Computable Bayesian Compression for Uniformly Discretizable Statistical Models

Łukasz Dębowski

Centrum Wiskunde & Informatica, 1098 XG Amsterdam, The Netherlands

Abstract. Supplementing Vovk and V'yugin's 'if' statement, we show that Bayesian compression provides the best enumerable compression for parameter-typical data if and only if the parameter is Martin-Löf random with respect to the prior. The result is derived for uniformly discretizable statistical models, introduced here. They feature the crucial property that given a discretized parameter, we can compute how much data is needed to learn its value with little uncertainty. Exponential families and certain nonparametric models are shown to be uniformly discretizable.

1 Introduction

Algorithmic information theory inspires an appealing interpretation of Bayesian inference [1,2,3,4]. Literally, a fixed individual parameter cannot have the property of being distributed according to a distribution but, when it is represented as a sequence of digits, the parameter is almost surely algorithmically random. Thus, if you believe that a parameter obeys a prior, it may rather mean that you suppose that the parameter is algorithmically random with respect to the prior. We want to argue that this interpretation is valid.

We will assume that the parameter $\boldsymbol{\theta}$ is, in some sense, effectively identifiable. Then one can disprove that a finite prefix of a fixed, not fully known $\boldsymbol{\theta}$ is algorithmically random by estimating the prefix and showing that there exists a shorter description of that prefix. Hence, Bayesian beliefs seem admissible scientific hypotheses according to the Popperian philosophy, cf. [1].

Secondly, it follows that the Bayesian measure $\int \boldsymbol{P_\theta} d\boldsymbol{Q}(\boldsymbol{\theta})$ gives the best enumerable compression of $\boldsymbol{P_\theta}$-typical data *if and only if* parameter $\boldsymbol{\theta}$ is algorithmically random with respect to prior $\boldsymbol{Q}$. This statement is useful when $\boldsymbol{P_\theta}$ is not computable for a fixed $\boldsymbol{\theta}$. Moreover, once we know where Bayesian compression fails, we should systematically adjust the prior to our hypotheses about the algorithmic complexity of $\boldsymbol{\theta}$ in an application.

As we will show, this *'if and only if'* result can be foreseen using the chain rule for prefix Kolmogorov complexity of finite objects [5], [6, Theorem 3.9.1]. The chain rule allows to relate randomness deficiencies for finite prefixes of the data and of the parameter in some specific statistical models, which we call uniformly discretizable. That yields a somewhat weaker *'if and only if'* statement. Subsequently, the statement can be strengthened using the dual chain rule for impossibility levels of infinite sequences [1, Theorem 1] and extensions of Lambalgen's

theorem for conditionally random sequences [7], [4, Theorem 4.2 and 5.3]. The condition of uniform discretization can be completely removed from the *'if'* part and relaxed to an effective identifiability of the parameter in the *'only if'* part. Namely, given a prefix of the parameter, we must be able to compute how much data is needed to learn its value with a fixed uncertainty.

The organization of this paper is as follows. In Section 2, we discuss quality of Bayesian compression for individual parameters and we derive the randomness deficiency bounds for prefixes of the parameter and the parameter-typical data. These bounds hold for the newly introduced class of uniformly discretizable statistical models. In Section 3, we show that exponential families are uniformly discretizable. The assumptions on the prior and the proof look familiar to statisticians working in minimum description length (MDL) inference [8,9]. An example of a 'nonparametric' uniformly discretizable model appears in Section 4. In the final Section 5, we prove that countable mixtures of uniformly discretizable models are uniformly discretizable if the Bayesian estimator consistently chooses the right submodel for the data.

The definition of uniformly discretizable models is given below. Condition (3) says that the parameter may be discretized to $m \geq \mu(n)$ digits for the sake of approximating the 'true' probability of data x^n. Condition (4) asserts that the parameter, discretized to m digits, can be predicted for all but finitely many m given data x^n of length $n \geq \nu(m)$. Functions μ and ν depend on a model.

To fix our notation in advance, we use a countable alphabet $\mathbb{X}$ and a finite $\mathbb{Y} = \{0, 1, ..., D - 1\}$, $D > 1$. The logarithm to base D is written as log. An italic $x \in \mathbb{X}^+$ is a string, a boldface $\boldsymbol{x} \in \mathbb{X}^{\mathbb{N}}$ is an infinite sequence. The n-th symbol of $\boldsymbol{x}$ is written as $x_n \in \mathbb{X}$ and x^n is the prefix of $\boldsymbol{x}$ of length n: $\boldsymbol{x} = x_1 x_2 x_3...$ and $x^n = x_1 x_2...x_n$. Capital boldface $\boldsymbol{Y} : \mathbb{X}^* \to \mathbb{R}$ denotes a distribution of strings normalized lengthwise, i.e., $0 \leq \boldsymbol{Y}(x)$, $\sum_a \boldsymbol{Y}(xa)\mathbf{1}_{\{|a|=n\}} = \boldsymbol{Y}(x)$, and $\boldsymbol{Y}(\lambda) = 1$ for the empty string λ. There is a unique measure on measurable sets of infinite sequences $\boldsymbol{x} \in \mathbb{X}^{\mathbb{N}}$, also denoted as $\boldsymbol{Y}$, such that $\boldsymbol{Y}(\{\boldsymbol{x} : x^n = x \text{ for } n = |x|\}) = \boldsymbol{Y}(x)$. Quantifier '$n$-eventually' means 'for all but finitely many $n \in \mathbb{N}$'.

Definition 1. *Fix a measurable subset $\boldsymbol{\Theta} \subset \mathbb{Y}^{\mathbb{N}}$. Let $\boldsymbol{P} : \mathbb{X}^* \times \boldsymbol{\Theta} \ni (x, \boldsymbol{\theta}) \mapsto \boldsymbol{P_\theta}(x) \in \mathbb{R}$ be a probability kernel, i.e., $\boldsymbol{P_\theta} : \mathbb{X}^* \to \mathbb{R}$ is a probability measure for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and the mapping $\boldsymbol{\theta} \mapsto \boldsymbol{P_\theta}$ is measurable. Let also $\boldsymbol{Q} : \mathbb{Y}^* \to \mathbb{R}$ be a probability measure on $\boldsymbol{\Theta}$, i.e., $\boldsymbol{Q}(\boldsymbol{\Theta}) = 1$. A Bayesian statistical model $(\boldsymbol{P}, \boldsymbol{Q})$ is called (μ, ν)-uniformly discretizable if it satisfies the following.*

(i) Define the measure $\boldsymbol{T} : \mathbb{X}^* \times \mathbb{Y}^* \to \mathbb{R}$ as

$$\boldsymbol{T}(x, \theta) := \int_{A(\theta)} \boldsymbol{P_\theta}(x) d\boldsymbol{Q}(\boldsymbol{\theta}), \tag{1}$$

where $A(\theta) := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \theta \text{ is the prefix of } \boldsymbol{\theta}\}$, and denote its other marginal

$$\boldsymbol{Y}(x) := \boldsymbol{T}(x, \lambda) = \int \boldsymbol{P_\theta}(x) d\boldsymbol{Q}(\boldsymbol{\theta}). \tag{2}$$

(ii) *Function* $\mu : \mathbb{N} \to \mathbb{R}$ *is nondecreasing and we require that for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\boldsymbol{P_\theta}$-*almost all* $\boldsymbol{x}$, *and* $m \geq \mu(n)$,

$$\lim_{n \to \infty} \frac{\log \left[\boldsymbol{Q}(\theta^m) \boldsymbol{P_\theta}(x^n) / \boldsymbol{T}(x^n, \theta^m) \right]}{\log m} = 0. \tag{3}$$

(iii) *Function* $\nu : \mathbb{N} \to \mathbb{R}$ *is nondecreasing and we require that for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\boldsymbol{P_\theta}$-*almost all* $\boldsymbol{x}$, *and* $n \geq \nu(m)$,

$$\lim_{m \to \infty} \boldsymbol{T}(x^n, \theta^m) / \boldsymbol{Y}(x^n) = 1. \tag{4}$$

Remark: A Bayesian model $(\tilde{\boldsymbol{P}}, \tilde{\boldsymbol{Q}})$ with a kernel $\tilde{\boldsymbol{P}} : \mathbb{X}^* \times \tilde{\boldsymbol{\Theta}} \to \mathbb{R}$ and a measure $\tilde{\boldsymbol{Q}}$ on $\tilde{\boldsymbol{\Theta}}$ will be called (ρ, μ, ν)-uniformly discretizable if $(\boldsymbol{P}, \boldsymbol{Q})$ is (μ, ν)-uniformly discretizable for a bijection $\rho : \tilde{\boldsymbol{\Theta}} \to \boldsymbol{\Theta}$, $\boldsymbol{P_\theta}(x) := \tilde{\boldsymbol{P}}_{\rho^{-1}(\boldsymbol{\theta})}(x)$, and $\boldsymbol{Q} := \tilde{\boldsymbol{Q}} \circ \rho^{-1}$. We will write '$(\rho, \mu(n), \nu(m))$-uniformly discretizable' when there are no convenient symbols for functions μ and ν.

A few words of comment to this definition are due. By condition (3), the support of prior $\boldsymbol{Q}$ equals $\boldsymbol{\Theta}$, i.e., $\boldsymbol{Q}(\theta^m) > 0$ for all m and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Condition (4) admits a consistent estimator if there is a function $\sigma : \mathbb{X}^* \to \mathbb{N}$, where $\nu(\sigma(x^n)) \leq n$, $\sigma(x^{n+1}) \geq \sigma(x^n)$, and $\lim_n \sigma(x^n) = \infty$. Define the discrete maximum likelihood estimator $\mathrm{MLE}(x; \sigma) := \mathrm{argmax}_{\theta \in \mathbb{Y}^m} \boldsymbol{T}(x, \theta)$ with $m = \sigma(x)$. The estimator is called consistent if $\mathrm{MLE}(x^n; \sigma) = \theta^{\sigma(x^n)}$ n-eventually for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\boldsymbol{P_\theta}$-almost all $\boldsymbol{x}$. This property is indeed satisfied.

Four models presented in Sections 3 and 4 satisfy a stronger condition.

Definition 2. *A* (μ, ν)-*uniformly discretizable model is called* μ-*uniformly discretizable if* ν *is recursive and* $\mu(\nu(m)) \leq m^\alpha$ *for an* $\alpha > 0$.

These models feature $\log \mu(n)$ close to the logarithm of Shannon redundancy $-\log \boldsymbol{Y}(x^n) + \log \boldsymbol{P_\theta}(x^n)$. A heuristic rationale is as follows. If we had $\mu \circ \nu = \mathrm{id}$, $-\log \boldsymbol{Q}(\theta^m) = \Omega(m)$, and we put $n = \nu(m)$ then

$$| -\log \boldsymbol{Y}(x^n) + \log \boldsymbol{P_\theta}(x^n) + \log \boldsymbol{Q}(\theta^m) | = o(\log m)$$

and hence $\mu(n) = m = O(-\log \boldsymbol{Y}(x^n) + \log \boldsymbol{P_\theta}(x^n))$. Whereas $-\log \boldsymbol{Q}(\theta^m) = \Omega(m)$ is a reasonable assumption, we rather observe $\mu(\nu(m)) > m$.

The present approach allows only discrete data. We hope, however, that uniformly discretizable models can be generalized to nondiscrete data so that consistency and algorithmic optimality of Bayesian procedures in density estimation could be characterized in a similar fashion, cf. [10]. Another interesting path of development is to integrate the algorithmic perspective on Bayesianism with the present MDL framework [8,9], where normalized maximum likelihood codes are discussed. By the algorithmic optimality of Bayesian compression, the normalized maximum likelihood measure, if it can be defined properly, should converge to the Bayesian measure $\int \boldsymbol{P_\theta} d\boldsymbol{Q}(\boldsymbol{\theta})$ in log-loss. We also suppose that reasonable luckiness functions, introduced to guarantee existence of modified normalized maximum likelihood codes [9, Section 11.3], may be close to algorithmic information between the parameter and the data.

2 Bounds for the Data and Parameter Complexity

We will use a universal computer with an oracle, which can compute certain functions $\mathbb{R} \to \mathbb{R}$. To make it clear which these are, we adopt the following definitions, cf. [11], [6, Sections 1.7 and 3.1], [1, Section 2], [4, Section 5]:

(i) A universal computer is an appropriate finite state machine that interacts with infinite tapes. The machine can move along the tapes in discrete steps, read and write on them single symbols from the finite set $\mathbb{Y}$, and announce the end of computation. We fix three one-sided tapes. At the beginning of computation, tape α contains a program, i.e., a string from a prefix-free subset of $\mathbb{Y}^+$, and tape β contains an oracle, i.e., an element of $(0\mathbb{Y}^*) \cup (1\mathbb{Y}^{\mathbb{N}})$. At the end of computation, tape γ contains an output, i.e., a string from $\mathbb{Y}^*$.

(ii) The prefix Kolmogorov complexity $K(y)$ of a string $y \in \mathbb{Y}^*$ is the minimal length of such a program on tape α that y is output on tape γ provided no symbol is read from tape β.

(iii) The conditional complexity $K(y|\delta)$ for $y \in \mathbb{Y}^*$ and $\delta \in \mathbb{Y}^* \cup \mathbb{Y}^{\mathbb{N}}$ is the minimal length of such a program on tape α that y is output on tape γ given 0δ or 1δ, respectively, as an oracle on tape β.

(iv) A function $f : \mathbb{Y}^* \cup \mathbb{Y}^{\mathbb{N}} \to \mathbb{Y}^*$ is *recursive* if there is such a program $z \in \mathbb{Y}^+$ that string $f(y)$ is output for all oracles $y \in \mathbb{Y}^* \cup \mathbb{Y}^{\mathbb{N}}$.

(v) Function ϕ is a *prefix code* if it is an injection and its image is prefix-free.

(vi) For certain prefix codes $\phi_{\mathbb{W}} : \mathbb{W} \to \mathbb{Y}^*$ and $\phi_{\mathbb{U}} : \mathbb{U} \to \mathbb{Y}^* \cup \mathbb{Y}^{\mathbb{N}}$ and arbitrary $w \in \mathbb{W}$ and $u \in \mathbb{U}$, we put $K(w) := K(\phi_{\mathbb{W}}(w))$ and $K(w|u) := K(\phi_{\mathbb{W}}(w)|\phi_{\mathbb{W}}(u))$. Fixing $\phi_{\mathbb{Y}^*}$ and $\phi_{\mathbb{Y}^* \cup \mathbb{Y}^{\mathbb{N}}}$ as identity functions, $f : \mathbb{U} \to \mathbb{W}$ is called *recursive* if so is $\phi_{\mathbb{W}} \circ f \circ \phi_{\mathbb{U}}^{-1}$.

(vii) Numbers obey special conventions. Integers are Elias-coded [12], whereas $\phi_{\mathbb{Q}}(p/q) := \phi_{\mathbb{Z}}(p)_{\mathbb{N}}(q)$ for every irreducible fraction p/q. To convert a real number from $(-\infty, \infty)$ into a one-sided sequence, we assume that $\phi_{\mathbb{R}}(\boldsymbol{r}) = \boldsymbol{\theta}$ satisfies $[1 + \exp(-\boldsymbol{r})] = \sum_{i=1}^{\infty} \theta_i D^{-i}$. This solves the problem of real arguments. A real-valued function $f : \mathbb{W} \to \mathbb{R}$ is called *enumerable* if there is a recursive function $g : \mathbb{W} \times \mathbb{N} \to \mathbb{Q}$ nondecreasing in k such that $\lim_k g(w, k) = f(w)$. A stronger condition, the f is called *recursive* if there is a recursive function $h : \mathbb{W} \times \mathbb{N} \to \mathbb{Q}$ such that $|f(w) - h(w, k)| < 1/k$.

(viii) Pairs (w, u) enjoy the code $\phi_{\mathbb{W} \times \mathbb{U}}(w, u) := \phi_{\mathbb{W}}(w)\phi_{\mathbb{W}}(u)$. This code cannot be used if w is real. In the Proposition 2 of Section 3, where we need to string real vectors, Cantor's code is used instead.

(ix) The concepts mentioned above are analogously extended to partial functions. Special care must be taken to assume computability of their domains, which is important to guarantee that the inverse of the Shannon-Fano-Elias code, used in Theorem 1, is recursive.

Last but not least, a *semimeasure* $\boldsymbol{U}$ is a function $\mathbb{X}^* \to \mathbb{R}$ that satisfies $0 \le \boldsymbol{U}(x)$, $\sum_a \boldsymbol{U}(xa)\mathbf{1}_{\{|a|=n\}} \le \boldsymbol{U}(x)$, and $\boldsymbol{U}(\lambda) \le 1$. Symbol $\overset{*}{<}$ denotes inequality up to a multiplicative constant.

Impossibility level

$$\mathcal{I}(\boldsymbol{x}; \boldsymbol{Y}) := \inf_{n \in \mathbb{N}} \frac{D^{-K(x^n)}}{\boldsymbol{Y}(x^n)} \tag{5}$$

is a natural measure of randomness deficiency for a sequence $\boldsymbol{x} \in \mathbb{X}^{\mathbb{N}}$ with respect to a recursive measure $\boldsymbol{Y}$, cf. [1], [6, Def. 4.5.10 and Thm. 4.5.5]. The respective set of $\boldsymbol{Y}$-Martin-Löf random sequences

$$\mathcal{L}_{\boldsymbol{Y}} := \{\boldsymbol{x} : \mathcal{I}(\boldsymbol{x}; \boldsymbol{Y}) < \infty\} \tag{6}$$

has two important properties.

Firstly, $\mathcal{L}_{\boldsymbol{Y}}$ is the maximal set of sequences on which no enumerable semimeasure outperforms a recursive measure $\boldsymbol{Y}$ more than by a multiplicative constant. Let $\boldsymbol{M}$ be the universal enumerable semimeasure [6, Section 4.5.1]. By [2, Theorem 1 and Lemma 3], we have

$$\mathcal{I}(\boldsymbol{x}; \boldsymbol{Y}) \overset{*}{<} \liminf_{n \to \infty} \frac{\boldsymbol{M}(x^n)}{\boldsymbol{Y}(x^n)} \overset{*}{<} \sup_{n \in \mathbb{N}} \frac{\boldsymbol{M}(x^n)}{\boldsymbol{Y}(x^n)} \overset{*}{<} [\mathcal{I}(\boldsymbol{x}; \boldsymbol{Y})]^{1+\epsilon} \tag{7}$$

for a fixed $\epsilon > 0$ and recursive $\boldsymbol{Y}$. By the definition of $\boldsymbol{M}$, $\boldsymbol{U}(x^n) \overset{*}{<} \boldsymbol{M}(x^n)$ for any enumerable (semi)measure $\boldsymbol{U}$. Hence $\sup_{n \in \mathbb{N}} \boldsymbol{U}(x^n)/\boldsymbol{Y}(x^n) < \infty$ if $\boldsymbol{x} \in \mathcal{L}_{\boldsymbol{Y}}$. Moreover, $\mathcal{L}_{\boldsymbol{Y}} = \mathcal{L}_{\boldsymbol{U}}$ if $\boldsymbol{Y}$ and $\boldsymbol{U}$ are *mutually equivalent* recursive measures, i.e., $\sup_{n \in \mathbb{N}} \boldsymbol{U}(x^n)/\boldsymbol{Y}(x^n) < \infty \iff \sup_{n \in \mathbb{N}} \boldsymbol{Y}(x^n)/\boldsymbol{U}(x^n) < \infty$ for all $\boldsymbol{x} \in \mathbb{X}^{\mathbb{N}}$.

Secondly, the set $\mathcal{L}_{\boldsymbol{Y}}$ has full measure $\boldsymbol{Y}$. The fact is well-known, cf. e.g. [1, Remark 2], and it can be seen easily using the auxiliary statement below, which strengthens Barron's result [13, Theorem 3.1]. Whereas $\boldsymbol{Y}(\mathcal{L}_{\boldsymbol{Y}}) = 1$ follows for $|B(\cdot)| = K(\cdot)$, we shall use this lemma later also for $|B(\cdot)| = K(\cdot|\boldsymbol{\theta})$.

Lemma 1 (no hypercompression). *Let $B : \mathbb{X}^* \to \mathbb{Y}^+$ be a prefix code. Then*

$$|B(x^n)| + \log \boldsymbol{Y}(x^n) > 0 \tag{8}$$

n-eventually for $\boldsymbol{Y}$-almost all sequences $\boldsymbol{x}$.

Proof. Consider the function $W(x) := D^{-|B(x)|}$. By the Markov inequality,

$$\boldsymbol{Y}\left((8) \text{ is false}\right) = \boldsymbol{Y}\left(\frac{W(x^n)}{\boldsymbol{Y}(x^n)} \geq 1\right) \leq \mathbf{E}_{\boldsymbol{x} \sim \boldsymbol{Y}}\left[\frac{W(x^n)}{\boldsymbol{Y}(x^n)}\right] = \sum_x \mathbf{1}_{\{|x|=n\}} W(x).$$

Hence $\sum_n \boldsymbol{Y}\left((8) \text{ is false}\right) \leq \sum_x D^{-|B(x)|} \leq 1 < \infty$ by the Kraft inequality. The claim now follows by the Borel-Cantelli lemma. $\square$

Now let $\boldsymbol{Y}$ be a recursive Bayesian measure (2). In a prototypical case, measures $\boldsymbol{P}_{\boldsymbol{\theta}}$ are not enumerable $\boldsymbol{Q}$-almost surely. But the data that are almost surely typical for these measures can be optimally compressed with the effectively computable measure $\boldsymbol{Y}$. That is, $\boldsymbol{P}_{\boldsymbol{\theta}}(\mathcal{L}_{\boldsymbol{Y}}) = 1$ holds $\boldsymbol{Q}$-almost everywhere, as implied by the following statement.

Lemma 2 (cf. [14, Section 9]). *Equality* $\boldsymbol{Y}(\mathcal{X}) = 1$ *for* $\boldsymbol{Y} = \int \boldsymbol{P_\theta} d\boldsymbol{Q}(\boldsymbol{\theta})$ *implies* $\boldsymbol{P_\theta}(\mathcal{X}) = 1$ *for* $\boldsymbol{Q}$*-almost all* $\boldsymbol{\theta}$.

Proof. Let $\mathcal{G}_n := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \boldsymbol{P_\theta}(\mathcal{X}) \geq 1 - 1/n\}$. We have $1 = \boldsymbol{Y}(\mathcal{X}) \leq \boldsymbol{Q}(\mathcal{G}_n) + \boldsymbol{Q}(\boldsymbol{\Theta} \setminus \mathcal{G}_n)(1 - 1/n) = 1 - n^{-1}\boldsymbol{Q}(\boldsymbol{\Theta} \setminus \mathcal{G}_n)$. Thus $\boldsymbol{Q}(\mathcal{G}_n) = 1$. By σ-additivity, $\boldsymbol{Q}(\mathcal{G}) = \inf_n \boldsymbol{Q}(\mathcal{G}_n) = 1$ follows for $\mathcal{G} := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \boldsymbol{P_\theta}(\mathcal{X}) = 1\} = \bigcap_n \mathcal{G}_n$. $\square$

Notably, the Bayesian compressor can be shown optimal exactly when the parameter is incompressible. Strictly speaking, we will obtain $\boldsymbol{P_\theta}(\mathcal{L_Y}) = 1$ if and only if $\boldsymbol{\theta}$ is Martin-Löf random with respect to $\boldsymbol{Q}$. This holds, of course, under some tacit assumptions. For instance, if we take $\boldsymbol{P_\theta} \equiv \boldsymbol{Y}$ then $\boldsymbol{P_\theta}(\mathcal{L_Y}) = 1$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. We may thus suppose that the *'if and only if'* statement holds provided the parameter can be effectively identified. The following two propositions form the first step to see what assumptions are needed exactly.

Lemma 3. *For a computer-dependent constant A, we have*

$$K(x|\boldsymbol{\theta}) \leq A + K(x|\theta^m, K(\theta^m)) + K(K(\theta^m)) + K(m). \tag{9}$$

Proof. A certain program for computing x given $\boldsymbol{\theta}$ operates as follows. It first calls a subroutine of length $K(m)$ to compute m and a subroutine of length $K(K(\theta^m))$ to compute $K(\theta^m)$. Then it reads the prefix θ^m of $\boldsymbol{\theta}$ and passes θ^m and $K(\theta^m)$ to a subroutine of length $K(x|\theta^m, K(\theta^m))$ which returns x. $\square$

Theorem 1. *Let $(\boldsymbol{P}, \boldsymbol{Q})$ be a Bayesian statistical model with a recursive prior $\boldsymbol{Q} : \mathbb{Y}^* \to \mathbb{R}$ and a recursive kernel $\boldsymbol{P} : \mathbb{X}^* \times \boldsymbol{\Theta} \to \mathbb{R}$.*

(i) If (3) holds for $\boldsymbol{P_\theta}$-almost all $\boldsymbol{x}$ then

$$K(x^n) + \log \boldsymbol{Y}(x^n) \geq K(\theta^m) + \log \boldsymbol{Q}(\theta^m) - 3\log m + o(\log m) \tag{10}$$

is also true for $\boldsymbol{P_\theta}$-almost all $\boldsymbol{x}$.

(ii) If (4) holds for a recursive $\tau : \mathbb{Y}^* \to \mathbb{N}$ and $n = \tau(\theta^m)$ then

$$K(x^n) + \log \boldsymbol{Y}(x^n) \leq K(\theta^m) + \log \boldsymbol{Q}(\theta^m) + O(1). \tag{11}$$

Proof. (i) For $\boldsymbol{P_\theta}$-almost all $\boldsymbol{x}$ we have both (3) and

$$K(x^n|\boldsymbol{\theta}) + \log \boldsymbol{P_\theta}(x^n) \geq 0 \tag{12}$$

n-eventually, by Lemma 1 for $|B(\cdot)| = K(\cdot|\boldsymbol{\theta})$. Applying Lemma 3 to these sequences yields

$$K(x^n|\theta^m, K(\theta^m)) + \log \boldsymbol{T}(x^n, \theta^m) - \log \boldsymbol{Q}(\theta^m)$$
$$\geq -K(K(\theta^m)) - K(m) + o(\log m) = -2\log m + o(\log m)$$

because $K(\theta^m) \leq m + \log m + o(\log m)$ and $K(m) \leq \log m + o(\log m)$. Since

$$K(x^n|\theta^m, K(\theta^m)) + K(\theta^m) = K(x^n, \theta^m) + O(1) \tag{13}$$

by the chain rule for prefix complexity [6, Theorem 3.9.1], we obtain

$$K(x^n, \theta^m) + \log \boldsymbol{T}(x^n, \theta^m) \geq K(\theta^m) + \log \boldsymbol{Q}(\theta^m) - 2 \log m + o(\log m).$$

In the following, we apply (13) with x^n and θ^m switched, and observe that

$$K(\theta^m | x^n, K(x^n)) \leq A + K(m) - \log \frac{\boldsymbol{T}(x^n, \theta^m)}{\boldsymbol{Y}(x^n)}$$

follows by conditional Shannon-Fano-Elias coding of θ^m of an arbitrary length given x^n, cf. [15, Section 5.9]. Hence (10) holds for $\boldsymbol{P_\theta}$-almost all $\boldsymbol{x}$.

(ii) By conditional Shannon-Fano-Elias coding of x^n given θ^m we obtain

$$K(x^n, \theta^m) \leq A' + K(\theta^m) - \log \frac{\boldsymbol{T}(x^n, \theta^m)}{\boldsymbol{Q}(\theta^m)}. \tag{14}$$

(This time, we need not specify the length of x^n separately since it can be computed from θ^m.) Substituting (4) into (14) and chaining the result with $K(x^n) \leq A'' + K(x^n, \theta^m)$ yields (11). □

Theorem 1 applies to uniformly discretizable models if we plug in $m \geq \mu(n)$ and $\tau(\theta^m) \geq \nu(m)$. Hence we obtain the first, less elegant dichotomy.

Proposition 1. *Let* $(\boldsymbol{P}, \boldsymbol{Q})$ *be a* μ-*uniformly discretizable model with a recursive prior* $\boldsymbol{Q} : \mathbb{Y}^* \to \mathbb{R}$ *and a recursive kernel* $\boldsymbol{P} : \mathbb{X}^* \times \boldsymbol{\Theta} \to \mathbb{R}$. *We have*

$$\boldsymbol{P_\theta}(\mathcal{L}_{\boldsymbol{Y}, \log \mu(n)}) = \begin{cases} 1 & \text{if } \boldsymbol{\theta} \in \mathcal{L}_{\boldsymbol{Q}, \log n}, \\ 0 & \text{if } \boldsymbol{\theta} \notin \mathcal{L}_{\boldsymbol{Q}, \log n}, \end{cases} \tag{15}$$

where the sets of $(\boldsymbol{Y}, g(n))$-*random sequences are defined as*

$$\mathcal{L}_{\boldsymbol{Y}, g(n)} := \left\{ \boldsymbol{x} : \inf_{n \in \mathbb{N}} \frac{K(x^n) + \log \boldsymbol{Y}(x^n)}{g(n)} > -\infty \right\}. \tag{16}$$

In particular, $\mathcal{L}_{\boldsymbol{Y}, 1} = \mathcal{L}_{\boldsymbol{Y}}$.

Theorem 1(ii) suffices to prove $\boldsymbol{P_\theta}(\mathcal{L}_{\boldsymbol{Y}}) = 0$ for $\boldsymbol{\theta} \notin \mathcal{L}_{\boldsymbol{Q}}$ but to show $\boldsymbol{P_\theta}(\mathcal{L}_{\boldsymbol{Y}}) = 1$ in the other case we need a stronger statement than Theorem 1(i). Here we can rely on the chain rule for conditional impossibility levels by Vovk and V'yugin [1, Theorem 1] and extensions of Lambalgen's theorem for conditionally random sequences by Takahashi [4]. For a recursive kernel $\boldsymbol{P}$, let us define by analogy the conditional impossibility level

$$\mathcal{I}(\boldsymbol{x}; \boldsymbol{P} | \boldsymbol{\theta}) := \inf_{n \in \mathbb{N}} \frac{D^{-K(x^n | \boldsymbol{\theta})}}{\boldsymbol{P_\theta}(x^n)} \tag{17}$$

and the set of conditionally random sequences

$$\mathcal{L}_{\boldsymbol{P} | \boldsymbol{\theta}} := \left\{ \boldsymbol{x} \in \mathbb{X}^{\mathbb{N}} : \mathcal{I}(\boldsymbol{x}; \boldsymbol{P} | \boldsymbol{\theta}) < \infty \right\}. \tag{18}$$

We have $\boldsymbol{P_\theta}(\mathcal{L}_{\boldsymbol{P}|\boldsymbol{\theta}}) = 1$ for all $\boldsymbol{\theta}$ by Lemma 1, as used in (12). Adjusting the proof of [6, Theorem 4.5.5] to computation with an oracle, we can show that the definition of $\mathcal{I}(\boldsymbol{x}; \boldsymbol{P}|\boldsymbol{\theta})$ given here is equivalent to the one given by [1], cf. [6, Def. 4.5.10]. Hence

$$\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left[\mathcal{I}(\boldsymbol{x}; \boldsymbol{P}|\boldsymbol{\theta}) \, \mathcal{I}(\boldsymbol{\theta}; \boldsymbol{Q}) \right] \overset{*}{<} \mathcal{I}(\boldsymbol{x}; \boldsymbol{Y}) \overset{*}{<} \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left[\mathcal{I}(\boldsymbol{x}; \boldsymbol{P}|\boldsymbol{\theta}) \, [\mathcal{I}(\boldsymbol{\theta}; \boldsymbol{Q})]^{1+\epsilon} \right] \qquad (19)$$

holds for $\boldsymbol{Y} = \int \boldsymbol{P_\theta} d\boldsymbol{Q}(\boldsymbol{\theta})$ and $\epsilon > 0$ by [1, Corollary 4].

Inequality (19) and Theorem 1(ii) imply the main claim of this article.

Theorem 2. *Let $(\boldsymbol{P}, \boldsymbol{Q})$ be a Bayesian statistical model with a recursive prior $\boldsymbol{Q} : \mathbb{Y}^* \to \mathbb{R}$ and a recursive kernel $\boldsymbol{P} : \mathbb{X}^* \times \boldsymbol{\Theta} \to \mathbb{R}$. Suppose that (4) holds for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\boldsymbol{P_\theta}$-almost all $\boldsymbol{x}$, and $n = \tau(\theta^m)$, where $\tau : \mathbb{Y}^* \to \mathbb{N}$ is recursive. Then we have*

$$\boldsymbol{P_\theta}(\mathcal{L}_{\boldsymbol{Y}}) = \begin{cases} 1 & \text{if } \boldsymbol{\theta} \in \mathcal{L}_{\boldsymbol{Q}}, \\ 0 & \text{if } \boldsymbol{\theta} \notin \mathcal{L}_{\boldsymbol{Q}}. \end{cases} \qquad (20)$$

The upper part of (20) can be strengthened as decomposition $\mathcal{L}_{\boldsymbol{Y}} = \bigcup_{\boldsymbol{\theta} \in \mathcal{L}_{\boldsymbol{Q}}} \mathcal{L}_{\boldsymbol{P}|\boldsymbol{\theta}}$, which holds for all recursive $\boldsymbol{P}$ and $\boldsymbol{Q}$ [4, Cor. 4.3 & Thm. 5.3]. (Our definition of a recursive $\boldsymbol{P}$ corresponds to 'uniformly computable' in [4].) We suppose that, under the assumption of Theorem 2, sets $\mathcal{L}_{\boldsymbol{P}|\boldsymbol{\theta}}$ are disjoint for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. This would strengthen the lower part of (20).

3   The Case of Exponential Families

As shown in [16], k-parameter exponential families exhibit Shannon redundancy $-\log \boldsymbol{Y}(x^n) + \log \boldsymbol{P_\theta}(x^n) = \frac{k}{2} \log n + \Theta(\log \log n)$. Here we shall prove that these models are uniformly discretizable with $\mu(n) = \left(\frac{k}{2} + \epsilon\right) \log n$ respectively. The result is established under a familiar condition. Namely, a prior $\tilde{\boldsymbol{Q}}$ on $\tilde{\boldsymbol{\Theta}} \subset \mathbb{R}^k$ is *universally lower-bounded* by the Lebesgue measure $\boldsymbol{\lambda}$ if for each $\boldsymbol{\vartheta} \in \tilde{\boldsymbol{\Theta}}$ there exists an open set $C \ni \boldsymbol{\vartheta}$ and a $w > 0$ such that $\tilde{\boldsymbol{Q}}(E) \geq w\boldsymbol{\lambda}(E)$ for every measurable $E \subset C$. This condition implies that $\tilde{\boldsymbol{\Theta}}$ is the support of $\tilde{\boldsymbol{Q}}$ and is satisfied, in particular, if $\tilde{\boldsymbol{Q}}$ and $\boldsymbol{\lambda}$ restricted to $\tilde{\boldsymbol{\Theta}}$ are mutually equivalent.

Let us write the components of vectors $\boldsymbol{\vartheta}, \boldsymbol{\vartheta}' \in \mathbb{R}^k$ as $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2, ..., \boldsymbol{\vartheta}_k)$ and their Euclidean distance as $|\boldsymbol{\vartheta}' - \boldsymbol{\vartheta}| := \sqrt{\sum_{l=1}^{k} (\boldsymbol{\vartheta}'_l - \boldsymbol{\vartheta}_l)^2}$.

Example 1 (an exponential family). Let the kernel $\tilde{\boldsymbol{P}} : \mathbb{X}^* \times \tilde{\boldsymbol{\Theta}} \ni (x, \boldsymbol{\vartheta}) \mapsto \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}}(x) \in \mathbb{R}$ represent a regular k-parameter exponential family. That is:

(i) Certain functions $p : \mathbb{X} \to (0, \infty)$ and $T : \mathbb{X} \to \mathbb{R}^k$ satisfy $\sum_{x \in \mathbb{X}} p(x) < \infty$ and $\forall_{\boldsymbol{\beta} \in \mathbb{R}^k \setminus \boldsymbol{0}} \forall_{c \in \mathbb{R}} \exists_{x \in \mathbb{X}} \sum_{l=1}^{k} \boldsymbol{\beta}_l T_l(x) \neq c$ (i.e., T has affinely independent components).

(ii) Let $Z(\boldsymbol{\beta}) := \sum_{x \in \mathbb{X}} p(x) \exp\left(\sum_{l=1}^{k} \boldsymbol{\beta}_l T_l(x)\right)$ and define measures

$$\tilde{\boldsymbol{P}}_{\boldsymbol{\beta}}(x^n) := \prod_{i=1}^{n} p(x_i) \exp\left(\sum_{l=1}^{k} \boldsymbol{\beta}_l T_l(x) - \ln Z(\boldsymbol{\beta})\right)$$

for $\boldsymbol{\beta} \in \boldsymbol{B} := \left\{\boldsymbol{\beta}' \in \mathbb{R}^k : Z(\boldsymbol{\beta}') < \infty\right\}$.

(iii) We require that $\boldsymbol{B}$ is open. (It is not empty since $\mathbf{0} \in \boldsymbol{B}$.) Under this condition, $\boldsymbol{\vartheta}(\cdot) : \boldsymbol{B} \ni \boldsymbol{\beta} \mapsto \boldsymbol{\vartheta}(\boldsymbol{\beta}) := \mathbf{E}_{\boldsymbol{x} \sim \tilde{\boldsymbol{P}}_{\boldsymbol{\beta}}} T(x_i) \in \mathbb{R}^k$ is a twice differentiable injection [17], [9]. Thus assume $\tilde{\boldsymbol{\Theta}} := \boldsymbol{\vartheta}(\boldsymbol{B})$ and put $\tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}} := \tilde{\boldsymbol{P}}_{\boldsymbol{\beta}(\boldsymbol{\vartheta})}$ for $\boldsymbol{\beta}(\cdot) := \boldsymbol{\vartheta}^{-1}(\cdot)$.

Additionally, let the prior $\tilde{\boldsymbol{Q}}$ be universally lower-bounded by the Lebesgue measure on $\mathbb{R}^k$ and let it satisfy $\tilde{\boldsymbol{Q}}(\tilde{\boldsymbol{\Theta}}) = 1$.

Proposition 2. *Use Cantor's code $\rho := \rho_s \circ \rho_n$, where $\rho_n : \tilde{\boldsymbol{\Theta}} \to (0,1)^k$ is a differentiable injection and $\rho_s : (0,1)^k \to \mathbb{Y}^{\mathbb{N}}$ satisfies $\rho_s(\boldsymbol{y}) = \theta_1 \theta_2 \theta_3...$ for any vector $\boldsymbol{y} \in (0,1)^k$ with components $\boldsymbol{y}_l = \sum_{i=1}^{\infty} \theta_{(i-1)k+l} D^{-i}$. Then the model $(\tilde{\boldsymbol{P}}, \tilde{\boldsymbol{Q}})$ is $\left(\rho, \left(\frac{k}{2} + \epsilon\right) \log n, D^{(2/k+\epsilon)m}\right)$-uniformly discretizable for $\epsilon > 0$.*

Proof. Let $\boldsymbol{\Theta} := \rho(\tilde{\boldsymbol{\Theta}})$, $\boldsymbol{P}_{\boldsymbol{\theta}}(x) := \tilde{\boldsymbol{P}}_{\rho^{-1}(\boldsymbol{\theta})}(x)$, $\boldsymbol{Q} := \tilde{\boldsymbol{Q}} \circ \rho^{-1}$, and $A(\theta) := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \theta$ is the prefix of $\boldsymbol{\theta}\}$. Consider a $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Firstly, let $m \geq \left(\frac{k}{2} + \epsilon\right) \log n$. We have (21) for $\boldsymbol{\vartheta} = \rho^{-1}(\boldsymbol{\theta})$ and $A_n = \rho^{-1}(A(\theta^m))$. Hence (3) holds by the Theorem 3(i) below. Secondly, let $n \geq D^{(2/k+\epsilon)m}$. We have (23) for $\boldsymbol{\vartheta} = \rho^{-1}(\boldsymbol{\theta})$ and $B_n = \rho^{-1}(A(\theta^m))$. Hence (4) follows by Theorem 3(ii). $\square$

The statement below may look more familiar for statisticians.

Theorem 3. *Fix a $\boldsymbol{\vartheta} \in \tilde{\boldsymbol{\Theta}}$ for the model specified in Example 1.*

(i) If we take sufficiently small measurable sets $A_n \subset \tilde{\boldsymbol{\Theta}}$ which satisfy

$$\limsup_{n \to \infty} \frac{\sup_{\boldsymbol{\vartheta}' \in A_n} |\boldsymbol{\vartheta}' - \boldsymbol{\vartheta}|}{\sqrt{n^{-1} \ln \ln n}} = 0 \tag{21}$$

and put $\tilde{\boldsymbol{P}}_n(x) := \int_{A_n} \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}'}(x) d\tilde{\boldsymbol{Q}}(\boldsymbol{\vartheta}') / \int_{A_n} d\tilde{\boldsymbol{Q}}(\boldsymbol{\vartheta}')$ then

$$\lim_{n \to \infty} \frac{\log \tilde{\boldsymbol{P}}_n(x^n) - \log \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}}(x^n)}{\ln \ln n} = 0 \tag{22}$$

for $\tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}}$-almost all $\boldsymbol{x}$.

(ii) On the other hand, if we take sufficiently large measurable sets

$$B_n \supset \left\{\boldsymbol{\vartheta}' \in \tilde{\boldsymbol{\Theta}} : |\boldsymbol{\vartheta}' - \boldsymbol{\vartheta}| \geq n^{-1/2+\alpha}\right\} \tag{23}$$

for an arbitrary $\alpha \in (0, 1/2)$ then

$$\lim_{n \to \infty} \left(\log \int \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}'}(x^n) d\tilde{\boldsymbol{Q}}(\boldsymbol{\vartheta}') - \log \int_{B_n} \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}'}(x^n) d\tilde{\boldsymbol{Q}}(\boldsymbol{\vartheta}')\right) = 0 \tag{24}$$

for $\tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}}$-almost all $\boldsymbol{x}$.

Proof. (i) Function $\hat{\boldsymbol{\vartheta}}(x^n) := n^{-1}\sum_{i=1}^n T(x_i)$ is the maximum likelihood estimator of $\boldsymbol{\vartheta}$, in the usual sense. Thus the Taylor expansion for any $\boldsymbol{\vartheta} \in \tilde{\boldsymbol{\Theta}}$ yields

$$\log \tilde{\boldsymbol{P}}_{\hat{\boldsymbol{\vartheta}}(x^n)}(x^n) - \log \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}}(x^n) = n\sum_{l,m=1}^k R_{lm}(\boldsymbol{\vartheta})S_{lm}(\boldsymbol{\vartheta}), \tag{25}$$

where $S_{lm}(\boldsymbol{\vartheta}) := (\boldsymbol{\vartheta}_l - \hat{\boldsymbol{\vartheta}}_l(x^n))(\boldsymbol{\vartheta}_m - \hat{\boldsymbol{\vartheta}}_m(x^n))$ and $R_{lm}(\boldsymbol{\vartheta}) := \int_0^1 (1-t)I_{lm}(t\boldsymbol{\vartheta} + (1-t)\hat{\boldsymbol{\vartheta}}(x^n))dt$, whereas the observed Fisher information matrix $I_{lm}(\boldsymbol{\vartheta}) := -n^{-1}\partial_{\boldsymbol{\vartheta}_l}\partial_{\boldsymbol{\vartheta}_m} \log \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}}(x^n)$ does not depend on n and x^n. Consequently,

$$\log \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}}(x^n) - \log \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}'}(x^n) =$$
$$n\sum_{l,m=1}^k \left[R_{lm}(\boldsymbol{\vartheta}')\left[S_{lm}(\boldsymbol{\vartheta})' - S_{lm}(\boldsymbol{\vartheta})\right] + \left[R_{lm}(\boldsymbol{\vartheta}') - R_{lm}(\boldsymbol{\vartheta})\right]S_{lm}(\boldsymbol{\vartheta})\right].$$

With C_n denote the intersection of $\tilde{\boldsymbol{\Theta}}$ and the smallest ball containing A_n and $\hat{\boldsymbol{\vartheta}}(x^n)$. Let $d_n := \left|\boldsymbol{\vartheta} - \hat{\boldsymbol{\vartheta}}(x^n)\right|$ and $a_n := \sup_{\boldsymbol{\vartheta}' \in A_n} |\boldsymbol{\vartheta}' - \boldsymbol{\vartheta}|$. Hence we bound

$$\left|\log \tilde{\boldsymbol{P}}_n(x^n) - \log \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}}(x^n)\right| \leq n\sum_{l,m=1}^k \left[|R_{lm}^+|a_n(2d_n + a_n) + |R_{lm}^+ - R_{lm}^-|d_n^2\right],$$

where $R_{lm}^+ := \sup_{\boldsymbol{\vartheta}' \in C_n} R_{lm}(\boldsymbol{\vartheta}')$ and $R_{lm}^- := \inf_{\boldsymbol{\vartheta}' \in C_n} R_{lm}(\boldsymbol{\vartheta}')$. By continuity of Fisher information $I_{lm}(\boldsymbol{\vartheta})$ as a function of $\boldsymbol{\vartheta}$, R_{lm}^+ and R_{lm}^- tend to $I_{lm}(\boldsymbol{\vartheta})$ for $n \to \infty$. On the other hand, the law of iterated logarithm

$$\limsup_{n\to\infty} \frac{\hat{\boldsymbol{\vartheta}}_l(x^n) - \boldsymbol{\vartheta}_l}{\sigma_l\sqrt{2n^{-1}\ln\ln n}} = 1 \tag{26}$$

is satisfied for $\tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}}$-almost all $\boldsymbol{x}$ with variance $\sigma_l^2 := \text{Var}_{\boldsymbol{x}\sim\tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}}} T_l(x_i)$ since the maximum likelihood estimator is unbiased, i.e., $\mathbf{E}_{\boldsymbol{x}\sim\tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}}} \hat{\boldsymbol{\vartheta}}(x^n) = \boldsymbol{\vartheta}$. Consequently, we obtain (22) for (21).

(ii) The proof applies Laplace approximation as in [18] or in the proof of Theorem 8.1 of [9, pages 248–251]. First of all, we have

$$\log \int \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}'}(x^n)d\tilde{\boldsymbol{Q}}(\boldsymbol{\vartheta}') - \log \int_{B_n} \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}'}(x^n)d\tilde{\boldsymbol{Q}}(\boldsymbol{\vartheta}') \leq \frac{\int_{\tilde{\boldsymbol{\Theta}}\setminus B_n} \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}'}(x^n)d\tilde{\boldsymbol{Q}}(\boldsymbol{\vartheta}')}{\int_{B_n} \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}'}(x^n)d\tilde{\boldsymbol{Q}}(\boldsymbol{\vartheta}')}.$$

In the following, we consider a sufficiently large n. Because of the law of iterated logarithm (26), $\hat{\boldsymbol{\vartheta}}(x^n)$ belongs to B_n for $\tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}}$-almost all $\boldsymbol{x}$. Hence the robustness property and the convexity of Kullback-Leibler divergence for exponential families [9, Eq. (19.12) and Proposition 19.2] imply a bound for the numerator

$$\int_{\tilde{\boldsymbol{\Theta}}\setminus B_n} \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}'}(x^n)d\tilde{\boldsymbol{Q}}(\boldsymbol{\vartheta}') \leq \sup_{\boldsymbol{\vartheta}'\in\tilde{\boldsymbol{\Theta}}\setminus B_n} \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}'}(x^n) \leq \sup_{\boldsymbol{\vartheta}'\in\partial B_n} \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}'}(x^n),$$

where ∂B_n is the boundary of B_n. Using (25) gives further

$$\sup_{\boldsymbol{\vartheta}'\in\partial B_n} \tilde{\boldsymbol{P}}_{\boldsymbol{\vartheta}'}(x^n) \leq \tilde{\boldsymbol{P}}_{\hat{\boldsymbol{\vartheta}}(x^n)}(x^n)\exp\left[-nR^-\delta^2\right]$$

with $R^- := \inf_{\vartheta' \in B_n} \left[\sum_{l=1}^k R_{lm}(\vartheta')S_{lm}(\vartheta') \right] /|\vartheta' - \hat{\vartheta}(x^n)|^2$ and $\delta := \inf_{\vartheta' \in \partial B_n}$ $|\vartheta' - \hat{\vartheta}(x^n)|$. Since the prior is universally lower-bounded by the Lebesgue measure, then (25) implies a bound for the denominator

$$\int_{B_n} \tilde{\boldsymbol{P}}_{\vartheta'}(x^n)d\tilde{\boldsymbol{Q}}(\vartheta') \geq w\tilde{\boldsymbol{P}}_{\hat{\vartheta}(x^n)}(x^n) \int_{|\boldsymbol{t}|<\delta} \exp\left[-nR^+|\boldsymbol{t}|^2\right] d\boldsymbol{t},$$

where $w > 0$ and $R^+ := \sup_{\vartheta' \in B_n} \left[\sum_{l=1}^k R_{lm}(\vartheta')S_{lm}(\vartheta') \right] /|\vartheta' - \hat{\vartheta}(x^n)|^2$. Hence we obtain an inequality for the ratio

$$\frac{\int_{\tilde{\boldsymbol{\Theta}} \setminus B_n} \tilde{\boldsymbol{P}}_{\vartheta}(x^n)d\tilde{\boldsymbol{Q}}(\vartheta)}{\int_{B_n} \tilde{\boldsymbol{P}}_{\vartheta}(x^n)d\tilde{\boldsymbol{Q}}(\vartheta)} \leq \frac{\sqrt{nR^+} \exp\left[-nR^-\delta^2/2\right]}{w \int_{|\boldsymbol{t}|<\delta\sqrt{nR^+}} \exp\left[-|\boldsymbol{t}|^2\right] d\boldsymbol{t}}.$$

The right-hand side tends to zero with $n \to \infty$ since $\delta = \Omega(n^{-1/2+\alpha})$ whereas R^+ and R^- tend to strictly positive constants by continuity and strictly positive definiteness of the Fisher information matrix. $\qquad\square$

4   Less Standard Examples

In this section we shall present less standard examples of statistical models. We begin with two very simple models.

Example 2 (the data are the parameter). Put $\boldsymbol{P}_{\boldsymbol{\theta}}(x^n) := \mathbf{1}_{\{x^n = \theta^n\}}$ for $\mathbb{X} = \mathbb{Y}$ and let $\boldsymbol{Q}(\theta) > 0$ for $\theta \in \mathbb{Y}^*$. This model is (n, m)-uniformly discretizable.

Example 3 (a singleton model). Each parameter $\boldsymbol{\theta}$ is random with respect to the prior $\boldsymbol{Q}$ concentrated on this parameter, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}\}$. The respective singleton model $(\boldsymbol{P}, \boldsymbol{Q})$ is $(0, 0)$-uniformly discretizable.

Now, a slightly more complex instance. Consider a class of stationary processes $(X_i)_{i \in \mathbb{Z}}$ of form $X_i := (K_i, \theta_{K_i})$, where the variables K_i are independent and distributed according to the hyperbolic distribution

$$P(K_i = k) = p(k) := \frac{k^{-1/\beta}}{\zeta(1/\beta)}, \quad k \in \mathbb{N}, \tag{27}$$

with a fixed $\beta \in (0, 1)$. This family of processes was introduced to model logical consistency of texts in natural language [19]. The distribution of variables X_i is equal to the measure $P(X_i \in \cdot) = \boldsymbol{P}_{\boldsymbol{\theta}}$ for the following Bayesian model.

Example 4 (an accessible description model). Put

$$\boldsymbol{P}_{\boldsymbol{\theta}}(x^n) := \prod_{i=1}^n p(k_i)\mathbf{1}_{\left\{z_i = \theta_{k_i}\right\}} \tag{28}$$

for $x_i = (k_i, z_i) \in \mathbb{N} \times \mathbb{Y}$ and let $\boldsymbol{Q}(\theta) > 0$ for $\theta \in \mathbb{Y}^*$.

For this model, Shannon information between the data and the parameter equals $\mathbf{E}_{(\boldsymbol{x},\boldsymbol{\theta})\sim\boldsymbol{T}}\left[-\log\boldsymbol{Y}(x^n)+\log\boldsymbol{P_\theta}(x^n)\right]=\Theta(n^\beta)$ asymptotically if $\boldsymbol{Q}(\theta)=D^{-|\theta|}$, cf. [19, Theorem 10]. As a consequence of the next statement, the accessible description model (28) is $(n^\upsilon,m^{1/\lambda})$-uniformly discretizable for

$$\upsilon > 2\beta/(1-\beta) \text{ and } \lambda < \beta.$$

Proposition 3. *For independent variables $(K_i)_{i\in\mathbb{Z}}$ with the distribution (27),*

$$\{K_1,K_2,...,K_n\}\setminus\{1,2,...,\lceil n^\upsilon\rceil\}=\emptyset, \tag{29}$$
$$\{1,2,...,\lfloor n^\lambda\rfloor\}\setminus\{K_1,K_2,...,K_n\}=\emptyset, \tag{30}$$

n-eventually almost surely.

Proof. To establish the first claim, put $U_n:=\lceil n^\upsilon\rceil$ and observe

$$P\left(\{K_1,K_2,...,K_n\}\setminus\{1,2,...,U_n\}\neq\emptyset\right)\leq\sum_{j=U_n+1}^\infty P(j\in\{K_1,K_2,...,K_n\})$$
$$=\sum_{j=U_n+1}^\infty 1-(1-p(j))^n\leq\sum_{j=U_n+1}^\infty np(j)$$
$$\leq\frac{n}{\zeta(1/\beta)}\int_{U_n}^\infty k^{-1/\beta}dk=\frac{n}{\zeta(1/\beta)}\frac{U_n^{1-1/\beta}}{1/\beta-1}\leq\frac{n^{-1-\epsilon}}{\zeta(1/\beta)(1/\beta-1)} \text{ for an } \epsilon>0.$$

Hence $\sum_{n=1}^\infty P\left(\{K_1,K_2,...,K_n\}\setminus\{1,2,...,U_n\}\neq\emptyset\right)<\infty$ so (29) holds by the Borel-Cantelli lemma. As for the second claim, put $L_n:=\lfloor n^\lambda\rfloor$ and observe

$$P\left(\{1,2,...,L_n\}\setminus\{K_1,K_2,...,K_n\}\neq\emptyset\right)\leq\sum_{j=1}^{L_n}P(j\notin\{K_1,K_2,...,K_n\})$$
$$=\sum_{j=1}^{L_n}(1-p(j))^n\leq L_n(1-p(L_n))^n=L_n\exp\left[n\log(1-p(L_n))\right]$$
$$\leq L_n\exp\left[-np(L_n)\right]\leq n^\beta\exp\left[-n^\epsilon\right] \text{ for an } \epsilon>0.$$

Hence $\sum_{n=1}^\infty P\left(\{1,2,...,L_n\}\setminus\{K_1,K_2,...,K_n\}\neq\emptyset\right)<\infty$ so (30) is also satisfied by the Borel-Cantelli lemma. $\square$

To use the above statement for the Bayesian model, notice first that $\boldsymbol{P_\theta}(x^n)>0$ for $\boldsymbol{P_\theta}$-almost all $\boldsymbol{x}$. Hence equalities $z_i=\theta_{k_i}$ and

$$\boldsymbol{T}(x^n,\theta^m)=\sum_{y^M\in\mathbb{Y}^M}\left(\prod_{i=1}^n p(k_i)\mathbf{1}_{\{z_i=y_{k_i}\}}\right)\left(\prod_{k=1}^m\mathbf{1}_{\{\theta_k=y_k\}}\right)\boldsymbol{Q}(y^M)$$
$$=\boldsymbol{P_\theta}(x^n)\sum_{y^M\in\mathbb{Y}^M}\left(\prod_{k\in\{k_1,k_2,...,k_n\}\cup\{1,2,...,m\}}\mathbf{1}_{\{\theta_k=y_k\}}\right)\boldsymbol{Q}(y^M)$$

hold for $\boldsymbol{P_\theta}$-almost all $\boldsymbol{x}$ with $M:=\max\{m,k_1,k_2,...,k_n\}$. Consequently,

$$\boldsymbol{Q}(\theta^m)\boldsymbol{P_\theta}(x^n)=\boldsymbol{T}(x^n,\theta^m) \quad\text{if}\quad \{k_1,k_2,...,k_n\}\setminus\{1,2,...,m\}=\emptyset, \tag{31}$$
$$\boldsymbol{T}(x^n,\theta^m)=\boldsymbol{Y}(x^n) \quad\text{if}\quad \{1,2,...,m\}\setminus\{k_1,k_2,...,k_n\}=\emptyset. \tag{32}$$

Thus the model given in Example 4 is $(n^\upsilon,m^{1/\lambda})$-uniformly discretizable.

The last example is not uniformly discretizable. It stems from the observation that any probability measure on $\mathbb{X}^\infty$ can be encoded with a single sequence from $\mathbb{Y}^\infty$. Such parameter is not identifiable, however.

Example 5 (a model that contains all distributions). For simplicity let $\mathbb{X} = \mathbb{N}$ and $\mathbb{Y} = \{0, 1\}$. The link between $\boldsymbol{\theta}$ and $\boldsymbol{P_\theta}$ will be established by imposing equalities $\boldsymbol{P_\theta}(\lambda) = 1$ and

$$\boldsymbol{P_\theta}(x^n) = \left(\boldsymbol{P_\theta}(x^{n-1}) - \sum_{y < x_n} \boldsymbol{P_\theta}(x^{n-1}y) \right) \cdot \sum_{k=1}^{\infty} \theta_{\phi(x^n, k)} 2^{-k}, \qquad (33)$$

where a recursive bijection $\phi : \mathbb{N}^+ \times \mathbb{N} \to \mathbb{N}$ is used. It is easy to see that $\boldsymbol{P_\theta}$ is a probability measure on $\mathbb{X}^\infty$ for each $\boldsymbol{\theta}$. Conversely, each probability measure on $\mathbb{X}^\infty$ equals $\boldsymbol{P_\theta}$ for at least one $\boldsymbol{\theta}$.

Let the prior be the uniform measure $\boldsymbol{Q}(\theta) := 2^{-|\theta|}$. Then the Bayesian measure $\boldsymbol{Y} = \int \boldsymbol{P_\theta}(x) d\boldsymbol{Q}(\boldsymbol{\theta})$ is recursive and equals

$$\boldsymbol{Y}(x^n) = \frac{1}{2} \left(\boldsymbol{Y}(x^{n-1}) - \sum_{y < x_n} \boldsymbol{Y}(x^{n-1}y) \right) \implies \log_2 \boldsymbol{Y}(x^n) = -\sum_{i=1}^{n} x_i.$$

Measure $\boldsymbol{Y}$ is not only optimal for all $\boldsymbol{Q}$-random $\boldsymbol{\theta}$, in the sense of $\boldsymbol{P_\theta}(\mathcal{L_Y}) = 1$, but it is also optimal for a certain $\boldsymbol{\theta} \notin \mathcal{L_Q}$ that satisfies $\boldsymbol{P_\theta} = \boldsymbol{Y}$. On the other hand, by the asymptotic equipartition property, $\boldsymbol{P_\theta}(\mathcal{L_Y}) = 0$ for stationary measures $\boldsymbol{P_\theta}$ that have a different entropy rate than $\boldsymbol{Y}$ [15, Section 15.7].

5 Countable Unions of Models

Bayesian mixtures of uniformly discretizable models are uniformly discretizable under the additional condition (34), which says that Bayesian model selection is consistent for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Let us write $\theta_k^m := \theta_k \theta_{k+1} ... \theta_m$. Moreover, define $\boldsymbol{T}^i$ and $\boldsymbol{Y}^i$ via (1)–(2) for models $(\boldsymbol{P}^i, \boldsymbol{Q}^i)$ substituted for $(\boldsymbol{P}, \boldsymbol{Q})$ respectively.

Theorem 4. *Let models $(\boldsymbol{P}^i, \boldsymbol{Q}^i)$ be (μ_i, ν_i)-uniformly discretizable with kernels $\boldsymbol{P}_{\boldsymbol{\theta}}^i(x)$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}^i$ and $i \in A$, a countable set. For a prefix code $c : A \to \mathbb{Y}^+$, put $\boldsymbol{\Theta} := \bigcup_{i \in A} c(i)\boldsymbol{\Theta}^i$. Consequently, denote $\mathrm{idx}(\boldsymbol{\theta}) := i$ and $\mathrm{trn}(\boldsymbol{\theta}) := \boldsymbol{\vartheta}$ for $\boldsymbol{\theta} = c(i)\boldsymbol{\vartheta} \in \boldsymbol{\Theta}$. Define the kernel $\boldsymbol{P_\theta}(x) := \boldsymbol{P}_{\mathrm{trn}(\boldsymbol{\theta})}^{\mathrm{idx}(\boldsymbol{\theta})}(x)$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and the prior $\boldsymbol{Q} := \sum_{i \in A} w(i)(\boldsymbol{Q}^i \circ \mathrm{trn})$ for $\sum_{i \in A} w(i) = 1$ and $w(i) > 0$. The model $(\boldsymbol{P}, \boldsymbol{Q})$ is (μ, ν)-uniformly discretizable provided*

$$\mu(n) := \sup_{i \in A}(|c(i)| + \mu_i(n)) < \infty,$$
$$\nu(m) := \sup_{i \in A} \nu_i(m - |c(i)|) < \infty,$$

and

$$\lim_{n \to \infty} \boldsymbol{Y}(x^n)/\boldsymbol{Y}^i(x^n) = w(i) \qquad (34)$$

for $i = \mathrm{idx}(\boldsymbol{\theta})$, $\boldsymbol{P_\theta}$-almost all $\boldsymbol{x}$, and all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Remark: Assuming recursive models and mutually singular $\boldsymbol{P}_{\boldsymbol{\vartheta}}^i$, convergence (34) may fail only for $\boldsymbol{\theta}$ that are not $\boldsymbol{Q}$-random, cf. [20]. Put $\mathcal{X} :=$

$\{\boldsymbol{x} : \lim_n \boldsymbol{Y}(x^n)/\boldsymbol{Y}^i(x^n) = w(i)\}$. By the ordinary martingale convergence, $\boldsymbol{Y}^i(\mathcal{X}) = 1$, whereas by convergence of recursive martingales [4, Theorem 3.1], $\mathcal{X} \supset \mathcal{L}_{\boldsymbol{Y}^i}$. Next, by [4, Cor. 4.3 & Thm. 5.3], we obtain $\mathcal{L}_{\boldsymbol{Y}^i} \supset \mathcal{L}_{\boldsymbol{P}^i|\vartheta}$ for $\boldsymbol{Q}^i$-random ϑ. Hence $\boldsymbol{P}_{\boldsymbol{\theta}}(\mathcal{X}) = 1$ if $\boldsymbol{\theta} \in \mathcal{L}_{\boldsymbol{Q}}$ and (35) holds true, in view of the Theorem 5 below.

Proof. Let $i = \mathrm{idx}(\boldsymbol{\theta})$. Observe that $\boldsymbol{T}(x^n, \theta^m) = w(i)\boldsymbol{T}^i(x^n, \theta^m_{|c(i)|+1})$ and $\boldsymbol{Q}(\theta^m) = w(i)\boldsymbol{Q}^i(\theta^m_{|c(i)|+1})$ if $m \geq |c(i)|$. Hence for $\boldsymbol{P}_{\boldsymbol{\theta}}$-almost all $\boldsymbol{x}$ and $m \geq \mu(n)$, we have

$$\left| \log \frac{\boldsymbol{Q}(\theta^m)\boldsymbol{P}_{\boldsymbol{\theta}}(x^n)}{\boldsymbol{T}(x^n, \theta^m)} \right| = \left| \log \frac{\boldsymbol{Q}^i(\theta^m_{|c(i)|+1})\boldsymbol{P}^i_{\mathrm{trn}(\boldsymbol{\theta})}(x^n)}{\boldsymbol{T}^i(x^n, \theta^m_{|c(i)|+1})} \right| = o(\log m).$$

On the other hand, for $\boldsymbol{P}_{\boldsymbol{\theta}}$-almost all $\boldsymbol{x}$ and $n \geq \nu(m)$,

$$\lim_{m \to \infty} \frac{\boldsymbol{T}(x^n, \theta^m)}{\boldsymbol{Y}(x^n)} = \lim_{m \to \infty} \left[\frac{w^i \boldsymbol{Y}^i(x^n)}{\boldsymbol{Y}(x^n)} \cdot \frac{\boldsymbol{T}^i(x^n, \theta^m_{|c(i)|+1})}{\boldsymbol{Y}^i(x^n)} \right] = 1.$$

$\square$

A complementary result says that the set of random parameters with respect to the mixture is the union of the respective sets for the combined models.

Theorem 5. *Consider the models from Theorem 4 and suppose that $\boldsymbol{Q}^i$ satisfy*

$$\boldsymbol{Q}^i(\theta^k)/\boldsymbol{Q}^i(\theta^m) \geq ac^{k-m} \tag{35}$$

for all $k \geq m \geq 0$ and certain constants $c < 1$ and $a > 0$. Then for $g(n) = \Omega(1)$ we have $\boldsymbol{\theta} \in \mathcal{L}_{\boldsymbol{Q},g(n)}$ if and only if $\mathrm{trn}(\boldsymbol{\theta}) \in \mathcal{L}_{\boldsymbol{Q}^{\mathrm{idx}(\boldsymbol{\theta})},g(n)}$.

Proof. Let $i = \mathrm{idx}(\boldsymbol{\theta})$. The claim is true if

$$\left| K(\theta^m) + \log \boldsymbol{Q}(\theta^m) - K(\theta^{|c(i)|+m}_{|c(i)|+1}) - \log \boldsymbol{Q}^i(\theta^{|c(i)|+m}_{|c(i)|+1}) \right| = O(1)$$

for $m \geq |c(i)|$. The latter condition is satisfied since $\left| K(\theta^m) - K(\theta^{|c(i)|+m}_{|c(i)|+1}) \right| \leq |c(i)| + O(1)$, whereas $\left| \log \boldsymbol{Q}(\theta^m) - \log \boldsymbol{Q}^i(\theta^{|c(i)|+m}_{|c(i)|+1}) \right| \leq |\log w(i)| + O(|c(i)|)$ by $\boldsymbol{Q}(\theta^m) = w(i)\boldsymbol{Q}^i(\theta^m_{|c(i)|+1})$ and (35). $\square$

These propositions may be useful when we seek a compressor that is optimal for all random and certain nonrandom parameters with respect to a given prior. A possible solution is to find priors against which the originally considered nonrandom parameters are random. Suppose that these priors and the original prior yield uniformly discretizable models and consistent Bayesian selection among these models is feasible. Then Theorems 2, 4, and 5 guarantee that the Bayesian mixture of all considered models achieves the best enumerable compression for all requested parameters and no so many others!

Acknowledgement

References

1. Vovk, V.G., V'yugin, V.V.: On the empirical validity of the Bayesian method. J. Roy. Statist. Soc. B 55, 253–266 (1993)
2. Vovk, V.G., V'yugin, V.V.: Prequential level of impossibility with some applications. J. Roy. Statist. Soc. B 56, 115–123 (1994)
3. Vitányi, P., Li, M.: Minimum description length induction, Bayesianism and Kolmogorov complexity. IEEE Trans. Inform. Theor. 46, 446–464 (2000)
4. Takahashi, H.: On a definition of random sequences with respect to conditional probability. Inform. Comput. 206, 1375–1382 (2008)
5. Gács, P.: On the symmetry of algorithmic information. Dokl. Akad. Nauk SSSR 15, 1477–1480 (1974)
6. Li, M., Vitányi, P.M.B.: An Introduction to Kolmogorov Complexity and Its Applications, 2nd edn. Springer, Heidelberg (1997)
7. van Lambalgen, M.: Random Sequences. PhD thesis, Universiteit van Amsterdam (1987)
8. Barron, A., Rissanen, J., Yu, B.: The minimum description length principle in coding and modeling. IEEE Trans. Inform. Theor. 44, 2743–2760 (1998)
9. Grünwald, P.D.: The Minimum Description Length Principle. MIT Press, Cambridge (2007)
10. Yu, B., Speed, T.P.: Data compression and histograms. Probab. Theor. Rel. Fields 92, 195–229 (1992)
11. Hopcroft, J.E., Ullman, J.D.: Introduction to Automata Theory, Languages and Computation. Addison-Wesley, Reading (1979)
12. Elias, P.: Universal codeword sets and representations for the integers. IEEE Trans. Inform. Theor. 21, 194–203 (1975)
13. Barron, A.R.: Logically Smooth Density Estimation. PhD thesis, Stanford University (1985)
14. Dawid, A.: Statistical theory: The prequential approach. J. Roy. Statist. Soc. A 147, 278–292 (1984)
15. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, Chichester (1991)
16. Li, L., Yu, B.: Iterated logarithmic expansions of the pathwise code lengths for exponential families. IEEE Trans. Inform. Theor. 46, 2683–2689 (2000)
17. Barndorff-Nielsen, O.E.: Information and Exponential Families. Wiley, Chichester (1978)
18. Jeffreys, H.: Theory of Probability, 3rd edn. Oxford University Press, Oxford (1961)
19. Dębowski, Ł.: On the vocabulary of grammar-based codes and the logical consistency of texts (2008) E-print, http://arxiv.org/abs/0810.3125
20. Csiszar, I., Shields, P.C.: The consistency of the BIC Markov order estimator. Ann. Statist. 28, 1601–1619 (2000)