Prequential Plug-In Codes that Achieve Optimal Redundancy Rates even if the Model is Wrong

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

***Abstract*— We analyse the prequential plug-in codes relative to one-parameter exponential families $\mathcal{M}$. We show that if data are sampled i.i.d. from some distribution outside $\mathcal{M}$, then the redundancy of any plug-in prequential code grows at rate larger than $\frac{1}{2}\ln n$ in the worst case. This means that plug-in codes, such as the Rissanen-Dawid ML code, may behave inferior to other important universal codes such as the 2-part MDL, Shtarkov and Bayes codes, for which the redundancy is always $\frac{1}{2}\ln n + O(1)$. However, we also show that a slight modification of the ML plug-in code, "almost" in the model, does achieve the optimal redundancy even if the the true distribution is outside $\mathcal{M}$.**

I. Introduction

We resolve two open problems from [1] concerning universal codes of the predictive plug-in type, also known as "prequential" codes. These codes were introduced independently by Rissanen [2] in the context of MDL learning and by Dawid [3], who proposed them as probability forecasting strategies rather than directly as codes. Roughly, the plug-in codes relative to parametric model $\mathcal{M} = \{M_\theta \mid \theta \in \Theta\}$ work by sequentially coding each outcome x_i based on an an estimator $\bar{\theta}_{i-1} = \bar{\theta}(x^{i-1})$ for all previous outcomes $x^{i-1} = x_1, \ldots, x_{i-1}$, leading to codelength (log loss) $-\ln M_{\bar{\theta}_{i-1}}(x_i)$, where M_θ denotes the probability density or mass function indexed by θ. If we take $\bar{\theta}_i = \hat{\theta}_i$ equal to the ML (maximum likelihood) estimator, we call the resulting code the "ML plug-in code".

There are many papers about the redundancy and/or expected regret for the ML plug-in codes, for a large variety of models including multivariate exponential families, ARMA processes, regression models and so on. Examples are [4], [5], [6]. In all these papers the ML plug-in code is shown to achieve an asymptotic expected regret or redundancy of $\frac{k}{2}\ln n + O(1)$, where k is the number of parameters of the model and n is the sample size. This matches the behaviour of the Shtarkov, Bayesian and two-part universal codes and is optimal in several ways, see [7]; since the ML plug-in codes are often easier to calculate than any of these other three codes, this appears to be a strong argument for using them in practical data compression and MDL-style model selection. Yet, more recently [8], [9], [10], it was shown that, at least for single-parameter exponential family models, when the data are generated i.i.d. $\sim P$, the redundancy in fact grows as $\frac{1}{2}\ln n \cdot \frac{\mathrm{var}_P X}{\mathrm{var}_M X}$, where M is the distribution in $\mathcal{M}$ that is closest to P in Kullback-Leibler divergence, i.e. it minimizes $D(P\|M)$; a related result for linear regression is in [11]. In contrast to the other cited works, [8], [9], [10], [11] do not assume that $P \in \mathcal{M}$: the model may be *misspecified*. Yet *if* $P \in \mathcal{M}$, then we have $M = P$ so that the redundancy grows like it does in the other universal models. But when $M \neq P$, the Shtarkov, Bayes and universal codes typically still achieve asymptotic expected regret $\frac{1}{2}\ln n$, whereas the plug-in codes behave differently. [8], [10] show that this leads to substantially inferior performance of the plug-in codes in practical MDL model selection.

A. The Two Open Problems/Conjectures

In general, the estimator for $\mathcal{M}$ based on x^{i-1} need not be an element of the parametric model $\mathcal{M}$; for example, we may think of the Bayesian predictive distribution as an estimator relative to $\mathcal{M}$, even though it is "out-model": rather than a single element of $\mathcal{M}$, it is a mixture of distributions in $\mathcal{M}$, each weighted by their posterior density (see Section IV for an example). We may thus re-interpret Bayesian universal codes as prequential codes based on "out-model" estimators. From now on, we reserve the term "prequential plug-*in* code", abbreviated to just "plug-in code", for codes based on "*in*-model" estimators, i.e. estimators required to lie within $\mathcal{M}$. When we call a code just "prequential", it may be sequentially constructed from either in-model or out-model estimators. [9] established a nonstandard redundancy, different from $(k/2)\ln n$, only for ML and closely related plug-in codes. [1, Open Problem Nr. 2] conjectured that a similar result should hold for *all* plug-in codes, even if they are based on in-model estimators very different from the ML estimator: the conjecture was that *no* plug-in code can achieve guaranteed redundancy of $(k/2)\ln n$ if data are i.i.d. $\sim P$ and $P \neq M$. Our first main result, Theorem 1 below, shows that, essentially, this conjecture is true for general one-parameter exponential families ($k = 1$). Specifically, the redundancy can become much larger than $(1/2)\ln n$ if $P \notin \mathcal{M}$.

The second related conjecture [1, Open Problem Nr. 3] concerned the fact that for the normal location family with constant variance σ^2, the Bayesian predictive distribution based on data x^{i-1} and a normal prior looks "almost" like an in-model estimator for x^{i-1}, and hence the resulting code looks "almost" like a plug-in code: the Bayes predictive distribution is equal to the normal distribution for X_i with mean equal to the ML estimator $\hat{\mu}(x^{i-1})$ but with a variance of order $\sigma^2 + O(1/n)$, i.e. slightly larger than the variance σ^2 of $P_{\hat{\mu}(x^{i-1})}$ (see Section IV for details). Since the Bayesian predictive distribution does achieve the redundancy $(1/2)\ln n$

even if $P \notin \mathcal{M}$, this means that if $\mathcal{M}$ is the normal location family, then there does exist an "almost" in-model estimator (i.e. a slight modification of the ML estimator) that does achieve $(1/2)\ln n$ even if $P \notin \mathcal{M}$. Although this example does not extend straightforwardly to other exponential families, [1] conjectured that there should nevertheless be some general definition for "almost" in-model estimators that achieve $(k/2)\ln n$ redundancy even if $P \notin \mathcal{M}$. Here we show that this conjecture is true, at least if $k = 1$: we propose the *slightly squashed* ML estimator, a modification of the ML estimator that puts it slightly outside model $\mathcal{M}$, and in Theorem 2 we show that this estimator achieves $(1/2)\ln n$ redundancy even if $P \notin \mathcal{M}$. This result is important in practice since, in contrast to the Bayesian predictive distribution, the slightly squashed ML estimator is in general just as easy to compute as the ML estimator itself.

II. Notation and Definitions

Throughout this text we use nats rather than bits as units of information. A sequence of outcomes $z_1, \ldots, z_n$ is abbreviated to z^n. We write E_P as a shorthand for $E_{Z\sim P}$, the expectation of Z under distribution P. When we consider a sequence of n outcomes independently distributed $\sim P$, we use E_P even as a shorthand for the expectation of $(Z_1, \ldots, Z_n)$ under the n-fold product distribution of P. Finally, $P(Z)$ denotes the probability mass function of P in case Z is discrete-valued, and it denotes the density of P, in case Z takes its value in a continuum. When we write 'density function of Z', then, if Z is discrete-valued, this should be read as 'probability mass function of Z'. Note however that in our second main result, Theorem 2 we do not assume that the data-generating distribution P admits a density.

Let $\mathcal{Z}$ be a set of outcomes, taking values either in a finite or countable set, or in a subset of k-dimensional Euclidean space for some $k \geq 1$. Let $X : \mathcal{Z} \to \mathbb{R}$ be a random variable on $\mathcal{Z}$, and let $\mathcal{X} = \{x \in \mathbb{R} \ : \ \exists z \in \mathcal{Z} : X(z) = x\}$ be the range of X. Exponential family models are families of distributions on $\mathcal{Z}$ defined relative to a random variable X (called 'sufficient statistic') as defined above, and a function $h : \mathcal{Z} \to [0, \infty)$. Let $\mathrm{Z}(\eta) := \int_{z\in\mathcal{Z}} e^{-\eta X(z)} h(z) dz$ (the integral to be replaced by a sum for countable $\mathcal{Z}$), and $\Theta_{\text{nat}} := \{\eta \in \mathbb{R} : \mathrm{Z}(\eta) < \infty\}$.

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

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

The statistic $X(z)$ is sufficient for η [12]. This suggests reparameterizing the distribution by the expected value of X, which is called the *mean value parameterization*. The function $\mu(\eta) = E_{M_\eta}[X]$ maps parameters in the natural parameterization to the mean value parameterization. It is a diffeomorphism (it is one-to-one, onto, infinitely often differentiable and has an infinitely often differentiable inverse) [12]. Therefore the mean value parameter space Θ_{mean} is also an open interval of $\mathbb{R}$. We write $\mathcal{M} = \{M_\mu \mid \mu \in \Theta_{\text{mean}}\}$ where M_μ is the distribution with mean value parameter μ.

We are now ready to define the plug-in universal model. This is a distribution on infinite sequences $z_1, z_2, \ldots \in \mathcal{Z}^\infty$, recursively defined in terms of the distributions of Z_{n+1} conditioned on $Z^n = z^n$, for all $n = 1, 2, \ldots$. In the definition, we use the notation $x_i := X(z_i)$. Note that we use the term "model" both for a single distribution ("plug-in universal model", a common phrase in information theory) and for a family of distributions ("statistical model", a common phrase in statistics).

Definition 2 (Plug-in universal model): Let $\mathcal{M} = \{M_\mu \mid \mu \in \Theta_{\text{mean}}\}$ be an exponential family with mean value parameter domain Θ_{mean}. Given $\mathcal{M}$, constant $\bar{\mu}_0 \in \Theta_{\text{mean}}$ and a sequence of functions $\bar{\mu}(z^1), \bar{\mu}(z^2), \ldots$, such that $\bar{\mu}(z^n) =: \bar{\mu}_n \in \Theta_{\text{mean}}$, we define the *plug-in universal model* (or *plug-in model* for short) U by setting, for all n, all $z^{n+1} \in \mathcal{Z}^{n+1}$:

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

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

We usually refer to plug-in universal model in terms of the codelength function of the corresponding plug-in universal code:

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

The most important plug-in model is the ML (*maximum likelihood*) plug-in model, defined as follows:

Definition 3 (ML plug-in model): Given $\mathcal{M}$ and constants $x_0 \in \Theta_{\text{mean}}$ and $n_0 > 0$, we define the *ML plug-in model* $\hat{U}$ by setting, for all n, all $z^{n+1} \in \mathcal{Z}^{n+1}$:

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

where

$$\hat{\mu}(z^n) = \hat{\mu}_n := \frac{x_0 \cdot n_0 + \sum_{i=1}^n x_i}{n + n_0}. \quad (2)$$

To understand this definition, note that for exponential families, for any sequence of data, the ordinary maximum likelihood parameter is given by the average $n^{-1}\sum x_i$ of the observed values of X [12]. Here we define our plug-in model in terms of a slightly modified maximum likelihood estimator that introduces a 'fake initial outcome' x_0 with multiplicity n_0 in order to avoid infinite code lengths for the first few outcomes (a well-known problem sometimes called the "inherent singularity" of predictive coding [7], [1]) and to ensure that the plug-in ML code of the first outcome is well-defined. In practice we can take $n_0 = 1$ but our result holds for any $n_0 > 0$.

Definition 4 (Relative redundancy): Following [13], [8], we define *relative redundancy* with respect to P of a code U that is universal on a model $\mathcal{M}$, as:

$$\mathcal{R}_U(n) := E_P[L_U(Z^n)] - \inf_{\mu \in \Theta_{\text{mean}}} E_P[-\ln M_\mu(Z^n)], \quad (3)$$

where L_U is the length function of U.

We use the term *relative redundancy* rather than just *redundancy* to emphasize that it measures redundancy relative to the element of the model that minimizes the codelength rather than to P, which is not necessarily an element of the model. From now on, we only consider P under which the data are i.i.d. Under this condition, let M_{μ^*} be the element of $\mathcal{M}$ that minimizes KL divergence to P:

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

where the equality follows from the definition of KL divergence. If M_{μ^*} exists, it is unique, and if $E_P[X] \in \Theta_{\text{mean}}$, then $\mu^* = E_P[X]$ [1, Ch. 17], and the relative redundancy satisfies

$$\mathcal{R}_U(n) = E_P[L_U(Z^n)] - E_P[-\ln M_{\mu^*}(Z^n)]. \quad (4)$$

III. First Result: Redundancy of Plug-In Codes

The three major types of universal codes, Bayes, NML and 2-part, achieve relative redundancies that are (in an appropriate sense) close to optimal. Specifically, under the conditions on $\mathcal{M}$ described above, and if data are i.i.d. $\sim P$, then, under some mild conditions on P, these universal codes satisfy:

$$\mathcal{R}_U(n) = \frac{1}{2} \ln n + O(1), \quad (5)$$

(where the $O(1)$ may depend on μ and the universal code used), whenever $P \in \mathcal{M}$ or $P \notin \mathcal{M}$. (5) is the famous 'k over 2 log n formula' ($k = 1$ in our case), refinements of which lie at the basis of practical approximations to MDL learning [1].

While it is known that for $P \in \mathcal{M}$, the fourth major type of universal code, the ML plug-in code, satisfies (5) as well, it was shown by [8], [9] that when P is not in the model, the ML plug-in code may behave suboptimally. Specifically, its relative redundancy satisfies:

$$\mathcal{R}_{\hat{U}}(n) = \frac{1}{2} \frac{\text{var}_P X}{\text{var}_{M_{\mu^*}} X} \ln n + O(1), \quad (6)$$

and can be significantly larger than (5), when the variance of P is large.

In this paper, we show that not only the ML plug-in code, but *every* plug-in code may behave suboptimally, when $P \notin \mathcal{M}$. In other words, modifying the ML estimator $\hat{\mu}_n$ or introducing any other sequence of estimators $\bar{\mu}_n$, and constructing the plug-in code based on that sequence will not help to satisfy (5). Thus the optimal redundancy can only be achieved by codes outside $\mathcal{M}$, unless $\mathcal{M}$ is the Bernoulli family (since we assume the data are i.i.d., in the Bernoulli case we must have that $P \in \mathcal{M}$; but the Bernoulli case is the *only* case in which we must have $P \in \mathcal{M}$).

Our main result, Theorem 1, concerns the case in which P is itself a member of some exponential family $\mathcal{P}$, but $\mathcal{P}$ is in general different than $\mathcal{M}$. Then, the suboptimal behavior of plug-in codes follows immediately as Corollary 1, stated further below.

Theorem 1: Let $\mathcal{M} = \{M_\mu \mid \mu \in \Theta_{\text{mean}}\}$ and $\mathcal{P} = \{P_\mu \mid \mu \in \Theta_{\text{mean}}\}$ be single parameter exponential families with the same sufficient statistic X and mean-value parameter space Θ_{mean}. Let U denote any plug-in model with respect to $\mathcal{M}$ based on the sequence of estimators $\bar{\mu}_0, \bar{\mu}_1, \bar{\mu}_2, \ldots$. Then, for Lebesgue almost all $\mu^* \in \Theta_{\text{mean}}$ (i.e. all apart from a Lebesgue measure zero set), for $X, X_1, X_2, \ldots$ i.i.d. $\sim P_{\mu^*} \in \mathcal{P}$:

$$\liminf_{n \to \infty} \frac{\mathcal{R}_U(n)}{\frac{1}{2} \ln n} \geq \frac{\text{var}_{P_{\mu^*}} X}{\text{var}_{M_{\mu^*}} X}.$$

Proof: (rough sketch) The proof is based on a theorem stated by Rissanen [14] (see also [1], Theorem 14.2), a special case of which says the following. Let $\Theta_0 \subset \Theta_{\text{mean}}$ be a closed, non-degenerate interval, $\mathcal{P}$ be defined as above, $P_\mu^{(n)}$ be a joint distribution of n outcomes generated i.i.d. from P_μ, Q be an arbitrary probabilistic source, i.e. a distribution on infinite sequences $z_1, z_2, \ldots \in \mathcal{Z}^\infty$, and let $Q^{(n)}$ be its restriction to the first n outcomes. Define: $g_n(\mu^*) = \frac{D(P_{\mu^*}^{(n)} \| Q^{(n)})}{\frac{1}{2} \ln n}$. Then for Lebesgue almost all $\mu^* \in \Theta_0$, $\liminf_{n \to \infty} g_n(\mu^*) \geq 1$.

We apply Rissanen's theorem by constructing a source Q, specifying the conditional probabilities $Q(z_{n+1} | z^n) := P_{\bar{\mu}_n}$, for every $n \geq 1$. We now have:

$$\begin{aligned} D(P_{\mu^*}^{(n)} \| Q^{(n)}) &= \sum_{i=0}^{n-1} E_{P_{\mu^*}} \left[\ln P_{\mu^*}(Z_{i+1}) - \ln Q(Z_{i+1} | Z^i) \right] \\ &= \sum_{i=1}^{n-1} E_{P_{\mu^*}} \left[D(P_{\mu^*} \| P_{\bar{\mu}_i}) \right]. \end{aligned} \quad (7)$$

To see how (7) is related to our case, let us first rewrite the redundancy in a more convenient form:

$$\mathcal{R}_U(n) = \sum_{i=0}^{n-1} E_{P_{\mu^*}} \left[D(M_{\mu^*} \parallel M_{\bar{\mu}_i}) \right]. \quad (8)$$

The derivation of (8) make use of a standard result in the theory of exponential families and can be found e.g. in [1].

Comparing (7) and (8), we see that although in both expressions, the expectation is taken with respect to P_{μ^*}, (7) is a statement about KL divergence between the members of $\mathcal{P}$, while (8) speaks about the members of $\mathcal{M}$. The trick, which allows us to relate both expressions, is to examine their second-order behavior. By expanding $D(P_{\mu^*} \| P_{\bar{\mu}_i})$ into a Taylor series around μ^*, we get:

$$D(P_{\mu^*} \| P_{\bar{\mu}_i}) \simeq 0 + D^{(1)}(\mu^*)(\bar{\mu}_i - \mu^*) + \frac{1}{2} D^{(2)}(\mu^*)(\bar{\mu}_i - \mu^*)^2,$$

where we abbreviated $D^{(k)}(\mu) = \frac{d^k}{d\mu^k} D(P_{\mu^*} \| P_\mu)$. The term $D^{(1)}(\mu^*)$ is zero, since $D(\mu^* \| \mu)$ as a function of μ has its minimum at $\mu = \mu^*$ [12]. As is well-known [12], for exponential families the term $D^{(2)}(\mu)$ coincides precisely with the Fisher information $I_{\mathcal{P}}(\mu)$ evaluated at μ. Another standard

result [12] for the mean-value parameterization says that for all μ, $I_{\mathcal{P}}(\mu) = \frac{1}{\mathrm{var}_{P_\mu} X}$. Therefore, we get $D(P_{\mu^*} \| P_{\bar{\mu}_i}) \simeq \frac{1}{2} \frac{(\bar{\mu}_i - \mu^*)^2}{\mathrm{var}_{P_{\mu^*}} X}$, and similarly, $D(M_{\mu^*} \| M_{\bar{\mu}_i}) \simeq \frac{1}{2} \frac{(\bar{\mu}_i - \mu^*)^2}{\mathrm{var}_{M_{\mu^*}} X}$, so that $D(M_{\mu^*} \| M_{\bar{\mu}_i}) \simeq D(P_{\mu^*} \| P_{\bar{\mu}_i}) \frac{\mathrm{var}_{P_{\mu^*}} X}{\mathrm{var}_{M_{\mu^*}} X}$, and using (7) and (8):

$$\mathcal{R}_U(n) \simeq D(P^{(n)}_{\mu^*} \| Q^{(n)}) \frac{\mathrm{var}_{P_{\mu^*}} X}{\mathrm{var}_{M_{\mu^*}} X}.$$

The last step of the proof is to use Rissanen's theorem and conclude that $\liminf_{n\to\infty} \frac{\mathcal{R}_U(n)}{\frac{1}{2}\ln n}$ is equal to

$$\liminf_{n\to\infty} \frac{D(P^{(n)}_{\mu^*} \| Q^{(n)})}{\frac{1}{2}\ln n} \frac{\mathrm{var}_{P_{\mu^*}} X}{\mathrm{var}_{M_{\mu^*}} X} \geq \frac{\mathrm{var}_{P_{\mu^*}} X}{\mathrm{var}_{M_{\mu^*}} X},$$

for Lebesgue almost all $\mu^* \in \Theta_0$, and thus for Lebesgue almost all $\mu^* \in \Theta_{\text{mean}}$. ■

We now use Theorem 1 to show that the redundancy of plug-in codes is suboptimal for all exponential families which satisfy the following very weak condition:

Condition 1: Let $\mathcal{M} = \{M_\mu \mid \mu \in \Theta_{\text{mean}}\}$ be a single parameter exponential family with sufficient statistic X and mean-value parameter space Θ_{mean}. We require that there exists another single-parameter exponential family $\mathcal{P} = \{P_\mu \mid \mu \in \Theta_{\text{mean}}\}$ with the same mean-value parameter space as $\mathcal{M}$, but with strictly larger variance than $\mathcal{M}$ for every $\mu \in \Theta_{\text{mean}}$.

The Condition 1 is widely satisfied among known exponential families. When $\mathcal{X} = [a, b]$, we define P_μ to be a "scaled" Bernoulli model, by putting all probability mass on $\{a, b\}$ in such a way that $E_{P_\mu} = \mu$. It is easy to show, that such distribution has the highest variance among all distributions defined on $[a, b]$ with a given mean value μ; therefore $\mathrm{var}_{P_\mu} X > \mathrm{var}_{M_\mu} X$, unless $\mathcal{M}$ is a "scaled" Bernoulli itself. When $\mathcal{X} = \mathbb{R}$, $\mathcal{P}$ can be chosen to be a normal family with fixed, sufficiently large variance σ^2. For $X = [0, \infty)$, $\mathcal{P}$ can be taken to be a gamma family with sufficiently large scale parameter. When $\mathcal{X} = \{0, 1, 2, \ldots\}$, $\mathcal{P}$ can be taken to be negative binomial (with expected "number of successes" sufficiently small).

Thus, we see that for all commonly used exponential families, except for Bernoulli, Condition 1 holds. On the other hand if $\mathcal{M}$ is Bernoulli, Corollary 1 is no longer relevant anyway, since then P must lie in $\mathcal{M}$.

Corollary 1: Let $\mathcal{M} = \{M_\mu \mid \mu \in \Theta_{\text{mean}}\}$ a single parameter exponential family with sufficient statistic X and mean-value parameter space Θ_{mean}, satisfying Condition 1. Let U denote any plug-in model with respect to $\mathcal{M}$ based on any sequence of estimators $\bar{\mu}_1, \bar{\mu}_2, \ldots$. Then, there exists a family of distributions $\mathcal{P} = \{P_\mu \mid \mu \in \Theta_{\text{mean}}\}$, such that for Lebesgue almost all $\mu^* \in \Theta_{\text{mean}}$, for $X, X_1, X_2, \ldots$ i.i.d. $\sim P_{\mu^*}$:

$$\liminf_{n\to\infty} \frac{\mathcal{R}_U(n)}{\ln n} \geq \frac{1}{2} \frac{\mathrm{var}_{P_{\mu^*}} X}{\mathrm{var}_{M_{\mu^*}} X} > \frac{1}{2},$$

so that the set of μ^* for which U achieves the regret $\frac{1}{2}\ln n + O(1)$ is a set of Lebesgue measure zero.

Proof: Immediate from Theorem 1. ■

IV. Second Result: Optimality of Squashed ML

We showed that every plug-in code, including the ML plug-in code, behaves suboptimally for 1-parameter families $\mathcal{M}$ unless $\mathcal{M}$ is Bernoulli. This fact does not, however, exclude the possibility that a small modification of the ML plug-in code, which puts the predictions slightly outside $\mathcal{M}$, will lead to the optimal redundancy (5). An argument supporting this claim comes from considering the Bayesian predictive distribution when $\mathcal{M}$ is the normal family with fixed variance σ^2. In this case, the Bayesian code based on prior $\mathcal{N}(\mu_0, \tau_0^2)$ has a simple form [1]:

$$U_{\text{Bayes}}(z_{n+1} \mid z^n) = f_{\mu_n, \tau_n^2 + \sigma^2}(z_{n+1}),$$

where f_{μ,σ^2} is the density of normal distribution $\mathcal{N}(\mu, \sigma^2)$,

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

Thus, the Bayesian predictive distribution is itself a Gaussian with mean equal to the modified maximum likelihood estimator (with $n_0 = \sigma^2/\tau_0^2$), 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" in-model code, which satisfies (5). This led [1] to conjecture that something similar holds for general exponential families. Here we show that this is indeed the case: we propose a simple modification of the ML plug-in universal model, obtained by predicting z_{n+1} using a slightly "squashed" version $M'_{\hat{\mu}_n}$ of the ML estimator $M_{\hat{\mu}_n}$, defined as:

$$M'_{\hat{\mu}_n}(z_{n+1}) := M_{\hat{\mu}_n}(z_{n+1}) \frac{1 + \frac{1}{2n} I_{\mathcal{M}}(\hat{\mu}_n)(x_{n+1} - \hat{\mu}_n)^2}{1 + \frac{1}{2n}},$$

where $\hat{\mu}_n$ is defined as in (2) and $I_{\mathcal{M}}(\mu)$ is the Fisher information for model $\mathcal{M}$. Note that $M'_{\hat{\mu}_n}(z_{n+1})(\cdot)$ represents a valid probability density: it is non-negative due to $I_{\mathcal{M}}(\hat{\mu}_n) > 0$ (property of exponential families), and it is properly normalized:

$$\int_{\mathcal{X}} M'_{\hat{\mu}_n}(z_{n+1})(z)dz = \left(1 + \tfrac{1}{2n}\right)^{-1} \Bigg(\int_{\mathcal{X}} M_{\hat{\mu}_n}(z)dz + \tfrac{1}{2n} I_{\mathcal{M}}(\hat{\mu}_n) \int_{\mathcal{X}} (X(z) - \hat{\mu}_n)^2 M_{\hat{\mu}_n}(z)dz \Bigg) = 1,$$

where the final equality follows because for exponential families, $I_{\mathcal{M}}(\mu) = (\mathrm{var}_{M_\mu} X)^{-1}$. While $M' \notin \mathcal{M}$, we have $D(M'_{\hat{\mu}_n} \| M_{\hat{\mu}_n}) = O(1/n)$, i.e. M' is "almost" in-model estimator.

Definition 5 (Squashed ML prequential model): Given $\mathcal{M}$, constants $x_0 \in \Theta_{\text{mean}}$ and $n_0 > 0$, we define the *slightly squashed ML prequential model* U by setting, for all n, all $z^{n+1} \in \mathcal{Z}^{n+1}$:

$$U(z_{n+1} \mid z^n) = M'_{\hat{\mu}_n}(z_{n+1}),$$

where M' is the slightly squashed ML estimator as above. The codelengths of the corresponding slightly squashed ML prequential code are not harder to calculate than those of the ordinary ML plug-in model and in some cases they are easier

to calculate than the lengths of the Bayesian universal code. On the other hand, we show below that the slightly squashed ML code always achieves the optimal redundancy, satisfying (5).

Theorem 2: 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 slightly squashed ML model with respect to $\mathcal{M}$. If $\mathcal{M}$ and P satisfy Condition 2 below, then:

$$\mathcal{R}_U(n) = \frac{1}{2} \ln n + O(1). \tag{9}$$

Condition 2: 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, that $\frac{d^2}{d\mu^2} I_{\mathcal{M}}(\mu) = O\left(\mu^{k-4}\right)$, $\frac{d^4}{d\mu^4} D(M_{\mu^*} \| M_\mu) = O(\mu^{k-6})$ and that either $I_{\mathcal{M}}(\mu)$ is constant or $I_{\mathcal{M}}(\mu) = O\left(\mu^{k/2-3}\right)$.
- If T is bounded from above by a constant g then $\frac{d^2}{d\mu^2} I_{\mathcal{M}}(\mu)$, $\frac{d^4}{d\mu^4} D(M_{\mu^*} \| M_\mu)$, and $I_{\mathcal{M}}(\mu)$ are polynomial in $1/(g - \mu)$.

The usefulness of Theorem 2 depends on the validity of Condition 2 among commonly used exponential families. As can be seen from Figure 1, for some standard exponential families, our condition applies whenever the fourth moment of P exists. *Proof: (of Theorem 2; rough sketch)* We express the relative redundancy of the slightly squashed ML plug-in code U by the sum of the relative redundancy of the ordinary ML plug-in code $\hat{U}$ and the difference in expected codelengths between U and $\hat{U}$:

$$\begin{aligned}
&\mathcal{R}_U(n) = E_P[L_U(Z^n)] - E_P[-\ln M_{\mu^*}(Z^n)] = \\
&E_P[L_U(Z^n) - L_{\hat{U}}(Z^n)] + \mathcal{R}_{\hat{U}}(n) = \\
&E_P[L_U(Z^n) - L_{\hat{U}}(Z^n)] + \tfrac{1}{2} \tfrac{\mathrm{var}_P X}{\mathrm{var}_{M_{\mu^*}} X} \ln n + O(1),
\end{aligned}$$

where the last equality follows from (6). We have:

$$\begin{aligned}
&L_U(Z^n) - L_{\hat{U}}(Z^n) = \\
&\textstyle\sum_{i=0}^{n-1} \left(-\ln U(Z_{i+1} \mid Z_i) + \ln \hat{U}(Z_{i+1} \mid Z_i) \right) = \\
&\textstyle\sum_{i=0}^{n-1} \left(\ln\left(1 + \frac{1}{2i}\right) - \ln\left(1 + \frac{1}{2i} I_{\mathcal{M}}(\hat{\mu}_i)(X_{i+1} - \hat{\mu}_i)^2\right) \right).
\end{aligned}$$

Since $\ln\left(1 + \frac{1}{2i}\right) = \frac{1}{2i} + O(i^{-2})$, we get $\sum_{i=0}^{n-1} \ln\left(1 + \frac{1}{2i}\right) = \frac{1}{2} \ln n + O(1)$. Denoting $V_i = \frac{1}{2i} I_{\mathcal{M}}(\hat{\mu}_i)(X_{i+1} - \hat{\mu}_i)^2$, we also get $\ln(1 + V_i) = V_i + O(i^{-2})$. Next, we consider $E_P[V_i]$:

$$\begin{aligned}
&E_P[V_i] = \tfrac{1}{2i} E_P\left[I_{\mathcal{M}}(\hat{\mu}_i)(X_{i+1} - \mu^* + \mu^* - \hat{\mu}_i)^2 \right] = \\
&\tfrac{1}{2i} E_P\left[I_{\mathcal{M}}(\hat{\mu}_i) \left(\mathrm{var}_{P_{\mu^*}} X + (\mu^* - \hat{\mu}_i)^2 \right) \right] = \\
&\tfrac{1}{2i} \left(\mathrm{var}_{P_{\mu^*}} X E_P\left[I_{\mathcal{M}}(\hat{\mu}_i) \right] + E_P\left[I_{\mathcal{M}}(\hat{\mu}_i)(\mu^* - \hat{\mu}_i)^2 \right] \right).
\end{aligned}$$

The second term $E_P\left[I_{\mathcal{M}}(\hat{\mu}_i)(\mu^* - \hat{\mu}_i)^2 \right]$ is $O(i^{-1})$ as $E_P[(\mu^* - \hat{\mu}_i)^2] = O(i^{-1})$ and $E[I_{\mathcal{M}}(\hat{\mu}_i)] = I_{\mathcal{M}}(\mu^*) + O(i^{-1})$ (follows from expanding $I_{\mathcal{M}}(\hat{\mu}_i)$ up to the first order around μ^*). Similarly, the first term is $(\mathrm{var}_{P_{\mu^*}} X) I_{\mathcal{M}}(\mu^*) + O(i^{-1})$. Thus, using $I_{\mathcal{M}}(\mu^*) = \frac{1}{\mathrm{var}_{M_{\mu^*}} X}$, we finally get:

$$E_P[-\ln(1+V_i)] = -E_P[V_i] + O(i^{-2}) = \frac{1}{2i} \frac{\mathrm{var}_{P_{\mu^*}} X}{\mathrm{var}_{M_{\mu^*}} X} + O(i^{-2}).$$

Fig. 1. Fisher information, its second derivative and a fourth derivative of the divergence for a number of exponential families. For the normal distribution with fixed mean we use mean 0 and the density of the squared outcomes is given as a function of the variance.

Distribution	$I(\mu)$	$\frac{d^2}{d\mu^2} I(\mu)$	$\frac{d^4}{d\mu^4} D(M_{\mu^*} \| M_\mu)$
Bernoulli	$\frac{1}{\mu(1-\mu)}$	$\frac{2}{\mu^3} + \frac{2}{(1-\mu)^3}$	$\frac{6\mu^*}{\mu^4} + \frac{6(1-\mu^*)}{(1-\mu)^4}$
Poisson	$\frac{1}{\mu}$	$\frac{2}{\mu^3}$	$\frac{6\mu^*}{\mu^4}$
Geometric	$\frac{1}{\mu(\mu-1)}$	$-\frac{2}{\mu^3} + \frac{2}{(1-\mu)^3}$	$\frac{6\mu^*}{\mu^4} - \frac{6(\mu^*+1)}{(\mu+1)^4}$
Gamma (fixed k)	$\frac{k}{\mu^2}$	$\frac{6k}{\mu^4}$	$-\frac{6k}{\mu^4} + \frac{24k\mu^*}{\mu^5}$
Normal (fixed mean)	$\frac{1}{2\mu^2}$	$\frac{3}{\mu^4}$	$-\frac{3}{\mu^4} + \frac{12\mu^*}{\mu^5}$
Normal (fixed variance)	σ^2	0	0

Taking all together, we see that the terms $\frac{\mathrm{var}_{P_{\mu^*}} X}{\mathrm{var}_{M_{\mu^*}} X}$ cancel and we finally get $R_U(n) = \frac{1}{2} \ln n + O(1)$. Condition 2 is necessary to ensure that all Taylor expansions above hold. ■

V. Future Work

In future work, we hope to extend our results concerning the slightly squashed ML estimator to the multi-parameter case and establish almost-sure variation of Theorem 2. We also plan to analyze the estimator in the individual sequence framework, along the lines of [15], [16].

References

[1] P. Grünwald, *The Minimum Description Length Principle*. Cambridge, MA: MIT Press, 2007.

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

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

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

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

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

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

[8] S. De Rooij and P. D. Grünwald, "MDL model selection using the ML plug-in code," in *Proceedings of the 2005 IEEE International Symposium on Information Theory (ISIT 2005)*, Adelaide, Australia, 2005.

[9] P. D. Grünwald and S. de Rooij, "Asymptotic log-loss of prequential maximum likelihood codes," in *Proc. of the 18th Annual Conference on Computational Learning Theory (COLT 2005)*, 2005, pp. 652–667.

[10] S. De Rooij and P. D. Grünwald, "An empirical study of MDL model selection with infinite parametric complexity," *Journal of Mathematical Psychology*, vol. 50, no. 2, pp. 180–192, 2006.

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

[12] O. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory*. Chichester, UK: Wiley, 1978.

[13] J. Takeuchi and A. R. Barron, "Robustly minimax codes for universal data compression," in *Proceedings of the Twenty-First Symposium on Information Theory and Its Applications (SITA '98)*, Gifu, Japan, 1998.

[14] J. Rissanen, "Stochastic complexity and modeling," *The Annals of Statistics*, vol. 14, pp. 1080–1100, 1986.

[15] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning and Games*. Cambridge University Press, 2006.

[16] M. Raginsky, R. F. Marcia, S. Jorge, and R. Willett, "Sequential probability assignment via online onvex programming using exponential families," in *Proceedings of the 2009 IEEE International Symposium on Information Theory*, Seoul, Korea, 2009.

APPENDIX

PROOF OF THEOREM 1

Before we show the main result, we need to prove the following lemmas.

Lemma 3: Let $\mathcal{M} = \{M_\mu \mid \mu \in \Theta_{\text{mean}}\}$ and $\mathcal{P} = \{P_\mu \mid \mu \in \Theta_{\text{mean}}\}$ be single parameter exponential families with the same sufficient statistic X and mean-value parameter space Θ_{mean}. Let $\Theta_0 \subset \Theta_{\text{mean}}$ be any non-degenerate closed interval. Let $X, X_1, X_2, \ldots$ be i.i.d. $\sim P_{\mu^*}$ for some $\mu^* \in \Theta_0$. Let $\bar{\mu}_0, \bar{\mu}_1, \bar{\mu}_2, \ldots$ be a sequence of estimators, such that $\bar{\mu}_i = \bar{\mu}_i(z^i)$ and $\bar{\mu}_i \in \Theta_0$ for all $i \geq 1$. Then, for Lebesgue almost all $\mu^* \in \Theta_0$:

$$\liminf_{n\to\infty} \frac{\sum_{i=0}^{n-1} E_{P_{\mu^*}}[(\bar{\mu}_i - \mu^*)^2]}{\ln n} \geq \underline{V}_{\mathcal{P}}$$

where $\underline{V}_{\mathcal{P}} := \inf_{\mu\in\Theta_0} \text{var}_{P_\mu} X$.

Proof: The proof is based on a theorem stated by Rissanen [14] (see also [1], Theorem 14.2), a special case of which says the following.

Let $\mathcal{P}$ and Θ_0 be defined as above, $P_\mu^{(n)}$ be a joint distribution of n outcomes generated i.i.d. from P_μ, Q be an arbitrary probabilistic source, i.e. a distribution on infinite sequences $z_1, z_2, \ldots \in \mathcal{Z}^\infty$, and let $Q^{(n)}$ be its restriction to the first n outcomes (marginalized over $z_{n+1}, z_{n+2}, \ldots$). Define:

$$g_n(\mu^*) = \inf_{n'\geq n} \left\{ \frac{D(P_{\mu^*}^{(n')} \| Q^{(n')})}{\frac{1}{2}\ln n'} \right\}. \tag{10}$$

Then for Lebesgue almost all $\mu^* \in \Theta_0$, $\lim_{n\to\infty} g_n(\mu^*) \geq 1$.

We construct the source Q by specifying the conditional probabilities:

$$Q(z_{n+1}|z^n) := P_{\bar{\mu}_n},$$

for every $n \geq 1$. This definition is valid, because $\bar{\mu}_n$ depends only on z^n. Now, we have:

$$\begin{aligned} D(P_{\mu^*}^{(n)} \| Q^{(n)}) &= E_{Z^n \sim P_{\mu^*}^{(n)}}[\ln P_{\mu^*}(Z^n) - \ln Q(Z^n)] \\ &= \sum_{i=0}^{n-1} E_{Z^i \sim P_{\mu^*}^{(i)}} \left[\ln P_{\mu^*}(Z_{i+1}) - \ln Q(Z_{i+1}|Z^i)\right] \\ &= \sum_{i=1}^{n-1} E_{Z^i \sim P_{\mu^*}^{(i)}} \left[D(P_{\mu^*} \| P_{\bar{\mu}_i})\right]. \end{aligned}$$

Expanding $D(P_{\mu^*} \| P_{\bar{\mu}_i})$ into a Taylor series around μ^* yeilds:

$$D(P_{\mu^*} \| P_{\bar{\mu}_i}) = 0 + D^{(1)}(\mu^*)(\bar{\mu}_i - \mu^*) + \frac{1}{2} D^{(2)}(\mu)(\bar{\mu}_i - \mu^*)^2,$$

for some μ between $\bar{\mu}_i$ and μ^*, where we abbreviated $D^{(k)}(\mu) = \frac{d^k}{d\mu^k} D(P_{\mu^*} \| P_\mu)$. The term $D^{(1)}(\mu^*)$ is zero, since $D(\mu^* \| \mu)$ as a function of μ has its minimum at $\mu = \mu^*$ [12]. As is well-known [12], for exponential families the term $D^{(2)}(\mu)$ coincides precisely with the Fisher information $I_{\mathcal{P}}(\mu)$ evaluated at μ. Another standard result [12] for the mean-value parameterization says that for all μ,

$$I_{\mathcal{P}}(\mu) = \frac{1}{\text{var}_{P_\mu} X}. \tag{11}$$

Therefore (using shorter notation $E_{P_{\mu^*}}$ for $E_{Z^i \sim P_{\mu^*}^{(i)}}$):

$$\begin{aligned} D(P_{\mu^*}^{(n)} \| Q^{(n)}) &= \tfrac{1}{2} \textstyle\sum_{i=0}^{n-1} E_{P_{\mu^*}} \left[\frac{(\bar{\mu}_i - \mu^*)^2}{\text{var}_{P_\mu} X}\right] \\ &\leq \tfrac{1}{2} \tfrac{1}{\underline{V}_{\mathcal{P}}} \textstyle\sum_{i=0}^{n-1} E_{P_{\mu^*}} \left[(\bar{\mu}_i - \mu^*)^2\right]. \end{aligned} \tag{12}$$

Note, that $\underline{V}_{\mathcal{P}} > 0$ is an infimum of a continuous and positive function on a compact set. From (10) and (12) we have:

$$\inf_{n'\geq n} \left\{ \frac{\frac{1}{2}\sum_{i=0}^{n'-1} E_{P_{\mu^*}}\left[(\bar{\mu}_i - \mu^*)^2\right]}{\frac{1}{2}\ln n'} \right\} \geq g_n(\mu^*) \underline{V}_{\mathcal{P}},$$

and thus Rissanen's theorem proves the lemma. ■

Lemma 4: Let $\mathcal{M}, \mathcal{P}, \Theta_0, X, X_1, X_2, \ldots$, be defined as in Lemma 3. Let U denote any plug-in model with respect to $\mathcal{M}$ based on the sequence of estimators $\bar{\mu}_1, \bar{\mu}_2, \ldots$. Then, for Lebesgue almost all $\mu^* \in \Theta_0$:

$$\liminf_{n\to\infty} \frac{\sum_{i=0}^{n-1} E_{P_{\mu^*}}[D(M_{\mu^*} \| M_{\bar{\mu}_i})]}{\ln n} \geq \frac{1}{2} \frac{\underline{V}_{\mathcal{P}}}{\overline{V}_{\mathcal{M}}},$$

for $\underline{V}_{\mathcal{P}} := \inf_{\mu\in\Theta_0} \text{var}_{P_\mu} X$ and $\overline{V}_{\mathcal{M}} := \sup_{\mu\in\Theta_0} \text{var}_{M_\mu} X$.

Proof: Notice, that now we do not restrict $\bar{\mu}_i$ to be in Θ_0, as in Lemma 3. Let us denote $\Theta_0 = [\mu_0, \mu_1]$. We define a truncated sequence of estimators $(\bar{\mu}_i')$ as follows:

$$\bar{\mu}_i' = \begin{cases} \mu_1 & \text{if } \bar{\mu}_i \geq \mu_1 \\ \bar{\mu}_i & \text{if } \mu_0 < \bar{\mu}_i < \mu_1 \\ \mu_0 & \text{if } \bar{\mu}_i \leq \mu_0 \end{cases},$$

so that $\bar{\mu}_i' \in \Theta_0$. Note, that $D(M_{\mu^*} \| M_{\bar{\mu}_i}) \geq D(M_{\mu^*} \| M_{\bar{\mu}_i'})$, as there exists $\lambda \in [0,1]$ such that we can express $\bar{\mu}_i' = \lambda\mu^* + (1-\lambda)\bar{\mu}_i$ and $D(M_{\mu^*} \| M_{\lambda\mu^* + (1-\lambda)\bar{\mu}_i})$ is strictly decreasing in λ [1]. Using this fact and expanding $D(M_{\mu^*} \| M_{\bar{\mu}_i'})$ into Taylor series as in Lemma 3, we get:

$$\begin{aligned} & E_{P_{\mu^*}}[D(M_{\mu^*} \| M_{\bar{\mu}_i})] \geq E_{P_{\mu^*}}[D(M_{\mu^*} \| M_{\bar{\mu}_i}')] \\ = \; & \frac{1}{2} E_{P_{\mu^*}} \left[\frac{(\bar{\mu}_i - \mu^*)^2}{\text{var}_{M_\mu} X}\right] \geq \frac{1}{2} \frac{1}{\overline{V}_{\mathcal{M}}} E_{P_{\mu^*}} \left[(\bar{\mu}_i - \mu^*)^2\right]. \end{aligned}$$

Summing over $i = 0, \ldots, n-1$ and using Lemma 3 finishes the proof. ■

Before we prove Theorem 1, we further need a simple lemma to rewrite the redundancy in a more convenient form:

Lemma 5: Let U and $\mathcal{M}$ be defined as in Theorem 1. We have:

$$\mathcal{R}_U(n) = \sum_{i=0}^{n-1} E_{P_{\mu^*}} \left[D(M_{\mu^*} \parallel M_{\bar{\mu}_i})\right].$$

The usefulness of this lemma comes from the fact that the KL divergence $D(\cdot\|\cdot)$ is defined as an expectation over M_{μ^*} rather than $P_{\mu*}$. The proof makes use of a standard result in the theory of exponential families and can be found e.g. in [1] (see also related Lemma 1 in [9]).

Proof: (of Theorem 1) Choose any $\mu^* \in \Theta$ and span around it a non-degenerate closed interval $\Theta_{\mu^*}' \subset \Theta_{\text{mean}}$, so that $\mu^* \in \text{int}\Theta_{\mu^*}'$. Fix some $\epsilon > 0$. It follows from general properties of exponential families (see, e.g., [12]) that $\text{var}_{M_\mu} X$ and $\text{var}_{P_\mu} X$ are continuous (with respect to μ), therefore

if we choose the interval Θ'_{μ^*} small enough, we will have $\frac{\underline{V}_{\mathcal{P}}}{\overline{V}_{\mathcal{M}}} > \frac{\mathrm{var}_{P_{\mu^*}} X}{\mathrm{var}_{M_{\mu^*}} X} - \epsilon$, with $\underline{V}_{\mathcal{P}} := \inf_{\mu \in \Theta'_{\mu^*}} \mathrm{var}_{P_\mu} X$ and $\overline{V}_{\mathcal{M}} := \sup_{\mu \in \Theta'_{\mu^*}} \mathrm{var}_{M_\mu} X$. Using Lemma 4 with $\Theta_0 = \Theta'_{\mu^*}$, and Lemma 5, we have for Lebesgue almost all $\mu \in \Theta'_{\mu^*}$.

$$\liminf_{n\to\infty} \frac{\mathcal{R}_U(n)}{\ln n} \geq \frac{1}{2}\frac{\underline{V}_{\mathcal{P}}}{\overline{V}_{\mathcal{M}}} > \frac{1}{2}\left(\frac{\mathrm{var}_{P_{\mu^*}} X}{\mathrm{var}_{M_{\mu^*}} X} - \epsilon\right).$$

Note, that w.l.o.g. Θ'_{μ^*} can be chosen to have rational ends. The family of all intervals $\Theta'_{\mu^*} \subset \Theta_{\text{mean}}$ with rational ends and rational μ^*, i.e. $\Xi = \{\Theta'_{\mu^*} = [\mu_0, \mu_1] \mid \mu^*, \mu_0, \mu_1 \in \Theta_{\text{mean}} \cap \mathbb{Q}\}$, is countable and covers Θ_{mean}, $\bigcup_{\Theta'_{\mu^*} \in \Xi} \Theta'_{\mu^*} = \Theta_{\text{mean}}$. Therefore, $\liminf_{n\to\infty} \frac{\mathcal{R}_U(n)}{\ln n} > \frac{\mathrm{var}_{P_{\mu^*}} X}{\mathrm{var}_{M_{\mu^*}} X} - \epsilon$ holds for Lebesgue almost all $\mu^* \in \Theta_{\text{mean}}$.

Since it holds for every $\epsilon > 0$, this also means that $\liminf_{n\to\infty} \frac{\mathcal{R}_U(n)}{\ln n} \geq \frac{\mathrm{var}_{P_{\mu^*}} X}{\mathrm{var}_{M_{\mu^*}} X}$ for Lebesgue almost all $\mu^* \in \Theta_{\text{mean}}$. To show this, assume the contrary, that the set $A = \left\{\mu^* \colon \liminf_{n\to\infty} \frac{\mathcal{R}_U(n)}{\ln n} < \frac{\mathrm{var}_{P_{\mu^*}} X}{\mathrm{var}_{M_{\mu^*}} X}\right\}$ has positive Lebesgue measure, $L(A) > 0$. Let $\epsilon_1, \epsilon_2, \ldots$ be any sequence of positive numbers converging to 0 and let us define $A_i = \left\{\mu^* \colon \liminf_{n\to\infty} \frac{\mathcal{R}_U(n)}{\ln n} < \frac{\mathrm{var}_{P_{\mu^*}} X}{\mathrm{var}_{M_{\mu^*}} X} - \epsilon_i\right\}$. Obviously, $A_1 \subset A_2 \subset \ldots$, and $\bigcup_i A_i = A$. From continuity of measure, we must have $L(A_i) > 0$ for i large enough, which is a contradiction.

Since this holds for every $\epsilon > 0$, the thesis follows. ■

Proof of Theorem 2

We will make use of the following two theorems, proofs of which can be found in [9].

Theorem 6: Let $X, X_1, \ldots$ be i.i.d., let $\hat{\mu}_n := (n_0 \cdot x_0 + \sum_{i=1}^n X_i)/(n + n_0)$ and $\mu^* = E[X]$. If the first k moments of X exist, then $E[(\hat{\mu}_n - \mu^*)^k] = O(n^{-\lceil \frac{k}{2} \rceil})$.

Theorem 7: Let $X, X_1, \ldots$ be i.i.d. random variables, define $\hat{\mu}_n := (n_0 \cdot x_0 + \sum_{i=1}^n X_i)/(n + n_0)$ and $\mu^* = E[X]$. Let $k \in \{0, 2, 4, \ldots\}$. If the first k moments exists then $P(|\hat{\mu}_n - \mu^*| \geq \delta) = O\left(n^{-\lceil \frac{k}{2} \rceil} \delta^{-k}\right)$.

Before we prove the main theorem, we need the following lemma:

Lemma 8: Fix any $s \in \{0, 2, 4\}$. Let $f(\mu)$ be some continuous function of μ. Suppose it holds for both $T := X$ and $T := -X$ that:

- 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 $f(\mu) = O(\mu^{k-s-2})$.
- If T is bounded from above by a constant g then $f(\mu)$ is polynomial in $1/(g - \mu)$.

Then the expression $E_P[f(\mu)(\hat{\mu}_i - \mu^*)^s]$, for μ between μ^* and $\hat{\mu}_i$, is of order $O(i^{-s/2})$.

Proof: The proof follows very closely part of the proof of Lemma 2 in [9]; we nevertheless give here a complete proof for the sake of clarity.

Let us denote $\delta_i := \hat{\mu}_i - \mu^*$. We 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 want to analyze asymptotic behavior of:

$$E_P\left[f(\mu){\delta_i}^s\right] = \sum_j P(\delta_i \in \Delta_j) E_P\left[f(\mu){\delta_i}^s \mid \delta_i \in \Delta_j\right].$$

If we can establish the proper asymptotic behavior $O(i^{-s/2})$ for all regions Δ_j for $j \geq 0$, then we can use a symmetrical argument to establish the behavior for Δ_- as well, so it suffices if we restrict ourselves to $j \geq 0$. First we show it for Δ_0. In this case, the basic idea is that since the remainder $f(\mu)$ is well-defined over the interval $\mu^* \leq \mu < \mu^* + a$, we can bound it by its extremum on that interval, namely $m := \sup_{\mu \in [\mu^*, \mu^* + a)} |f(\mu)|$. Now we get:

$$\left|P(\delta_i \in \Delta_0) E\left[f(\mu){\delta_i}^s \mid \delta_i \in \Delta_0\right]\right| \;\leq\; 1 \cdot E\left[{\delta_i}^s \left|f(\mu)\right|\right],$$

which is less or equal than $mE\left[{\delta_i}^s\right]$. Using Theorem 6 we find that $E[{\delta_i}^s]$ is $O(i^{-s/2})$, which is what we want. Theorem 6 requires that the first four moments of P exist, but this is guaranteed to be the case: either the outcomes are bounded from both sides, in which case all moments necessarily exist, or the existence of the required moments is part of the condition on the main theorem.

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

$$\sum_{j=1}^{\infty} P(\delta_i \in \Delta_j) E\left[f(\mu){\delta_i}^s \mid \delta_i \in \Delta_j\right] = O(i^{-s/2}) \qquad (13)$$

We bound this expression from above. The δ_i in the expectation is at most $a + j$. Furthermore $f(\mu) = O(\mu^{k-s-2})$ by assumption, where $\mu \in [a + j - 1, a + j)$. Depending on k and s, both boundaries could maximize this function, but it is easy to check that in both cases the resulting function is $O(j^{k-s-2})$. So we bound (13) from the above by:

$$\sum_{j=1}^{\infty} P(|\delta_i| \geq a + j - 1)(a + j)^s O(j^{k-s-2}).$$

Since we know from the condition on the main theorem that the first $k \geq 4$ moments exist, we can apply Theorem 7 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 $O(i^{-\frac{k}{2}}) \sum_j O(j^{-2})$, which is of order $O(i^{-s/2})$, since the sum $\sum_j O(j^{-2})$ converges and $k \geq s$.

Now we consider the case where the outcomes are bounded from above by g. This case is more complicated, since now we have made no extra assumptions as to existence of the moments of P. Of course, if the outcomes are bounded from both sides, then all moments necessarily exist, but if the outcomes are unbounded from below this may not be true. To remedy this, we map all outcomes into a new domain in such a way that all moments of the transformed variables

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

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

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 7 is safe. It reveals that the summed probability is $O(i^{-\frac{k}{2}})$ for any even $k \in \mathbb{N}$. Now we bound $f(\mu)$ which is $O((g-\mu)^{-m})$ for some $m \in \mathbb{N}$ by the condition on the main theorem. Here we use that $\mu \leq \hat{\mu}_i$; the latter is maximized if all outcomes equal the bound g, in which case the estimator equals $g - n_0(g - x_0)/(i + n_0) = g - O(i^{-1})$. Putting all of this together, we get $\sup |f(\mu)| = O((g-\mu)^{-m}) = O(i^m)$; if we plug this into the equation we obtain:

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

This is of order $O(i^{-s/2})$ if we choose $k \geq 6m + s$. We can do this because the construction of $g(\cdot)$ ensures that all moments exist, and therefore certainly the first $6m + s$. ■

We can now proceed to prove the theorem:

Proof: (of Theorem 2) We express the relative redundancy of the squashed ML plug-in code U by the sum of the relative redundancy of the ordinary ML plug-in code $\hat{U}$ and the difference in expected codelengths between U and $\hat{U}$:

$$\begin{aligned} & \mathcal{R}_U(n) = E_P[L_U(Z^n)] - E_P[-\ln M_{\mu^*}(Z^n)]. \\ =\ & E_P[L_U(Z^n) - L_{\hat{U}}(Z^n)] + \mathcal{R}_{\hat{U}}(n) \\ =\ & E_P[L_U(Z^n) - L_{\hat{U}}(Z^n)] + \frac{1}{2}\frac{\mathrm{var}_P X}{\mathrm{var}_{M_{\mu^*}} X} \ln n + O(1), \end{aligned}$$

where the last equality follows from (6), which is valid under the conditions imposed on $\frac{d^4}{d\mu^4} D(M_{\mu^*} \| M_\mu)$ (see Condition 1 in [9] for details). We have:

$$\begin{aligned} & L_U(Z^n) - L_{\hat{U}}(Z^n) \\ =\ & \textstyle\sum_{i=0}^{n-1} \left(-\ln U(Z_{i+1} \mid Z_i) + \ln \hat{U}(Z_{i+1} \mid Z_i)\right) \\ =\ & \textstyle\sum_{i=0}^{n-1} \left(\ln\left(1 + \frac{1}{2i}\right) - \ln\left(1 + \frac{1}{2i} I_{\mathcal{M}}(\hat{\mu}_i)(X_{i+1} - \hat{\mu}_i)^2\right)\right). \end{aligned}$$

Since $\ln\left(1 + \frac{1}{2i}\right) = \frac{1}{2i} + O(i^{-2})$, we have:

$$\sum_{i=0}^{n-1} \ln\left(1 + \frac{1}{2i}\right) = \frac{1}{2}\ln n + O(1). \tag{14}$$

To analyze the second term in the sum, we use the fact that for arbitary a:

$$-a \leq -\ln(1+a) \leq -a + \frac{1}{2}a^2,$$

which follows e.g. from expanding the logarithm into Taylor expansion up to the second order. In our case, $a = V_i := \frac{1}{2i} I_{\mathcal{M}}(\hat{\mu}_i)(X_{i+1} - \hat{\mu}_i)^2$. We will show that $E_P[V_i^2]$ is $O(i^{-2})$, and then $E_P[-\ln(1+V_i)] = -E_P[V_i] + O(i^{-2})$. We have:

$$\begin{aligned} & E_P\left[V_i^2\right] = \tfrac{1}{4i^2} E_P\left[I_{\mathcal{M}}^2(\hat{\mu}_i)\left(X_{i+1} - \hat{\mu}_i\right)^4\right] \\ =\ & \tfrac{1}{4i^2} E_{X^i \sim P}\left[I_{\mathcal{M}}^2(\hat{\mu}_i) E_{X_{i+1}\sim P}\left[\left(X_{i+1} - \mu^* + \mu^* - \hat{\mu}_i\right)^4\right]\right] \\ =\ & \tfrac{1}{4i^2} E_P\left[I_{\mathcal{M}}^2(\hat{\mu}_i)\left(\mathrm{m}_P^{(4)} - 4\delta_i \mathrm{m}_P^{(3)} + 6\delta_i^2 \mathrm{var}_{P_{\mu^*}} X + \delta_i^4\right)\right], \end{aligned}$$

where $\mathrm{m}_P^{(k)}$ is $E_P[(X - \mu^*)^k]$, the k-th central moment of P_{μ^*}, and $\delta_i = \hat{\mu}_i - \mu^*$. We will show that the terms under expectation are bounded. If $I_{\mathcal{M}}(\hat{\mu}_i)$ is constant, then we apply Theorem 6 with $k = 1$, $k = 2$ and $k = 4$ to the second, third and fourth term, respectively and thus all the terms under expectation are $O(1)$. If $I_{\mathcal{M}}(\hat{\mu}_i)$ is not constant, then by Condition 2 the assumptions of Lemma 8 are satisfied with $f(\mu) = I_{\mathcal{M}}^2(\mu)$ and $s = 0, 2, 4$. Applying the lemma subsequently to the first, third and fourth term (with $s = 0, 2, 4$, respectively), we see that all those terms are $O(1)$. The second term is also $O(1)$ by applying Lemma 8 once again with $f(\mu) = \mu I_{\mathcal{M}}^2(\mu)$ and $s = 0$ (assumptions are again satisfied by Condition 2). Thus, we showed that $E_P[V_i^2] = O(i^{-2})$.

Next, we consider $E_P[V_i]$:

$$\begin{aligned} & E_P[V_i] = \frac{1}{2i} E_P\left[I_{\mathcal{M}}(\hat{\mu}_i)(X_{i+1} - \mu^* + \mu^* - \hat{\mu}_i)^2\right] \\ =\ & \frac{1}{2i} E_P\left[I_{\mathcal{M}}(\hat{\mu}_i)\left(\mathrm{var}_{P_{\mu^*}} X + \delta_i^2\right)\right] \\ =\ & \frac{1}{2i}\left(\mathrm{var}_{P_{\mu^*}} X E_P\left[I_{\mathcal{M}}(\hat{\mu}_i)\right] + E_P\left[I_{\mathcal{M}}(\hat{\mu}_i)\delta_i^2\right]\right). \end{aligned}$$

The second term $E_P\left[I_{\mathcal{M}}(\hat{\mu}_i)\delta_i^2\right]$ is $O(i^{-1})$ by Lemma 8 applied with $f(\mu) = I_{\mathcal{M}}(\mu)$ and $s = 2$. To analyze the first term we expand $I_{\mathcal{M}}(\hat{\mu}_i)$ into Taylor series around μ^*:

$$E_P\left[I_{\mathcal{M}}(\hat{\mu}_i)\right] = I_{\mathcal{M}}(\mu^*) + E_P\left[\frac{d}{d\mu} I_{\mathcal{M}}(\mu^*)\delta_i + \frac{d^2}{d^2\mu} I_{\mathcal{M}}(\mu)\delta_i^2\right],$$

for some μ between μ^* and $\hat{\mu}_i$. The linear term in the expansion is $O(i^{-1})$ by Theorem 6 applied with $k = 1$. The quadratic term is $O(i^{-1})$ by applying Lemma 8 with $f(\mu) = \frac{d^2}{d^2\mu} I_{\mathcal{M}}(\mu)$ and $s = 2$; Condition 2 guarantees that assumptions of the lemma are satisfied. Thus, using (11):

$$E_P[V_i] = \frac{1}{2i} I_{\mathcal{M}}(\mu^*) \mathrm{var}_{P_{\mu^*}} X + O(i^{-2}) = \frac{1}{2i}\frac{\mathrm{var}_{P_{\mu^*}} X}{\mathrm{var}_{M_{\mu^*}} X} + O(i^{-2}),$$

so that:

$$E_P[-\ln(1+V_i)] = -\frac{1}{2i}\frac{\mathrm{var}_{P_{\mu^*}} X}{\mathrm{var}_{M_{\mu^*}} X} + O(i^{-2}), \tag{15}$$

Taking together (15) and (14) we have:

$$L_U(Z^n) - L_{\hat{U}}(Z^n) = \frac{1}{2}\ln n - \frac{1}{2}\frac{\mathrm{var}_{P_{\mu^*}} X}{\mathrm{var}_{M_{\mu^*}} X}\ln n + O(1),$$

and thus:

$$R_U(n) = \frac{1}{2}\ln n + O(1).$$

■