with more parameters.

Instead of the worst-case regret, we analyse the *redundancy*, a closely related concept. We find that the redundancy of PML can be quite different from that of the other main universal codes. For all these codes, the redundancy is $\frac{1}{2}c \ln n + O(1)$ for some c, but while $c = 1$ for Bayes, NML and 2-part codes (under regularity conditions on P^* and $\mathcal{M}$), we show here that for PML, any $c > 0$ is possible, depending on P^* and $\mathcal{M}$.

There are a plethora of results concerning the redundancy and/or the regret for PML, for a large variety of models including multivariate exponential families, ARMA processes, regression models and so on. Examples are [70, 44, 97, 57]. In all these papers it is shown that either the regret or the redundancy grows as $\frac{k}{2} \ln n + o(\ln n)$, either in expectation or almost surely. Thus, these results already indicate that $c = 1$ for those models. The reason that these results do

not contradict ours, is that they invariably concern the case where the generating P^* is in $\mathcal{M}$, so that automatically $\mathrm{var}_{M^*}(X) = \mathrm{var}_{P^*}(X)$.

As discussed in Section 3.4, the result has interesting consequences for parameter estimation and practical data compression, but the most important and surprising consequence is for MDL learning and model selection, where our result implies that PML may behave suboptimally *even if one of the models under consideration is correct!*

In Section 3.1 we informally state and explain our result. Section 3.2 contains the formal statement of our main result (Theorem 3.2.3), as well as a proof. In Section 3.3 we show that our results remain valid to some extent if "redundancy" is replaced by "expected regret" (Theorem 3.3.1). We discuss further issues, including relevance of the result, in Section 3.4. Section 3.5 states and proves various lemmas needed in the proofs of Theorems 3.2.3 and 3.3.1.

3.1 Main Result, Informally

Suppose $\mathcal{M} = \{P_\theta \ : \ \theta \in \Theta\}$ is a k-dimensional parametric family of distributions, and $Z_1, Z_2, \ldots$ are i.i.d. according to some distribution $P^* \in \mathcal{M}$. A code is universal for $\mathcal{M}$ if it is almost as efficient at coding outcomes from P^* as the best element of $\mathcal{M}$. (As in previous chapters, we sometimes use codes and distributions interchangeably.) In the introductory chapter we measured the overhead incurred on the first n outcomes $z^n = z_1, \ldots, z_n$ in terms of the worst-case regret

$$\max_{z^n} \mathcal{R}(L, \mathcal{M}, z^n) \quad = \quad \max_{z^n} \left(L(z^n) - \inf_{L' \in \mathcal{M}} L'(z^n) \right),$$

but in this chapter we consider the redundancy instead. We define the *redundancy* of a distribution Q with respect to a model $\mathcal{M}$ as

$$\mathfrak{R}(P^*, Q, \mathcal{M}, n) \quad := \quad \underset{Z^n \sim P^*}{E}[-\ln Q(Z^n)] - \inf_{\theta \in \Theta} \underset{Z^n \sim P^*}{E}[-\ln P_\theta(Z^n)], \qquad (3.1)$$

where we use nats rather than bits as units of information to simplify equations. We omit the first three arguments of the redundancy if they are clear from context. These and other notational conventions are detailed in Section 3.2. The redundancy is a close lower bound on the *expected* regret, see Section 3.3. We do not know the exact relationship to worst-case regret.

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

$$\mathfrak{R}(P^*, Q, \mathcal{M}, n) = \frac{k}{2} \ln n + O(1), \qquad (3.2)$$

where the $O(1)$ may depend on P^*, $\mathcal{M}$ and the universal code Q that is used. This is the famous "k over 2 log n formula", refinements of which lie at the basis of most practical approximations to MDL learning, see [38].

Often, the source distribution P^* is assumed to be an element of the model. If such is the case, then by the information inequality [25] the second term of (3.1) is minimised for $P_\theta = P^*$, so that

$$\mathfrak{R}(P^*, Q, \mathcal{M}, n) = E_{P^*}[-\ln Q(Z^n)] - E_{P^*}[-\ln P^*(Z^n)]. \tag{3.3}$$

Thus, (3.3) can be interpreted as the expected number of additional nats one needs to encode n outcomes if one uses the code corresponding to Q instead of the optimal (Shannon-Fano) code with lengths $-\ln P^*(Z^n)$. For a good universal code this quantity is small for all or most $P^* \in \mathcal{M}$.

In this chapter we consider the case where the data are i.i.d. according to an *arbitrary* P^* not necessarily in the model $\mathcal{M}$. It is now appropriate to rename the redundancy to *relative* redundancy, since we measure the number of nats we lose compared to the best element in the model, rather than compared to the generating distribution P^*. The definition (3.1) remains unchanged. It can no longer be rewritten as (3.3) however: Assuming it exists and is unique, let P_{θ^*} be the element of $\mathcal{M}$ that minimises KL divergence to P^*:

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

where the equality follows from the definition of the KL divergence [25]. Then the relative redundancy satisfies

$$\mathfrak{R}(P^*, Q, \mathcal{M}, n) = E_{P^*}[-\ln Q(Z^n)] - E_{P^*}[-\ln P_{\theta^*}(Z^n)]. \tag{3.4}$$

It turns out that for the NML, 2-part MDL and Bayes codes, the relative redundancy (3.4) with $P^* \notin \mathcal{M}$, still satisfies (3.2), under some conditions on $\mathcal{M}$ and P^* (Section 3.3). In this chapter we show that (3.2) does *not* hold for PML. The PML code with length function L works by sequentially predicting Z_{i+1} using a (slightly modified) ML or Bayesian MAP estimator $\hat\theta_i = \hat\theta(z^i)$ based on the past data, that is, the first i outcomes z^i. The total code length $L(z^n)$ on a sequence z^n is given by the sum of the individual "predictive" code lengths (log losses): $L(z^n) = \sum_{i=0}^{n-1}[-\ln P_{\hat\theta_i}(z_{i+1})]$. In our main theorem, we show that if and $\mathcal{M}$ is a regular one-parameter exponential family ($k = 1$), then

$$\mathfrak{R}(P^*, Q, \mathcal{M}, n) = \tfrac{1}{2}\frac{\mathrm{var}_{P^*}X}{\mathrm{var}_{P_{\theta^*}}X}\ln n + O(1), \tag{3.5}$$

where X is the sufficient statistic of the family. Example 6 below illustrates the phenomenon. Note that if $P^* \in \mathcal{M}$, then $P_{\theta^*} = P^*$ and (3.5) becomes the familiar expression. The result holds as long as $\mathcal{M}$ and P^* satisfy a mild condition that is stated and discussed in the next section. Section 3.4 discusses the consequences of

this result for compression, estimation and model selection, as well as its relation to the large body of earlier results on PML coding.

Example 6. Let $\mathcal{M}$ be the family of Poisson distributions, parameterised by their mean μ. Since neither the NML universal code nor Jeffreys' prior are defined for this model it is attractive to use the PML code as a universal code for this model. The ML estimator $\hat{\mu}_i$ is the empirical mean of $z_1, \ldots, z_i$.

Suppose Z, Z_1, Z_2, ... are i.i.d. according to a degenerate P with $P(Z = 4) = 1$. Since the sample average is a sufficient statistic for the Poisson family, $\hat{\mu}_i$ will be equal to 4 for all $i \geq 1$. On the other hand, μ^*, the parameter (mean) of the distribution in $\mathcal{M}$ closest to P in KL-divergence, will be equal to 4 as well. Thus the relative redundancy (3.4) of the PML code is given by

$$\mathfrak{R}(P^*, Q, \mathcal{M}, n) = -\ln P_{\hat{\mu}_0}(4) + \ln P_4(4) + \sum_{i=1}^{n-1} [-\ln P_4(4) + \ln P_4(4)] = O(1),$$

assuming an appropriate definition of $\hat{\mu}_0$. In the case of the Poisson family, we have $Z = X$ in (3.5). Thus, since $\mathrm{var}_{P^*} Z = 0$, this example agrees with (3.5).

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

Example 7. The Hardy-Weinberg model deals with genotypes of which the alleles are assumed independently Bernoulli distributed according to some parameter p. There are four combinations of alleles, usually denoted "aa", "AA", "aA", "Aa"; but since "aA" and "Aa" result in the same genotype, the Hardy-Weinberg model is defined as a probability distribution on three outcomes. We model this by letting X be a random variable on the underlying space, that maps "aA" and "Aa" to the same value: $X(aa) = 0$, $X(aA) = X(Aa) = \frac{1}{2}$ and $X(AA) = 1$. Then $P(X = 0) = (1-p)^2$, $P(X = \frac{1}{2}) = 2p(1-p)$ and $P(X = 1) = p^2$. The Hardy-Weinberg model is an exponential family with sufficient statistic X. To see this, note that for any parameter $p \in [0, 1]$, we have $EX = \mu = P(A) = p$, so we can parameterise the model by the mean of X. The variance of the distribution with parameter μ is $\frac{1}{2}\mu(1-\mu)$. Now suppose that we code data in a situation where the Hardy-Weinberg model is *wrong* and the genotypes are in fact distributed according to $P(X = \frac{1}{2}) = P(X = 1) = \frac{1}{2}$ and $P(X = 0) = 0$, such that mean and variance of X are $\frac{3}{4}$ and $\frac{2}{32}$ respectively. The closest distribution in the model has the same mean (since the mean is a sufficient statistic), and variance $\frac{3}{32}$. Thus PML will achieve a redundancy of $\frac{1}{3} \ln n$ rather than $\frac{1}{2} \ln n$ (up to $O(1)$).

3.2 Main Result, Formally

In this section, we define our quantities of interest and we state and prove our main result. Throughout this text we use nats rather than bits as units of information. Outcomes are capitalised if they are to be interpreted as random variables instead of instantiated values. Let P^* be a distribution on some set $\mathcal{Z}$, which can be either finite or countably infinite, or a subset of k-dimensional Euclidean space for some $k \geq 1$. A sequence of outcomes $z_1, \ldots, z_n$ is abbreviated to z^n. Let $X : \mathcal{Z} \to \mathbb{R}^k$ be a random vector. We write $E_{P^*}[X]$ as a shorthand for $E_{X \sim P^*}[X]$. When we consider a sequence of n outcomes independently distributed $\sim P^*$, we even use E_{P^*} as a shorthand for the expectation of $(X_1, \ldots, X_n)$ under the n-fold product distribution of P^*. Finally, $P^*(X)$ denotes the probability mass function of P^* in case X is discrete-valued, and the density of P^* in case X takes values in a continuum. When we write "density function of X", then, if X is discrete-valued, this should be read as "probability mass function of X". Note however that in our main result, Theorem 3.2.3 below, we do not assume that the data-generating distribution P^* admits a density.

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

Definition 3.2.1 (Exponential family)**.** The *single parameter exponential family* [48] with *sufficient statistic* X and *carrier* h is the family of distributions with densities $P_\eta(z) := \frac{1}{\mathrm{Z}(\eta)} e^{-\eta X(z)} h(z)$, where $\eta \in \Theta_\eta$. Θ_η is called the *natural parameter space*. The family is called *regular* if Θ_η is an open interval of $\mathbb{R}$.

In the remainder of this text we only consider single parameter, regular exponential families with a 1-to-1 parameterisation; this qualification will henceforth be omitted. Examples include the Poisson, geometric and multinomial families, and the model of all Gaussian distributions with a fixed variance or mean. In the first four cases, we can take X to be the identity, so that $X = Z$ and $\mathcal{X} = \mathcal{Z}$. In the case of the normal family with fixed mean, σ^2 becomes the sufficient statistic and we have $\mathcal{Z} = \mathbb{R}$, $\mathcal{X} = [0, \infty)$ and $X = Z^2$.

The statistic $X(z)$ is sufficient for η [48]. This suggests reparameterising the distribution by the expected value of X, which is called the *mean value parameterisation*. The function $\mu(\eta) = E_{P_\eta}[X]$ maps parameters in the natural parameterisation to the mean value parameterisation. It is a diffeomorphism (it is one-to-one, onto, infinitely often differentiable and has an infinitely often differentiable inverse) [48]. Therefore the mean value parameter space Θ_μ is also an open interval of $\mathbb{R}$. We note that for some models (such as Bernoulli and Poisson), the parameter space is usually given in terms of the a non-open set

of mean-values (e.g., $[0, 1]$ in the Bernoulli case). To make the model a regular exponential family, we have to restrict the set of parameters to its own interior. Henceforth, whenever we refer to a standard statistical model such as Bernoulli or Poisson, we assume that the parameter set has been restricted in this sense.

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

Definition 3.2.2 (PML distribution)**.** Let Θ_μ be the mean value parameter domain of an exponential family $\mathcal{M} = \{P_\mu \mid \mu \in \Theta_\mu\}$. Given $\mathcal{M}$ and constants $x_0 \in \Theta_\mu$ and $n_0 > 0$, we define the *PML distribution* U by setting, for all n, all $z^{n+1} \in \mathcal{Z}^{n+1}$:

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

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

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

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

We henceforth abbreviate $\hat{\mu}(z^n)$ to $\hat{\mu}_n$. We usually refer to the PML distribution in terms of the corresponding code length function

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

To understand this definition, note that for exponential families, for any sequence of data, the ordinary maximum likelihood parameter is given by the average $n^{-1} \sum x_i$ of the observed values of X [48]. Here we define our PML distribution in terms of a slightly modified maximum likelihood estimator that introduces a "fake initial outcome" x_0 with multiplicity n_0 in order to avoid infinite code lengths for the first few outcomes (a well-known problem called by Rissanen the "inherent singularity" of predictive coding [71, 36]) and to ensure that the probability of the first outcome is well-defined for the PML distribution. In practice we can take $n_0 = 1$ but our result holds for any $n_0 > 0$. The justification of our modification to the ML estimator is discussed further in Section 3.4.

Theorem 3.2.3 (Main result)**.** *Let $X, X_1, X_2, \ldots$ be i.i.d. $\sim P^*$, with $E_{P^*}[X] = \mu^*$. Let $\mathcal{M}$ be a single parameter exponential family with sufficient statistic X and μ^* an element of the mean value parameter space. Let U denote the PML distribution with respect to $\mathcal{M}$. If $\mathcal{M}$ and P^* satisfy Condition 3.2.4 below, then*

$$\mathfrak{R}(P^*, U, \mathcal{M}, n) = \tfrac{1}{2} \frac{\mathrm{var}_{P^*} X}{\mathrm{var}_{P_{\mu^*}(X)}} \ln n + O(1). \tag{3.6}$$

Comparing this to (3.5), note that P_{μ^*} is the element of $\mathcal{M}$ achieving smallest expected code length, i.e. it achieves $\inf_{\mu\in\Theta_\mu} D(P^*\|P_\mu)$ [48].

Condition 3.2.4. *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 $\frac{\mathrm{d}^4}{\mathrm{d}\mu^4} D(P_{\mu^*}\|P_\mu) = O\left(\mu^{k-6}\right)$.*
- *If T is bounded from above by a constant g then $\frac{\mathrm{d}^4}{\mathrm{d}\mu^4} D(P_{\mu^*}\|P_\mu)$ is polynomial in $1/(g-\mu)$.*

Roughly, this condition expresses a trade-off between the data generating distribution P^* and the model. If the model is well-behaved, in the sense that the fourth order derivative of the KL divergence does not grow too fast with the parameter, then we do not require many moments of P^* to exist. Vice versa if the model is not well-behaved, then the theorem only holds for very specific P^*, of which many moments exist.

The condition holds for most single-parameter exponential families that are relevant in practice. To illustrate, in Figure 3.2 we give the fourth derivative of the divergence for a number of common exponential families explicitly. All parameters beside the mean are treated as fixed values. Note that to interpret the mean 0 normal distributions as a 1-parameter exponential family the density we had to set $X(z) = z^2$, so that its mean $E[X]$ is actually the variance $E[Z^2]$ of the normal distribution. As can be seen from the figure, for these exponential families, our condition applies whenever at least the first four moments of P^* exist: a quite weak condition on the data generating distribution.

Proof of Theorem 3.2.3. For exponential families, we have

$$\begin{aligned} & E_{P^*}[-\ln P_\mu(Z)] - E_{P^*}[-\ln P_{\mu'}(Z)] \\ = \;& \eta(\mu) E_{P^*}[X(Z)] + \ln \mathrm{Z}(\eta(\mu)) + E_{P^*}[-\ln h(Z)] \\ - \;& \eta(\mu') E_{P^*}[X(Z)] - \ln \mathrm{Z}(\eta(\mu')) - E_{P^*}[-\ln h(Z)] \\ = \;& E_{P^*}[-\ln P_\mu(X)] - E_{P^*}[-\ln P_{\mu'}(X)], \end{aligned}$$

so that $\mathfrak{R}(P^*, U, \mathcal{M}, n) = E_{P^*}[L_U(X^n)] - \inf_\mu E_{P^*}[-\ln P_\mu(X^n)]$. This means that relative redundancy, which is the sole quantity of interest in the proof, depends only on the sufficient statistic X, not on any other aspect of the outcome that may influence Z. Thus, in the proof of Theorem 3.2.3 as well as all the Lemmas and Propositions it makes use of, we will never mention Z again. Whenever we refer to a "distribution" we mean a distribution of random variable X, and we also think of the data generating distribution P^* in terms of the distribution it induces on X rather than Z. Whenever we say "the mean" without further qualification, we refer to the mean of the random variable X. Whenever we

Figure 3.1 $\frac{\mathrm{d}^4}{\mathrm{d}\mu^4} D(P_{\mu^*} \| P_\mu)$ for a number of exponential families.

	$P_{\mu^*}(x)$	$\frac{\mathrm{d}^4}{\mathrm{d}\mu^4} D(P_{\mu^*} \| P_\mu)$
Bernoulli	$(\mu^*)^x (1-\mu^*)^{(1-x)}$	$\frac{6\mu^*}{\mu^4} + \frac{6(1-\mu^*)}{(1-\mu)^4}$
Poisson	$\frac{e^{\mu^*} \mu^{*x}}{x!}$	$\frac{6\mu^*}{\mu^4}$
Geometric	$\theta^x(1-\theta) = \frac{(\mu^*)^x}{(\mu^*+1)^{x+1}}$	$\frac{6\mu^*}{\mu^4} - \frac{6(\mu^*+1)}{(\mu+1)^4}$
Exponential	$\frac{1}{\mu^*} e^{-x/\mu^*}$	$-\frac{6}{\mu^4} + \frac{24\mu^*}{\mu^5}$
Normal (fixed mean = 0)	$\frac{1}{\sqrt{2\pi\mu^* x}} e^{-\frac{x}{2\mu^*}}$	$-\frac{3}{\mu^4} + \frac{12\mu^*}{\mu^5}$
Normal (fixed variance = 1)	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu^*)^2}$	0
Pareto	$\frac{ab^a}{x^{a+1}}$ for $b = \frac{a-1}{a}\mu^*$	$\frac{6a}{\mu^4}$

refer to the Kullback-Leibler (KL) divergence between P and Q, we refer to the KL divergence between the distributions they induce for X. The reader who is confused by this may simply restrict attention to exponential family models for which $Z = X$, and consider X and Z identical.

The proof refers to a number of theorems and lemmas which will be developed in Section 3.5. In the statement of all these results, we assume, as in the statement of Theorem 3.2.3, that $X, X_1, X_2, \ldots$ are i.i.d. $\sim P^*$ and that $\mu^* = E_{P^*}[X]$. If X takes its values in a countable set, then all integrals in the proof should be read as the corresponding sums.

The redundancy can be rewritten further as the sum of the expected risk for each outcome (Lemma 3.5.6). We obtain

$$\mathfrak{R}(P^*, U, \mathcal{M}, n) = \sum_{i=0}^{n-1} \underset{\hat{\mu}_i \sim P^*}{E} \left[D(P_{\mu^*} \| P_{\hat{\mu}_i}) \right]. \tag{3.7}$$

Here, the estimator $\hat{\mu}_i$ is a random variable that takes on values according to P^*, while the optimal parameter value μ^* is fixed (and determined by P^*). We will write $D(\mu^* \| \hat{\mu}_i)$ as shorthand notation for P_{μ^*} and $P_{\hat{\mu}_i}$.

We now first rewrite the divergence. We abbreviate $\delta_i := \hat{\mu}_i - \mu^*$ and $D^{(k)}(\mu) := \frac{d^k}{d\mu^k} D(P_{\mu^*} \| P_\mu)$. That is, $D^{(k)}(\mu)$ is the k-th derivative of the function $f(\mu) := D(P_{\mu^*} \| P_\mu)$. Taylor-expansion of the divergence around μ^* yields

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

The last term is the remainder term of the Taylor expansion, in which $\ddot{\mu}_i \in [\mu^*, \hat{\mu}_i]$. The second term $D^{(1)}(\mu^*)$ is zero, since $D(\mu^* \| \mu)$ reaches its minimum at $\mu = \mu^*$. For the third term we observe that

$$D^{(2)}(\mu) = \frac{\mathrm{d}^2}{\mathrm{d}\mu^2} E[\ln P_{\mu^*}(X) - \ln P_{\mu}(X)] = -\frac{\mathrm{d}^2}{\mathrm{d}\mu^2} E[\ln P_{\mu}(X)],$$

which is equal to the Fisher information. Fisher information is defined as $I(\theta) := E\left[(\frac{\mathrm{d}}{\mathrm{d}\theta} \ln f(X \mid \theta))^2\right]$, but as is well known [48], for exponential families this is equal to $-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} E\left[\ln f(X \mid \theta)\right]$, which matches $D^{(2)}(\cdot)$ exactly. Furthermore, for the mean value parameterisation $I(\mu) = 1/\mathrm{var}_{P_\mu}(X)$. We obtain

$$D(P_{\mu^*} \| P_{\hat{\mu}_i}) = \frac{1}{2} {\delta_i}^2 / \mathrm{var}_{P_{\mu^*}}(X) + \frac{1}{6} {\delta_i}^3 D^{(3)}(\mu^*) + \frac{1}{24} {\delta_i}^4 D^{(4)}(\ddot{\mu}_i). \tag{3.8}$$

We plug this expression back into (3.7), giving

$$\mathfrak{R}(P^*, U, \mathcal{M}, n) = \frac{1}{2\mathrm{var}_{P_{\mu^*}}(X)} \sum_{i=0}^{n-1} E_{P^*}\left[{\delta_i}^2\right] + R(n), \tag{3.9}$$

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

$$R(n) = \sum_{i=0}^{n-1} E_{P^*}\left[\frac{1}{6} {\delta_i}^3 D^{(3)}(\mu^*) + \frac{1}{24} {\delta_i}^4 D^{(4)}(\ddot{\mu}_i)\right], \tag{3.10}$$

and where μ and δ_i are random variables; note that although μ is not indexed it does depend on the index i. In Lemma 3.5.8 we show that $R(n) = O(1)$, giving:

$$\mathfrak{R}(P^*, U, \mathcal{M}, n) = O(1) + \frac{1}{2\mathrm{var}_{P_{\mu^*}}(X)} \sum_{i=0}^{n-1} E_{P^*}\left[(\hat{\mu}_i - \mu^*)^2\right]. \tag{3.11}$$

Note that $\hat{\mu}_i$ is almost the ML estimator. This suggests that each term in the sum of (3.11) should be almost equal to the variance of the ML estimator, which is $\mathrm{var}X/i$. Because of the slight modification that we made to the estimator, we get a correction term of $O(i^{-2})$ as established in Theorem 3.5.2:

$$\begin{aligned} \sum_{i=0}^{n-1} E_{P^*}\left[(\hat{\mu}_i - \mu^*)^2\right] &= \sum_{i=0}^{n-1} O\left((i+1)^{-2}\right) + \mathrm{var}_{P^*}(X) \sum_{i=0}^{n-1} (i+1)^{-1} \\ &= O(1) + \mathrm{var}_{P^*}(X) \ln n \end{aligned} \tag{3.12}$$

The combination of (3.11) and (3.12) completes the proof. □

3.3 Redundancy vs. Regret

The "goodness" of a universal code relative to a model $\mathcal{M}$ can be measured in several ways: rather than using redundancy (as we did here), one can also choose to measure code length differences in terms of *regret*, where one has a further choice between *expected regret* and *worst-case regret* [4]. Here we only discuss the implications of our result for the expected regret measure.

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

$$L_U(z^n) - \inf_{\theta\in\Theta}[-\ln P_\theta(z^n)]. \tag{3.13}$$

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

$$E_{Z^n\sim P^*}[\mathcal{R}(P^*, U, \mathcal{M}, n)] = E_{P^*}[L_U(Z^n) - \inf_{\theta\in\Theta}[-\ln P_\theta(Z^n)]].$$

To quantify the difference with redundancy, consider the function

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

and note that for any universal code, $\mathfrak{R} - E[\mathcal{R}] = d(n)$. In case $P^* \in \mathcal{M}$, then under regularity conditions on $\mathcal{M}$ and its parameterisation, it can be shown [23] that

$$\lim_{n\to\infty} d(n) = \frac{k}{2}, \tag{3.14}$$

where k is the dimension of $\mathcal{M}$. In our case, where P^* is not necessarily in $\mathcal{M}$, we have the following :

Theorem 3.3.1. *Let $\mathcal{X}$ be finite. Let P^*, P_μ and μ^* be as in Theorem 3.2.3. Then*

$$\lim_{n\to\infty} d(n) = \frac{1}{2}\frac{\mathrm{var}_{P^*}X}{\mathrm{var}_{P_{\mu^*}}X}. \tag{3.15}$$

Once we are dealing with 1-parameter families, in the special case that $P^* \in \mathcal{M}$, this result reduces to (3.14). We suspect that, under a condition similar to Condition 3.2.4, the same result still holds for general, not necessarily finite or countable or bounded $\mathcal{X}$, but we do not know the details. In any case, our result is sufficient to show that in some cases (namely, if $\mathcal{X}$ is finite), we have

$$\mathcal{R}(P^*, U, \mathcal{M}, n) = \frac{1}{2}\frac{\mathrm{var}_{P^*}X}{\mathrm{var}_{P_{\mu^*}}X}\ln n + O(1),$$

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

Incidentally, we can use Theorem 3.3.1 to substantiate the claim in Section 3.1, which stated that the Bayes (equipped with a strictly positive differentiable prior), NML and 2-part codes still achieve relative redundancy of $\frac{1}{2}\ln n$ if $P^* \neq \mathcal{M}$, at least if $\mathcal{X}$ is finite. Let us informally explain why this is the case. It is easy to show that Bayes, NML and (suitably defined) 2-part codes achieve regret $\frac{1}{2}\ln n + O(1)$ for *all* sequences $z_1, z_2, \ldots$ such that $\hat{\theta}(z^n)$ is bounded away from the boundary of the parameter space Θ_μ, for all large n [4, 38]. It then follows using, for example, the Chernoff bound that these codes must also achieve expected regret $\frac{1}{2}\ln n + O(1)$ for *all* distributions P^* on $\mathcal{X}$ that satisfy $E_{P^*}[X] = \mu^* \in \Theta_\mu$. Theorem 3.3.1 then shows that they also achieve relative redundancy $\frac{1}{2}\ln n + O(1)$ for *all* distributions P^* on $\mathcal{X}$ that satisfy $E_{P^*}[X] = \mu^* \in \Theta_\mu$. We omit further details.

3.4 Discussion

3.4.1 A Justification of Our Modification of the ML Estimator

A prequential code cannot be defined in terms of the ordinary ML estimator ($n_0 = 0$ in Definition 3.2.2) for two reasons. First, the ML estimator is undefined until the first outcome has been observed. Second, it may achieve infinite code lengths on the observed data. A simple example is the Bernoulli model. If we first observe $z_1 = 0$ and then $z_2 = 1$, the code length of z_2 according to the ordinary ML estimator of z_2 given z_1 would be $-\ln P_{\hat{\mu}(z_1)}(z_2) = -\ln 0 = \infty$. There are several ways to resolve this problem. We choose to add an "fake initial outcome". Another possibility that has been suggested (e.g., [26]) is to use the ordinary ML estimator, but to postpone using it until after m outcomes have been observed, where m is the smallest number such that $-\ln P_{\hat{\mu}(z^m)}(Z_{m+1})$ is guaranteed to be finite, no matter what value Z_{m+1} is realized. The first m outcomes may then be encoded by repeatedly using some code L_0 on outcomes of $\mathcal{Z}$, so that for $i \leq m$, the code length of z_i does not depend on the outcomes z^{i-1}. In the Bernoulli example, one could for example use the code corresponding to $P(Z_i = 1) = 1/2$, until and including the first i such that z^i includes both a 0 and a 1. It then takes i bits to encode the first z^i outcomes, no matter what they are. After that, one uses the prequential code with the standard ML estimator. It is easy to see (by slight modification of the proof) that our theorem still holds for this variation of prequential coding. Thus, our particular choice for resolving the startup problem is not crucial to obtaining our result. The advantage of our solution is that, as we now show, it allows us to interpret our modified ML estimator as a Bayesian MAP and Bayesian mean estimator as well, thereby showing that the same behaviour

can be expected for such estimators.

3.4.2 Prequential Models with Other Estimators

An attractive property of our generalisation of the ML estimator is that it actually also generalises two other commonly used estimators, namely the Bayesian maximum a-posteriori and Bayesian mean estimators.

The Bayesian maximum a-posteriori estimator can always be interpreted as an ML estimator based on the sample and some additional "fake data", provided that a conjugate prior is used ([12]; see also the notion of *ESS (Equivalent Sample Size) Priors* discussed in, for example, [51]). Therefore, the prequential ML model as defined above can also be interpreted as a prequential MAP model for that class of priors, and the whole analysis carries over to that setting.

For the Bayesian mean estimator, the relationship is slightly less direct. However, it follows from the work of Hartigan [42, Chapter 7] on the so-called "maximum likelihood prior", that by slightly modifying conjugate priors, we can construct priors such that the Bayesian mean also becomes of the form of our modified ML estimator.

Our whole analysis thus carries over to prequential codes based on these estimators. In fact, we believe that our result holds quite generally:

Conjecture 3.4.1. *Let $\mathcal{M}$ be a regular exponential family with sufficient statistic X and let $\mathcal{P}$ be the set of distributions on $\mathcal{Z}$ such that $E_{P^*}[X^4]$ exists. There exists* no "in-model" estimator *such that the corresponding prequential code achieves redundancy $\frac{1}{2}\ln n + O(1)$ for all $P^* \in \mathcal{P}$.*

Here, by an *in-model estimator* we mean an algorithm that takes as input any sample of arbitrary length and outputs a $P_\mu \in \mathcal{M}$. Let us contrast this with "out-model estimators": fix some prior on the parameter set Θ_μ and let $P(\mu \mid z_1, \ldots, z_{n-1})$ be the Bayesian posterior with respect to this prior and data $z_1, \ldots, z_{n-1}$. One can think of the Bayesian predictive distribution $P(z_n \mid z_1, \ldots, z_{n-1}) := \int_{\mu \in \Theta_\mu} P_\mu(z_n) P(\mu \mid z_1, \ldots, z_{n-1}) d\mu$ as an *estimate* of the distribution of Z, based on data $z_1, \ldots, z_{n-1}$. But unlike estimators as defined in the conjecture above, the resulting *Bayesian predictive estimator* will in general not be a member of $\mathcal{M}$, but rather of its convex closure: we call it an *out-model estimator*. The redundancy of the Bayesian universal model is equal to the *accumulated Kullback-Leibler (KL) risk* of the Bayesian predictive estimator [36]. Thus, the accumulated KL risk of the Bayesian predictive estimator is $\frac{1}{2}\ln n + O(1)$ even under misspecification. Thus, if our conjecture above holds true, then in-model estimators behave in a fundamentally different way from out-model estimators in terms of their asymptotic risk.

Example 8. The well-known Laplace and Krichevsky-Trofimov estimators for the Bernoulli model [38] define PML distributions according to Definition 3.2.2:

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

3.4.3 Practical Significance for Model Selection

As mentioned in the introduction, there are many results showing that in various contexts, if $P^* \in \mathcal{M}$, then the prequential ML code achieves optimal redundancy. These results strongly suggest that it is a very good alternative for (or at least approximation to) the NML or Bayesian codes in MDL model selection. Indeed, quoting Rissanen [71]:

> "If the encoder does not look ahead but instead computes the best parameter values from the past string, only, using an algorithm which the decoder knows, then no preamble is needed. The result is a *predictive* coding process, one which is quite different from the sum or integral formula in the stochastic complexity.[1] And it is only because of a certain inherent singularity in the process, as well as the somewhat restrictive requirement that the data must be ordered, that we do not consider the resulting predictive code length to provide another competing definition for the stochastic complexity, but rather regard it as an approximation."

Our result however shows that the prequential ML code may behave quite differently from the NML and Bayes codes, thereby strengthening the conclusion that it should not be taken as a definition of stochastic complexity. Although there is only a significant difference if data are distributed according to some $P^* \notin \mathcal{M}$, the difference is nevertheless very relevant in an MDL model selection context with disjoint models, even if one of the models under consideration *does* contain the "true" P^*. To see this, suppose we are comparing two models $\mathcal{M}_1$ and $\mathcal{M}_2$ for

[1] The stochastic complexity is the code length of the data $z_1, \ldots, z_n$ that can be achieved using the NML code.

the same data, and in fact, $P^* \in \mathcal{M}_1 \cup \mathcal{M}_2$. For concreteness, assume $\mathcal{M}_1$ is the Poisson family and $\mathcal{M}_2$ is the geometric family. We want to decide which of these two models best explains the data. According to the MDL Principle, we should associate with each model a universal code (preferably the NML code). We should then pick the model such that the corresponding universal code length of the data is minimised. Now suppose we use the prequential ML code lengths rather than the NML code lengths. Without loss of generality suppose that $P^* \in \mathcal{M}_1$. Then $P^* \notin \mathcal{M}_2$. This means that the code length relative to $\mathcal{M}_1$ behaves essentially like the NML code length, but the code length relative to $\mathcal{M}_2$ behaves differently – at least as long as the variances do not match (which for example, is forcibly the case if $\mathcal{M}_1$ is Poisson and $\mathcal{M}_2$ is geometric). This introduces a bias in the model selection scheme. In the previous chapter we found experimentally that the error rate for model selection based on the prequential ML code decreases more slowly than when other universal codes are used. Even though in some cases the redundancy grows *more slowly* than $\frac{1}{2} \ln n$, so that the prequential ML code is more efficient than the NML code, we explained that model selection based on the prequential ML codes must nevertheless always behave worse than Bayesian and NML-based model selection. The practical relevance of this phenomenon stems from the fact that the prequential ML code lengths are often a lot easier to compute than the Bayes or NML codes. They are often used in applications [61, 52], so that is important to determine when this can be done safely.

3.4.4 Theoretical Significance

The result is also of theoretical-statistical interest: our theorem can be re-interpreted as establishing bounds on the asymptotic *Kullback-Leibler risk* of density estimation using ML and Bayes estimators under misspecification ($P^* \notin \mathcal{M}$). Our result implies that, under misspecification, the KL risk of estimators such as ML, which are required to lie in the model $\mathcal{M}$, behaves in a fundamentally different way from the KL risk of estimators such as the Bayes predictive distribution, which are not restricted to lie in $\mathcal{M}$. Namely, we can think of every universal model U defined as a random process on infinite sequences as an *estimator* in the following way: define, for all n,

$$\breve{P}_n := \Pr_U(Z_{n+1} = \cdot \mid Z_1 = z_1, \ldots, Z_n = z_n),$$

a function of the sample $z_1, \ldots, z_n$. $\breve{P}_n$ can be thought of as the "estimate of the true data generating distribution upon observing $z_1, \ldots, z_n$". In case U is the prequential ML model, $\breve{P}_n = P_{\hat{\mu}_n}$ is simply our modified ML estimator. However, universal models other than PML, $\breve{P}_n$ does not always lie in $\mathcal{M}$. An example is the Bayesian universal code defined relative to some prior w. This code has lengths $L'(z^n) := -\ln \int P_\mu(z^n) w(\mu) \, \mathrm{d}\mu$ [38]. The corresponding estimator is the *Bayesian posterior predictive distribution* $P_{\mathrm{Bayes}}(z_{i+1} \mid z^i) := \int P_\mu(z_{i+1}) w(\mu \mid z^i) \, \mathrm{d}\mu$ [38].

The Bayesian predictive distribution is a mixture of elements of $\mathcal{M}$. We will call standard estimators like the ML estimator, which are required to lie in $\mathcal{M}$, *in-model* estimators. Estimators like the Bayesian predictive distribution will be called *out-model*.

Let now $\breve{P}_n$ be any estimator, in-model or out-model. Let $\breve{P}_{z^n}$ be the distribution estimated for a particular realized sample z^n. We can measure the closeness of $\breve{P}_{z^n}$ to P_{μ^*}, the distribution in $\mathcal{M}$ closest to P^* in KL-divergence, by considering the *extended KL divergence*

$$D^*(P_{\mu^*} \| \breve{P}_{z^n}) = E_{Z \sim P^*}[-\ln \breve{P}_{z^n}(Z) - [-\ln P_{\mu^*}(Z)]].$$

We can now consider the *expected* KL divergence between P_{μ^*} and $\breve{P}_n$ after observing a sample of length n:

$$E_{Z_1,\ldots,Z_n \sim P^*}[D^*(P_{\mu^*} \| \breve{P}_n)]. \tag{3.16}$$

In analogy to the definition of "ordinary" *KL risk* [4], we call (3.16) the *extended KL risk*. We recognise the redundancy of the PML distribution as the accumulated expected KL risk of our modified ML estimator (see Proposition 3.5.7 and Lemma 3.5.6). In exactly the same way as for PML, the redundancy of the Bayesian code can be re-interpreted as the accumulated KL risk of the Bayesian predictive distribution. With this interpretation, our Theorem 3.2.3 expresses that under misspecification, the cumulative KL risk of the ML estimator differs from the cumulative KL risk of the Bayes estimator by a term of $\Theta(\ln n)$. If our conjecture that *no* in-model estimator can achieve redundancy $\frac{1}{2}\ln n + O(1)$ for all μ^* and all P^* with finite variance is true (Section 3.4.2), then it follows that the KL risk for in-model estimators behaves in a fundamentally different way from the KL risk for out-model estimators, and that out-model estimators are needed to achieve the optimal constant $c = 1$ in the redundancy $\frac{1}{2}c\ln n + O(1)$.

3.5 Building Blocks of the Proof

The proof of Theorem 3.2.3 is based on Lemma 3.5.6 and Lemma 3.5.8. These Lemmas are stated and proved in Sections 3.5.2 and 3.5.3, respectively. The proofs of Theorem 3.2.3 and Theorem 3.3.1, as well as the proof of both Lemmas, are based on a number of generally useful results about probabilities and expectations of deviations between the average and the mean of a random variable. We first list list these deviation-related results.

3.5.1 Results about Deviations between Average and Mean

Lemma 3.5.1. *Let $X, X_1, X_2, \ldots$ be i.i.d. with mean 0. Then $E\left[\left(\sum_{i=1}^n X_i\right)^2\right] = n\mathrm{var}(X)$.*

Proof. The lemma is clearly true for $n = 0$. Suppose it is true for some n. Abbreviate $S_n := \sum_{i=1}^n X_i$. We have $E[S_n] = \sum EX = 0$. Now we find $E\left[S_{n+1}^2\right] = E\left[(S_n + X)^2\right] = E\left[S_n^2\right] + 2E[S_n]EX + E[X^2] = (n+1)\mathrm{var}(X)$. The proof follows by induction. □

Theorem 3.5.2. *Let $X, X_1, \ldots$ be i.i.d. random variables, define $\hat{\mu}_n := (n_0 \cdot x_0 + \sum_{i=1}^n X_i)/(n + n_0)$ and $\mu^* = E[X]$. If $\mathrm{var}X < \infty$, then $E\left[(\hat{\mu}_n - \mu^*)^2\right] = O\left((n+1)^{-2}\right) + \mathrm{var}(X)/(n+1)$.*

Proof. We define $Y_i := X_i - \mu^*$; this can be seen as a new sequence of i.i.d. random variables with mean 0 and $\mathrm{var}Y = \mathrm{var}X$. We also set $y_0 := x_0 - \mu^*$. Now we have:

$$\begin{aligned}
E\left[(\hat{\mu}_n - \mu^*)^2\right] &= E\left[\left(n_0 \cdot y_0 + \sum_{i=1}^n Y_i\right)^2\right](n+n_0)^{-2} \\
&= E\left[(n_0 \cdot y_0)^2 + 2n_0 \cdot y_0 \sum_{i=1}^n Y_i + \left(\sum_{i=1}^n Y_i\right)^2\right](n+n_0)^{-2} \\
&= O\left((n+1)^{-2}\right) + E\left[\left(\sum_{i=1}^n Y_i\right)^2\right](n+n_0)^{-2} \\
&\overset{(*)}{=} O\left((n+1)^{-2}\right) + n\mathrm{var}(Y)(n+n_0)^{-2} \\
&= O\left((n+1)^{-2}\right) + \mathrm{var}(X)/(n+1),
\end{aligned}$$

where $(*)$ follows by Lemma 3.5.1. □

The following theorem is of some independent interest.

Theorem 3.5.3. *Suppose $X, X_1, X_2, \ldots$ are i.i.d. with mean 0. If the first $k \in \mathbb{N}$ moments of X exist, then we have Then $E\left[\left(\sum_{i=1}^n X_i\right)^k\right] = O\left(n^{\lfloor \frac{k}{2} \rfloor}\right)$.*

Remark It follows as a special case of Theorem 2 of [98] that $E\left[|\sum_{i=1}^n X_i|^k\right] = O(n^{\frac{k}{2}})$ which almost proves this lemma and which would in fact be sufficient for our purposes. We use this lemma instead which has an elementary proof.

Proof. We have:

$$E\left[\left(\sum_{i=1}^n X_i\right)^k\right] = E\left[\sum_{i_1=1}^n \cdots \sum_{i_k=1}^n X_{i_1} \cdots X_{i_k}\right] = \sum_{i_1=1}^n \cdots \sum_{i_k=1}^n E\left[X_{i_1} \cdots X_{i_k}\right].$$

We define the *frequency sequence* of a term to be the sequence of exponents of the different random variables in the term, in decreasing order. For a frequency sequence $f_1, \ldots, f_m$, we have $\sum_{i=1}^{m} f_i = k$. Furthermore, using independence of the different random variables, we can rewrite $E[X_{i_1} \cdots X_{i_k}] = \prod_{i=1}^{m} E[X^{f_i}]$ so the value of each term is determined by its frequency sequence. By computing the number of terms that share a particular frequency sequence, we obtain:

$$E\left[\left(\sum_{i=1}^{n} X_i\right)^k\right] = \sum_{f_1+\ldots+f_m=k} \binom{n}{m} \binom{k}{f_1, \ldots, f_m} \prod_{i=1}^{m} E[X^{f_i}].$$

To determine the asymptotic behaviour, first observe that the frequency sequence $f_1, \ldots, f_m$ of which the contribution grows fastest in n is the longest sequence, since for that sequence the value of $\binom{n}{m}$ is maximised as $n \to \infty$. However, since the mean is zero, we can discard all sequences with an element 1, because the for those sequences we have $\prod_{i=1}^{m} E[X^{f_i}] = 0$ so they contribute nothing to the expectation. Under this constraint, we obtain the longest sequence for even k by setting $f_i = 2$ for all $1 \leq i \leq m$; for odd k by setting $f_1 = 3$ and $f_i = 2$ for all $2 \leq i \leq m$; in both cases we have $m = \left\lfloor \frac{k}{2} \right\rfloor$. The number of terms grows as $\binom{n}{m} \leq n^m/m! = O(n^m)$; for $m = \left\lfloor \frac{k}{2} \right\rfloor$ we obtain the upper bound $O(n^{\lfloor \frac{k}{2} \rfloor})$. The number of frequency sequences is finite and does not depend on n; since the contribution of each one is $O\left(n^{\lfloor \frac{k}{2} \rfloor}\right)$, so must be the sum. □

Theorem 3.5.4. *Let $X, X_1, \ldots$ be i.i.d. random variables, define $\hat{\mu}_n := (n_0 \cdot x_0 + \sum_{i=1}^{n} X_i)/(n + n_0)$ and $\mu^* = E[X]$. If the first k moments of X exist, then $E[(\hat{\mu}_n - \mu^*)^k] = O(n^{-\lceil \frac{k}{2} \rceil})$.*

Proof. The proof is similar to the proof for Theorem 3.5.2. We define $Y_i := X_i - \mu^*$; this can be seen as a new sequence of i.i.d. random variables with mean 0, and $y_0 := x_0 - \mu^*$. Now we have:

$$\begin{aligned} E\left[(\hat{\mu}_n - \mu^*)^k\right] &= E\left[\left(n_0 \cdot y_0 + \sum_{i=1}^{n} Y_i\right)^k\right] (n + n_0)^{-k} \\ &= O\left(n^{-k}\right) \sum_{p=0}^{k} \binom{k}{p} (n_0 \cdot y_0)^p E\left[\left(\sum_{i=1}^{n} Y_i\right)^{k-p}\right] \\ &= O\left(n^{-k}\right) \sum_{p=0}^{k} \binom{k}{p} (n_0 \cdot y_0)^p \cdot O\left(n^{\lfloor \frac{k-p}{2} \rfloor}\right). \end{aligned}$$

In the last step we used Theorem 3.5.3 to bound the expectation. We sum $k+1$ terms of which the term for $p = 0$ grows fastest in n, so the expression is $O(n^{-\lceil \frac{k}{2} \rceil})$ as required. □

Theorem 3.5.4 concerns the *expectation* of the deviation of $\hat{\mu}_n$. We also need a bound on the *probability* of large deviations. To do that we have the following separate theorem:

Theorem 3.5.5. *Let $X, X_1, \ldots$ be i.i.d. random variables, define $\hat{\mu}_n := (n_0 \cdot x_0 + \sum_{i=1}^{n} X_i)/(n + n_0)$ and $\mu^* = E[X]$. Let $k \in \{0, 2, 4, \ldots\}$. If the first k moments exists then $P(|\hat{\mu}_n - \mu^*| \geq \delta) = O\left(n^{-\frac{k}{2}} \delta^{-k}\right)$.*

Proof.

$$\begin{aligned} P^*(|\hat{\mu}_n - \mu^*| \geq \delta) &= P^*\left((\hat{\mu}_n - \mu^*)^k \geq \delta^k\right) \\ &\leq E\left[(\hat{\mu}_n - \mu^*)^k\right] \delta^{-k} \quad \text{(by Markov's inequality)} \\ &= O\left(n^{-\frac{k}{2}} \delta^{-k}\right) \quad \text{(by Theorem 3.5.4)} \end{aligned}$$ □

3.5.2 Lemma 3.5.6: Redundancy for Exponential Families

Lemma 3.5.6. *Let U be a PML distribution model and $\mathcal{M}$ be an exponential family as in Theorem 3.2.3. We have*

$$\mathfrak{R}(P^*, U, \mathcal{M}, n) = \sum_{i=0}^{n-1} \underset{\hat{\mu}_i \sim P^*}{E} \left[D(P_{\mu^*} \parallel P_{\hat{\mu}_i}) \right].$$

(Here, the notation $\hat{\mu}_i \sim P^*$ means that we take the expectation with respect to P^* over data sequences of length i, of which $\hat{\mu}_i$ is a function.)

Proof. We have:

$$\arg\min_{\mu} E_{P^*}\left[-\ln P_\mu(X^n)\right] = \arg\min_{\mu} E_{P^*}\left[\ln \frac{P_{\mu^*}(X^n)}{P_\mu(X^n)}\right] = \arg\min_{\mu} D(P_{\mu^*} \parallel P_\mu).$$

In the last step we used Proposition 3.5.7 below. The divergence is minimised when $\mu = \mu^*$ [48], so we find that:

$$\begin{aligned} &\mathfrak{R}(P^*, U, \mathcal{M}, n) = E_{P^*}[-\ln U(X^n)] - E_{P^*}[-\ln P_{\mu^*}(X^n)] = E_{P^*}\left[\ln \frac{P_{\mu^*}(X^n)}{U(X^n)}\right] \\ &= E_{P^*}\left[\sum_{i=0}^{n-1} \ln \frac{P_{\mu^*}(X_i)}{P_{\hat{\mu}_i}(X_i)}\right] = \sum_{i=0}^{n-1} E_{P^*}\left[\ln \frac{P_{\mu^*}(X_i)}{P_{\hat{\mu}_i}(X_i)}\right] = \sum_{i=0}^{n-1} \underset{\hat{\mu}_i \sim P}{E} \left[D(P_{\mu^*} \parallel P_{\hat{\mu}_i}) \right]. \end{aligned} \tag{3.17}$$

Here, the last step again follows from Proposition 3.5.7. □

Proposition 3.5.7. *Let $X \sim P^*$ with mean μ^*, and let P_μ index an exponential family with sufficient statistic X, so that P_{μ^*} exists. We have:*

$$E_{P^*}\left[-\ln \frac{P_{\mu^*}(X)}{P_\theta(X)}\right] = D(P_{\mu^*} \parallel P_\theta)$$

Proof. Let $\eta(\cdot)$ denote the function mapping parameters in the mean value parameterisation to the natural parameterisation. (It is the inverse of the function $\mu(\cdot)$ which was introduced in the discussion of exponential families.) By working out both sides of the equation we find that they both reduce to:

$$\eta(\mu^*)\mu^* + \ln \mathrm{Z}(\eta(\mu^*)) - \eta(\theta)\mu^* - \ln \mathrm{Z}(\eta(\theta)).$$ □

3.5.3 Lemma 3.5.8: Convergence of the sum of the remainder terms

Lemma 3.5.8. *Let $R(n)$ be defined as in (3.10). Then*

$$R(n) = O(1).$$

Proof. We omit irrelevant constants. We abbreviate $\frac{\mathrm{d}^k}{\mathrm{d}\mu^k} D(P_{\mu^*} \| P_\mu) = D^{(k)}(\mu)$ as in the proof of Theorem 3.2.3. First we consider the third order term. We write $E_{\delta_i \sim P^*}$ to indicate that we take the expectation over data which is distributed according to P^*, of which δ_i is a function. We use Theorem 3.5.4 to bound the expectation of ${\delta_i}^3$; under the condition that the first three moments exist, which is assumed to be the case, we obtain:

$$\sum_{i=0}^{n-1} \underset{\delta_i \sim P^*}{E}\left[{\delta_i}^3 D^{(3)}(\mu^*)\right] = D^{(3)}(\mu^*) \sum_{i=0}^{n-1} E[\delta_i^3] = D^{(3)}(\mu^*) \sum_{i=0}^{n-1} O((i+1)^{-2}) = O(1).$$

(The constants implicit in the big-ohs are the same across terms.)

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

$$\sum_{i=0}^{n-1} \underset{\delta_i \sim P^*}{E}\left[{\delta_i}^4 D^{(4)}(\ddot{\mu}_i)\right] \;=\; \sum_{i=0}^{n-1} \sum_j P^*(\delta_i \in \Delta_j) \underset{\delta_i \sim P^*}{E}\left[{\delta_i}^4 D^{(4)}(\ddot{\mu}_i) \mid \delta_i \in \Delta_j\right]$$

If we can establish that the sum converges for all regions Δ_j for $j \geq 0$, then we can use a symmetrical argument to establish convergence for Δ_- as well, so it suffices if we restrict attention to $j \geq 0$. First we show convergence for Δ_0. In

this case, the basic idea is that since the remainder $D^{(4)}(\ddot{\mu}_i)$ is well-defined over the interval $\mu^* \leq \mu < \mu^* + a$, we can bound it by its extremum on that interval, namely $m := \sup_{\mu \in [\mu^*, \mu^*+a)} \left| D^{(4)}(\ddot{\mu}_i) \right|$. Now we get:

$$\left| \sum_{i=0}^{n-1} P^*(\delta_i \in \Delta_0) E\left[{\delta_i}^4 D^{(4)}(\ddot{\mu}_i) \mid \delta_i \in \Delta_0 \right] \right| \leq \left| \sum_{i=0}^{n-1} 1 \cdot E\left[{\delta_i}^4 \left| D^{(4)}(\ddot{\mu}_i) \right| \right] \right| \leq \left| m \sum_i E\left[{\delta_i}^4 \right] \right|. \quad (3.18)$$

Using Theorem 3.5.4 we find that $E[{\delta_i}^4]$ is $O((i+1)^{-2})$ of which the sum converges. Theorem 3.5.4 requires that the first four moments of P^* exist, but this is guaranteed to be the case: either the outcomes are bounded from both sides, in which case all moments necessarily exist, or the existence of the required moments is part of the condition on the main theorem.

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

$$\sum_{i=0}^{n-1} \sum_{j=1}^{\infty} P^*(\delta_i \in \Delta_j) E\left[{\delta_i}^4 D^{(4)}(\ddot{\mu}_i) \mid \delta_i \in \Delta_j \right].$$

To show convergence, we bound the absolute value of this expression from above. The δ_i in the expectation is at most $a + j$. Furthermore $D^{(4)}(\ddot{\mu}_i) = O(\mu^{k-6})$ by assumption on the main theorem, where $\mu \in [a + j - 1, a + j)$. Depending on k, both boundaries could maximise this function, but it is easy to check that in both cases the resulting function is $O(j^{k-6})$. So we get:

$$\left| \ldots \right| \leq \sum_{i=0}^{n-1} \sum_{j=1}^{\infty} P^*(|\delta_i| \geq a + j - 1)(a + j)^4 O(j^{k-6}).$$

Since we know from the condition on the main theorem that the first $k \geq 4$ moments exist, we can apply Theorem 3.5.5 to find that $P(|\delta_i| \geq a + j - 1) = O(i^{-\lceil \frac{k}{2} \rceil}(a + j - 1)^{-k}) = O(i^{-\frac{k}{2}})O(j^{-k})$ (since k has to be even); plugging this into the equation and simplifying we obtain $\sum_i O(i^{-\frac{k}{2}}) \sum_j O(j^{-2})$. For $k \geq 4$ this expression converges.

Now we consider the case where the outcomes are bounded from above by g. This case is more complicated, since now we have made no extra assumptions as to existence of the moments of P^*. Of course, if the outcomes are bounded from both sides, then all moments necessarily exist, but if the outcomes are unbounded from below this may not be true. We use a trick to remedy this: we map all outcomes

into a new domain in such a way that all moments of the transformed variables are guaranteed to exist. Any constant x^- defines a mapping $g(x) := \max\{x^-, x\}$. Furthermore we define the random variables $Y_i := g(X_i)$, the initial outcome $y_0 := g(x_0)$ and the mapped analogues of μ^* and $\hat{\mu}_i$, respectively: $\mu^\dagger$ is defined as the mean of Y under P and $\tilde{\mu}_i := (y_0 \cdot n_0 + \sum_{j=1}^{i} Y_j)/(i + n_0)$. Since $\tilde{\mu}_i \geq \hat{\mu}_i$, we can bound:

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

By choosing x^- small enough, we can bring $\mu^\dagger$ and μ^* arbitrarily close together; in particular we can choose x^- such that $a + \mu^* - \mu^\dagger > 0$ so that application of Theorem 3.5.5 is safe. It reveals that the summed probability is $O(i^{-\frac{k}{2}})$. Now we bound $D^{(4)}(\ddot{\mu}_i)$ which is $O((g-\mu)^{-m})$ for some $m \in \mathbb{N}$ by the condition on the main theorem. Here we use that $\ddot{\mu}_i \leq \hat{\mu}_i$; the latter is maximised if all outcomes equal the bound g, in which case the estimator equals $g - n_0(g - x_0)/(i + n_0) = g - O(i^{-1})$. Putting all of this together, we get $\sup \left| D^{(4)}(\ddot{\mu}_i) \right| = O((g-\mu)^{-m}) = O(i^m)$; if we plug this into the equation we obtain:

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

This converges if we choose $k \geq m/2$. As the construction of the mapping $g(\cdot)$ ensures that all moments exist, the first $m/2$ moments certainly must exist. This completes the proof. □

3.6 Proof of Theorem 3.3.1

We use the same conventions as in the proof of Theorem 3.2.3. Specifically, we concentrate on the random variables $X_1, X_2, \ldots$ rather than $Z_1, Z_2, \ldots$, which is justified by Equation (3.7). Let $f(x^n) = -\ln P_{\mu^*}(x^n) - [\inf_{\mu \in \Theta_\mu} - \ln P_\mu(x^n)]$. Within this section, $\hat{\mu}(x^n)$ is defined as the ordinary ML estimator. Note that, if x^n is such that its ML estimate is defined, then $f(x^n) = -\ln P_{\mu^*}(x^n) + \ln P_{\hat{\mu}(x^n)}(x^n)$.

Note $d(n) = E_{P^*}[f(X^n)]$. Let $h(x)$ be the carrier of the exponential family under consideration (see Definition 3.2.1). Without loss of generality, we assume

$h(x) > 0$ for all x in the finite set $\mathcal{X}$. Let $a_n^2 = n^{-1/2}$. We can write

$$\begin{aligned} d(n) \;=\; E_{P^*}[f(X^n)] \;=\; & \pi_n \quad E_{P^*}[f(X^n) \mid (\mu^* - \hat{\mu}_n)^2 \geq a_n^2] \\ + \; & (1-\pi_n) \quad E_{P^*}[f(X^n) \mid (\mu^* - \hat{\mu}_n)^2 < a_n^2], \end{aligned} \tag{3.19}$$

where $\pi_n = P^*((\mu^* - \hat{\mu}_n)^2 \geq a_n^2)$. We determine $d(n)$ by bounding the two terms on the right of (3.19). We start with the first term. Since X is bounded, all moments of X exists under P^*, so we can bound π_n using Theorem 3.5.5 with $k = 8$ and $\delta = a_n = n^{-1/4}$. (Note that the theorem in turn makes use of Theorem 3.5.4 which remains valid when we use $n_0 = 0$.) This gives

$$\pi_n = O(n^{-2}). \tag{3.20}$$

Note that for all $x^n \in \mathcal{X}^n$, we have

$$0 \leq f(x^n) \leq \sup_{x^n \in \mathcal{X}^n} f(x^n) \leq \sup_{x^n \in \mathcal{X}^n} -\ln P_{\mu^*}(x^n) \leq nC, \tag{3.21}$$

where C is some constant. Here the first inequality follows because $\hat{\mu}$ maximises $\ln P_{\hat{\mu}(x^n)}(x^n)$ over μ; the second is immediate; the third follows because we are dealing with discrete data, so that $P_{\hat{\mu}}$ is a probability mass function, and $P_{\hat{\mu}}(x^n)$ must be ≤ 1. The final inequality follows because μ^* is in the interior of the parameter space, so that the natural parameter $\eta(\mu^*)$ is in the interior of the natural parameter space. Because X is bounded and we assumed $h(x) > 0$ for all $x \in \mathcal{X}$, it follows by the definition of exponential families that $\sup_{x \in \mathcal{X}} -\ln P_{\mu^*}(x) < \infty$.

Together (3.20) and (3.21) show that the expression on the first line of (3.19) converges to 0, so that (3.19) reduces to

$$d(n) = (1-\pi_n) E_{P^*}[f(X^n) \mid (\mu^* - \hat{\mu}_n)^2 < a_n^2] + O(n^{-1}). \tag{3.22}$$

To evaluate the term inside the expectation further we first Taylor approximate $f(x^n)$ around $\hat{\mu}_n = \hat{\mu}(x^n)$, for given x^n with $(\mu^* - \hat{\mu}_n)^2 < a_n^2 = 1/\sqrt{n}$. We get

$$f(x^n) = -(\mu^* - \hat{\mu}_n) \frac{d}{d\mu} \ln P_{\hat{\mu}_n}(x^n) + n \frac{1}{2} (\mu^* - \hat{\mu}_n)^2 I(\mu_n), \tag{3.23}$$

where I is the Fisher information (as defined in the proof of the main theorem) and μ_n lies in between μ^* and $\hat{\mu}$, and depends on the data x^n. Since the first derivative of μ at the ML estimate $\hat{\mu}$ is 0, the first-order term is 0. Therefore $f(x^n) = \frac{1}{2} n (\mu^* - \hat{\mu}_n)^2 I(\mu_n)$, so that

$$\frac{1}{2} n g(n) \inf_{\mu \in [\mu^* - a_n, \mu^* + a_n]} I(\mu) \leq E_{P^*}[f(X^n) \mid (\mu^* - \hat{\mu}_n)^2 < a_n^2] \leq \frac{1}{2} n g(n) \sup_{\mu \in [\mu^* - a_n, \mu^* + a_n]} I(\mu),$$

where we abbreviated $g(n) := E_{P^*}[(\mu^* - \hat{\mu}_n)^2 \mid (\mu^* - \hat{\mu}_n)^2 < a_n^2]$. Since $I(\mu)$ is smooth and positive, we can Taylor-approximate it as $I(\mu^*) + O(n^{-\frac{1}{4}})$, so we obtain the bound:

$$E_{P^*}[f(X^n) \mid (\mu^* - \hat{\mu}_n)^2 < a_n^2] = n g(n) \left(\frac{1}{2} I(\mu^*) + O(n^{-\frac{1}{4}}) \right). \tag{3.24}$$

To evaluate $g(n)$, note that we have

$$E_{P^*}[(\mu^* - \hat{\mu}_n)^2] = \pi_n E_{P^*}[(\mu^* - \hat{\mu}_n)^2 \mid (\mu^* - \hat{\mu}_n)^2 \geq a_n^2] + (1 - \pi_n)g(n). \quad (3.25)$$

Using Theorem 3.5.2 with $n_0 = 0$ we rewrite the expectation on the left hand side as $\mathrm{var}_{P^*}X/n$. Subsequently reordering terms we obtain:

$$g(n) = \frac{(\mathrm{var}_{P^*}X)/n - \pi_n E_{P^*}[(\mu^* - \hat{\mu}_n)^2 \mid (\mu^* - \hat{\mu}_n)^2 \geq a_n^2]}{1 - \pi_n}. \quad (3.26)$$

Plugging this into bound (3.24), and multiplying both sides by $1 - \pi_n$, we get:

$$(1 - \pi_n)E_{P^*}[f(X^n) \mid (\mu^* - \hat{\mu}_n)^2 < a_n^2] = \\ \left(\mathrm{var}_{P^*}X - n\pi_n E_{P^*}[(\mu^* - \hat{\mu}_n)^2 \mid (\mu^* - \hat{\mu}_n)^2 \geq a_n^2]\right)\left(\frac{1}{2}I(\mu^*) + O(n^{-\frac{1}{4}})\right).$$

Since X is bounded, the expectation on the right must lie between 0 and some constant C. Using $\pi_n = O(n^{-2})$ and the fact that $I(\mu^*) = 1/\mathrm{var}_{P_{\mu^*}}X$, we get

$$(1 - \pi_n)E_{P^*}[f(X^n) \mid (\mu^* - \hat{\mu}_n)^2 < a_n^2] = \frac{1}{2}\frac{\mathrm{var}_{P^*}X}{\mathrm{var}_{P_{\mu^*}}X} + O(n^{-\frac{1}{4}}).$$

The result follows if we combine this with (3.22).

3.7 Conclusion and Future Work

In this paper we established two theorems about the relative redundancy, defined in Section 3.1:

1. A particular type of universal code, the *prequential ML code* or *ML plug-in code*, exhibits behaviour that we found unexpected. While other important universal codes such as the NML/Shtarkov and Bayesian codes, achieve a regret of $\frac{1}{2}\ln n$, where n is the sample size, the prequential ML code achieves a relative redundancy of $\frac{1}{2}\frac{\mathrm{var}_{P^*}X}{\mathrm{var}_{P_{\mu^*}}X}\ln n$. (Sections 3.1 and 3.2.)

2. At least for finite sample spaces, the relative redundancy is very close to the expected regret, the difference going to $\frac{1}{2}\frac{\mathrm{var}_{P^*}X}{\mathrm{var}_{P_{\mu^*}}X}$ as the sample size increases (Section 3.3, Theorem 3.3.1). In future work, we hope to extend this theorem to general 1-parameter exponential families with arbitrary sample spaces.

There is a substantial amount of literature in which the regret for the prequential ML 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.

The first result is robust with respect to slight variations in the definition of the prequential ML 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. Our framework thus also covers prequential codes that use other point estimators such as the Bayesian MAP and mean estimators defined relative to a large class of reasonable priors. In Section 3.4.2 we conjecture that no matter what in-model estimator is used, the prequential model cannot yield a relative redundancy of $\frac{1}{2} \ln n$ independently of the variance of the data generating distribution.