Universal Coding on Infinite Alphabets: Exponentially Decreasing Envelopes

Dominique Bontemps

Universal Coding on Infinite Alphabets: Exponentially Decreasing Envelopes

Dominique Bontemps

***Abstract*—This paper deals with the problem of universal lossless coding on a countable infinite alphabet. It focuses on some classes of sources defined by an envelope condition on the marginal distribution, namely exponentially decreasing envelope classes with exponent α.**

The minimax redundancy of exponentially decreasing envelope classes is proved to be equivalent to $\frac{1}{4\alpha \log e} \log^2 n$. Then a coding strategy is proposed, with a Bayes redundancy equivalent to the maximin redundancy. At last, an adaptive algorithm is provided, whose redundancy is equivalent to the minimax redundancy.

***Index Terms*—Data compression, universal coding, infinite countable alphabets, redundancy, Bayes, adaptive compression.**

I. Introduction

COMPRESSION of data is broadly used in our daily life: from the movies we watch to the office documents we produce. In this article, we are interested in lossless source coding on an unknown alphabet. This has applications in areas such as language modeling or lossless multimedia codecs.

First, we present briefly the problematics of source coding. More details are available in general textbooks, like [1]. Then we make a short review of preceding results, in which we situate the topic of this article, exponentially decreasing envelope classes, and we announce our results.

A. Source coding

Consider a finite or countably infinite alphabet $\mathcal{X}$. A source on $\mathcal{X}$ is a probability distribution $\boldsymbol{P}$, on the set $\mathcal{X}^{\mathbb{N}}$ of infinite sequences of symbols from $\mathcal{X}$. Its marginal distributions are denoted by P^n, $n \geq 1$ (for $n = 1$, we only note P). The scope of lossless source coding is to encode a sequence of symbols $X_{1:n}$, generated according to P^n, into a sequence of bits as small as possible. The algorithm has to be uniquely decodable.

The binary entropy $H(P^n) = \mathbb{E}_{P^n}[-\log_2 P^n(X_{1:n})]$ is known to be a lower bound for the expected codelength of $X_{1:n}$. From now on, $\log$ denotes the logarithm taken to base 2, while $\ln$ is used to denote the natural logarithm. Since arithmetic coding based on P^n encodes a message $x_{1:n}$ with $\lceil -\log P^n(x_{1:n}) \rceil + 1$ bits, this lower bound can be achieved within two bits. Then, the expected redundancy measures the mean number of extra bits, in addition to the entropy, a coding strategy uses to encode X^n. In the sequel, we use the word *redundancy* instead of *expected redundancy*.

Furthermore, together with Kraft-McMillan inequality, arithmetic coding provides an almost perfect correspondence between coding algorithms and probability distributions on $\mathcal{X}^n$. In this setting, if an algorithm is associated to the probability distribution Q^n, its expected redundancy reduces to the Kullbach-Liebler divergence between P^n and Q^n

$$D(P^n; Q^n) = \mathbb{E}_{P^n}\left[\log \frac{P^n(X_{1:n})}{Q^n(X_{1:n})}\right].$$

We call this quantity (expected) redundancy of the distribution Q^n (with respect to P^n).

Unfortunately, the true statistics of the source are not known in general, but P^n is supposed to belong to some large class of sources Λ (for instance, the class of all iid sources, or the class of Markov sources). In this paper, the maximum redundancy

$$R_n(Q^n; \Lambda) = \sup_{\boldsymbol{P} \in \Lambda} R_n(Q^n; P^n)$$

measures how well a coding probability Q^n behave on an entire class Λ. With this point of view, the best coding probability is a *minimax* coding probability, that achieves the *minimax redundancy*

$$R_n(\Lambda) = \inf_{Q^n} R_n(Q^n; \Lambda).$$

Another way to measure the ability of a class of sources to be efficiently encoded is the *Bayes redundancy*

$$R_{n,\mu}(\Lambda) = \inf_{Q^n} \int_\Lambda R_n(Q^n; P^n)\, \mathrm{d}\mu(\boldsymbol{P})$$

where μ is a prior distribution on Λ endowed with the topology of weak convergence and the Borel σ-field. Only one coding strategy achieves the Bayes redundancy: the Bayes mixture

$$M_{n,\mu}(x_{1:n}) = \int_{\Theta_{C,\alpha}} P^n_{\boldsymbol{\theta}}(x_{1:n})\, \mathrm{d}\mu(\boldsymbol{\theta}).$$

When Λ is a class of iid sources on the set $\mathcal{X} = \mathbb{N}_+$ of positive integers, there is a natural parametrization of Λ by $P_{\boldsymbol{\theta}}(j) = \theta_j$, with $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots) \in \Theta_\Lambda$. Θ_Λ is then a subset of

$$\Theta = \left\{ \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots) \in [0,1]^{\mathbb{N}} : \sum_{i \geq 1} \theta_i = 1 \right\}$$

and it is endowed with the topology of pointwise convergence. In this case we write μ as a prior on Θ_Λ.

Minimax redundancy and Bayes redundancy are linked by an important relation [2], [3]; it is written here in the context of iid sources on a finite or countably infinite alphabet, but Haussler [4] has shown that it can be generalized for all classes of stationary ergodic processes on a complete separable metric space.

D. Bontemps is PhD fellowship holder at Laboratoire de Mathématiques, Univ. Paris-Sud 11, Orsay, 91405 France.

Theorem 1: Let Λ be a class of iid sources, such that the parameter set Θ_Λ is a measurable subset of Θ. Let $n \geq 1$. Then

$$R_n(\Lambda) = \sup_\mu R_{n,\mu}(\Lambda),$$

where the supremum is taken over all (Borel) probability measures on Θ_Λ.

The quantity $\sup_\mu R_{n,\mu}(\Lambda)$ is called *maximin redundancy*. A prior whose Bayes redundancy corresponds to the maximin redundancy is said to be maximin, or least favorable.

Theorem 1 says that maximin redundancy and minimax redundancy are the same. It provides a tool to calculate the minimax redundancy.

Before speaking about known results, let us make mention of other two notions.

With an asymptotic point of view, a sequence of coding probabilities $(Q_n)_{n\geq 1}$ is said to be weakly universal if the per-symbol redundancy tends to 0 on Λ: $\sup_{\boldsymbol{P}\in\Lambda} \lim_{n\to\infty} \frac{1}{n} D(P^n; Q^n) = 0$.

Instead of the expected redundancy, many authors consider individual sequences. In this case, the *minimax regret*

$$R_n^*(\Lambda) = \inf_{Q^n} \sup_{\boldsymbol{P}\in\Lambda} \sup_{x_{1:n}\in\mathcal{X}^n} \log \frac{P^n(x_{1:n})}{Q^n(x_{1:n})}$$

plays the role that the minimax redundancy plays with the expected redundancy.

B. Exponentially decreasing envelope classes

In the case of a finite alphabet of size k, many classes of sources have been studied in the literature, for which estimates of the redundancy have been provided. In particular we have the class of all iid sources (see [5]–[10], and references therein), whose minimax redundancy is

$$\frac{k-1}{2} \log \frac{n}{2\pi e} + \log \frac{\Gamma(1/2)^k}{\Gamma(k/2)} + o(1).$$

This last class can be seen as a particular case of a $(k-1)$-dimensional class of iid sources on a (possibly) bigger alphabet, for which we have a similar result under certain conditions (see [11]–[13]). Similar results are still available for classes of Markov processes and finite memory tree sources on a finite alphabet (see [5], [14]–[16]), and for k-dimensional classes of even non-iid sources on an arbitrary alphabet (see [17]).

The results become less precise when one considers infinite dimensional classes on a finite alphabet. A typical example is the class of renewal processes, for which we do not have an equivalent of the expected redundancy, but we know that it is lower and upper bounded by a constant times $\sqrt{n}$ (see [18], [19]).

Eventually, it is well known that the class of stationary ergodic sources is weakly universal (see [1]). However, Shields [20] showed that this class does not admit non-trivial universal redundancy rates.

In the case of a countably infinite alphabet, the situation is significantly different. Even the class of all iid sources is not weakly universal (see [21], [22]). Kieffer characterized weakly universal classes in [21] (see also [22], [23]):

Proposition 1: A class Λ of stationary sources on $\mathbb{N}_+$ is weakly universal if and only if there exists a probability distribution Q on $\mathbb{N}_+$ such that for every $\boldsymbol{P} \in \Lambda$, $D(P;Q) < \infty$.

In the literature, we find two main ways to deal with infinite alphabets. The first one [24]–[32] separates the message into two parts: a description of the symbols appearing in the message, and the *pattern* they form. Then the compression of patterns is studied.

A second approach [23], [33]–[36] studies collections of sources satisfying Kieffer's condition, and proposes compression algorithms for these classes. A result from [36] indicates us such a way:

Proposition 2: Let Λ be a class of iid sources over $\mathbb{N}_+$. Let $\hat{p}$ be defined by $\hat{p}(x) = \sup_{\boldsymbol{P}\in\Lambda} P(x)$. Then the minimax regret verifies

$$R_n^*(\Lambda) < \infty \Leftrightarrow \sum_{x\in\mathbb{N}_+} \hat{p}(x) < \infty.$$

It is therefore quite natural to consider classes of iid sources with envelope conditions on the marginal distribution. In this article we study specific classes of iid sources introduced by [36], and called *exponentially decreasing envelope classes*.

Definition 1: Let C and α be positive numbers satisfying $C > e^{2\alpha}$. The exponentially decreasing envelope class $\Lambda_{Ce^{-\alpha\cdot}}$ is the class of sources defined by

$$\Lambda_{Ce^{-\alpha\cdot}} = \{\boldsymbol{P} : \forall k \geq 1,\ P(k) \leq Ce^{-\alpha k} \text{ and } \boldsymbol{P} \text{ is stationary and memoryless.}\}$$

The first condition addresses mainly the queue of the distribution of X_1; it means that great numbers must be rare enough. It does not mean that the distribution is geometrical: if C is big enough, many other distributions are possible. Furthermore we will see that the exact value of C does not change significantly the minimax redundancy, unlike α.

Since in this paper we are going to only talk about exponentially decreasing envelope classes, we simplify the notations $R_n(Q^n; \Lambda_{Ce^{-\alpha\cdot}})$, $R_n(\Lambda_{Ce^{-\alpha\cdot}})$, and $R_{n,\mu}(\Lambda_{Ce^{-\alpha\cdot}})$ into $R_n(Q^n; C, \alpha)$, $R_n(C,\alpha)$, and $R_{n,\mu}(C,\alpha)$ respectively. The subset of Θ corresponding to $\Lambda_{Ce^{-\alpha\cdot}}$ is denoted by

$$\Theta_{C,\alpha} = \{\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots) \in [0,1]^{\mathbb{N}} : \sum_{i\geq 1} \theta_i = 1 \text{ and } \forall i \geq 1,\ \theta_i \leq Ce^{-\alpha i}\}. \quad (1)$$

We obtain three main results about these classes.

In Section II we calculate the minimax redundancy of exponentially decreasing envelope classes, and we find that it is equivalent to $\frac{1}{4\alpha \log e} \log^2 n$ as n tends to the infinity. This rate is interesting for two main reasons. Up to our knewledge, exponentially decreasing envelope classes are the first family of classes on an infinite alphabet for which an equivalent of the minimax redundancy is known. Then, even the rate is new: until now only rates in $\log n$ or in $\sqrt{n}$ have been obtained.

Once the minimax redundancy of a class of sources is known, there are two main approaches to find an optimal universal code: either you look for a maximin Bayes prior, or for a minimax algorithm. These two issues are studied here.

In Section III we are concerned with the problem of finding a maximin Bayes prior. We construct a sequence of Bayes

priors whose Bayes redundancy is equivalent to the maximin redundancy, as the length n of the message tends to the infinity.

Then we are interested in finding a minimax coding algorithm. Section IV proposes a new adaptive coding algorithm, and in Section V we show that the maximum redundancy of this new algorithm is equivalent to the minimax redundancy of exponentially decreasing envelope classes.

Eventually, the Appendix contains some lemmas and corollaries used in the main analysis.

II. Minimax redundancy

In this section we state our main result. Theorem 2 below gives an equivalent of the minimax redundancy of exponentially decreasing envelope classes. To get it, we use a result due to Haussler and Opper [37].

Theorem 2: Let C and α be positive numbers satisfying $C > e^{2\alpha}$. The minimax redundancy of the exponentially decreasing envelope class $\Lambda_{Ce^{-\alpha\cdot}}$ verifies

$$R_n(C,\alpha) \underset{n\to\infty}{\sim} \frac{1}{4\alpha\log e}\log^2 n.$$

Theorem 2 improves on a preceding result of [36, Theorem 7]. In that article the following bounds of the minimax redundancy of exponentially decreasing envelope classes are given:

$$\begin{aligned}\frac{1}{8\alpha\log e}\log^2 n\,(1+o(1)) &\le R_n(C,\alpha)\\ &\le \frac{1}{2\alpha\log e}\log^2 n + O(1).\end{aligned}$$

In subsection II-A we outline the work done in [37], and then we use it in subsection II-B to prove Theorem 2. Eventually, we discuss in subsection II-C the adaptation of this method to other envelope classes.

A. From metric entropy to minimax redundancy

To study the redundancy of a class of sources, [37] considers the Hellinger distance between the first marginal distributions of each source. Bounds on the minimax redundancy are provided in terms of the metric entropy of the set of the first marginal distributions, with respect to the Hellinger distance. As a consequence, that method can be applied only to iid sources. However it is very efficient in the case of exponentially decreasing envelope classes.

First, we need to define the Hellinger distance and the metric entropy. In the case of sources on a countably infinite alphabet, the Hellinger distance can be defined in the following way:

Definition 2: Let P and Q two probability distributions on $\mathbb{N}_+$. Then the Hellinger distance between P and Q is defined by

$$h(P,Q) = \sqrt{\sum_{k\ge1}\left(\sqrt{P(k)}-\sqrt{Q(k)}\right)^2}.$$

A related metric can be defined on the parameter set Θ:

$$d(\boldsymbol{\theta},\boldsymbol{\theta}') = h(P_{\boldsymbol{\theta}},P_{\boldsymbol{\theta}'}) = \sqrt{\sum_{k\ge1}\left(\sqrt{\theta_k}-\sqrt{\theta'_k}\right)^2}.$$

From a metric we can define the *metric entropy* of a set. We need to define first some numbers.

Definition 3: Let (S,ρ) be any complete separable metric space. Let $\epsilon>0$ be a positive number.

1) We denote by $\mathcal{D}_\epsilon(S,\rho)$ the cardinality of the smallest finite partition of S with sets of diameter at most ϵ, or we set $\mathcal{D}_\epsilon(S,\rho)=\infty$ if no such finite partition exists.
2) The metric entropy of (S,ρ) is defined by
$$\mathcal{H}_\epsilon(S,\rho) = \ln \mathcal{D}_\epsilon(S,\rho).^{1}$$
3) An *ϵ-cover* of S is a subset $A\subset S$ such that, for all x in S, there is an element y of A with $\rho(x,y)<\epsilon$. The *covering number* $\mathcal{N}_\epsilon(S,\rho)$ is the cardinality of the smallest finite ϵ-cover of S, or we define $\mathcal{N}_\epsilon(S,\rho)=\infty$ if no finite ϵ-cover exists.
4) An *ϵ-separated subset* of S is a subset $A\subset S$ such that, for all distinct x, y in A, $\rho(x,y)>\epsilon$. The *packing number* $\mathcal{M}_\epsilon(S,\rho)$ is the cardinality of the largest finite ϵ-separated subset of S, or we define $\mathcal{M}_\epsilon(S,\rho)=\infty$ if arbitrary large ϵ-separated subsets exist.

The following lemma explains how these numbers are linked. It is a classical result that can be found for instance in [38].

Lemma 1: Let (S,ρ) be any complete separable metric space. For all $\epsilon>0$,

$$\mathcal{M}_{2\epsilon}(S,\rho)\le\mathcal{D}_{2\epsilon}(S,\rho)\le\mathcal{N}_\epsilon(S,\rho)\le\mathcal{M}_\epsilon(S,\rho).$$

Lemma 1 enables us to choose the most convenient number to calculate the metric entropy.

From the metric entropy one can define the notion of metric dimension, which generalises the classical notion of dimension. But the metric entropy lets us know in some way how dense are the elements in a set, even infinite dimensional.

Another quantity that [37] uses is the *minimax risk for the $(1+\lambda)$-affinity*

$$R^{\text{minimax}}_{1,\rho_{1+\lambda}}(\Lambda) = \inf_Q \sup_{\boldsymbol{\theta}\in\Theta_\lambda}\sum_{k\ge1}P_{\boldsymbol{\theta}}(k)^{1+\lambda}Q(k)^{-\lambda},$$

defined for all $\lambda>0$.
More precisions about the $(1+\lambda)$-affinity are given in [37]. See also [39] for a special regard payed to envelope classes.

In the case of an envelope class Λ_f defined by an integrable envelope function f, it is easy to see that $R^{\text{minimax}}_{1,\rho_{1+\lambda}}<\infty$ for all $\lambda>0$. Indeed the choice

$$Q(k) = \frac{f(k)}{\sum_{l\ge1}f(l)}$$

leads to the relation

$$R^{\text{minimax}}_{1,\rho_{1+\lambda}} \le \left(\sum_{k\ge1}f(k)\right)^\lambda.$$

[1] We follow [37] in this definition of the metric entropy. Several authors use a slightly different definition, based on the covering number or the packing number.

We can now write a slightly modified version [2] of Theorem 5 of [37] in the context of data compression on an infinite alphabet.

Theorem 3: Let Λ be a class of iid sources on $\mathbb{N}_+$, such that the parameter set Θ_Λ is a measurable subset of Θ. Assume that there exists $\lambda > 0$ such that $R^{\text{minimax}}_{1,\rho_{1+\lambda}} < \infty$. Let $h(x)$ be a continuous, nondecreasing function defined on the positive reals such that, for all $\gamma \geq 0$ and $C > 0$,

1)
$$\lim_{x\to\infty} \frac{h\left(Cx\left(h(x)\right)^\gamma\right)}{h(x)} = 1$$
and
2)
$$\lim_{x\to\infty} \frac{h\left(Cx(\ln x)^\gamma\right)}{h(x)} = 1.$$

Then

1) If
$$\mathcal{H}_\epsilon(\Theta_\Lambda, d) \underset{\epsilon\to 0}{\sim} h\left(\frac{1}{\epsilon}\right),$$
then
$$R_n(\Lambda) \underset{n\to\infty}{\sim} (\log e)\, h(\sqrt{n}).\ ^3$$
2) If, for some $\alpha > 0$ and $c > 0$,
$$\liminf_{\epsilon\to 0} \frac{\mathcal{H}_\epsilon(\Theta_\Lambda, d)}{(1/\epsilon)^\alpha h(1/\epsilon)} \geq c,$$
then
$$\liminf_{n\to\infty} \frac{R_n(\Lambda)}{n^{\alpha/(\alpha+2)}\left[h(n^{1/(\alpha+2)})\right]^{2/(\alpha+2)}} > 0.$$
3) If, for some $\alpha > 0$ and $C > 0$,
$$\limsup_{\epsilon\to 0} \frac{\mathcal{H}_\epsilon(\Theta_\Lambda, d)}{(1/\epsilon)^\alpha h(1/\epsilon)} \leq C,$$
then
$$\limsup_{n\to\infty} \frac{R_n(\Lambda)}{(n\ln n)^{\alpha/(\alpha+2)}\left[h(n^{1/(\alpha+2)})\right]^{2/(\alpha+2)}} < \infty.$$

The conditions concerning the function h mean that h cannot grow too fast. For instance, h can grow like $C(\ln x)^\beta$, with $\beta \geq 0$.

The first case in the theorem corresponds to the one we use for exponentially decreasing envelope classes. In this case, the fast decreasing envelope produces a "not too big" metric entropy. The theorem gives us an equivalent of the minimax redundancy of the class of sources when n goes to the infinity. This turns out very useful, as it improves a preceding result of [36]. However it is only an asymptotic result, without any convergence speed.

The second and the third items correspond to bigger classes of sources. In these cases the result is a bit less interesting: it gives a speed for the growth of the redundancy, but without the associated constant factor. Furthermore there is a gap of $(\ln n)^{\alpha/(\alpha+2)}$ beetween the lower bound of point 2 and the upper bound of point 3. However it allows us to retrieve more or less a result of [36] for another type of envelope classes.

We develop now these applications.

[2]The separation of the upper and lower bounds have no effect on the proof given by Haussler and Opper. A complete justification is available in [39].

[3]The $(\log e)$ factor comes from the use of the logarithm with basis 2 in the definition of R_n.

B. The minimax redundancy of exponentially decreasing envelope classes

Here we want to prove Theorem 2 by applying Theorem 3. Thus we have to calculate the metric entropy of exponentially decreasing envelope classes. This is done by Proposition 3:

Proposition 3: Let C and α be positive numbers satisfying $C > e^{2\alpha}$. The metric entropy of the parameters set $\Theta_{C,\alpha}$ verifies
$$\mathcal{H}_\epsilon(\Theta_{C,\alpha}, d) = (1 + o(1))\frac{1}{\alpha}\ln^2(1/\epsilon),$$
where $o(1)$ is a function $g(\epsilon)$ such that $g(\epsilon) \to 0$ as $\epsilon \to 0$.

Proof of Theorem 2: Just apply Theorem 3, with $h(x) = \frac{1}{\alpha}\ln^2(x)$, to get the result. ∎

Proof of Proposition 3: We start with general considerations. Let Λ_f be the envelope class defined by the integrable envelope function f. Let Θ_f be the corresponding parameter set
$$\Theta_f = \{\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots) \in [0,1]^{\mathbb{N}} : \sum_{i\geq 1} \theta_i = 1 \text{ and } \forall i \geq 1,\ \theta_i \leq f(i)\}.$$

The function $\boldsymbol{\theta} \mapsto (\sqrt{\theta_1}, \sqrt{\theta_2}, \ldots)$ is an isometry between the metric space (Θ_f, d) and the subset $A_f \cap \{\|x\| = 1\}$ of ℓ^2, equipped with the classical euclidean norm $\|\cdot\|$, where A_f is defined by
$$A_f = \{(x_k)_{k\in\mathbb{N}^*} \in \ell^2 :\ \forall k \in \mathbb{N}^*,\ 0 \leq x_k \leq \sqrt{f(k)}\}. \quad (2)$$
The metric entropy of (Θ_f, d) can be calculated in this space.

The next idea we apply is to truncate some coordinates, to work in a finite dimensional space instead of ℓ^2. Together with an adequate use of Lemma 1, this helps us to obtain upper and lower bounds for the metric entropy of (Θ_f, d). We start with the upper bound.

Lemma 2: Let Λ_f be the envelope class defined by the integrable envelope function f, and let $\epsilon > 0$ be a positive number. Let N_ϵ denote the integer
$$N_\epsilon = \inf\left\{n \geq 1 : \sum_{k\geq n+1} f(k) \leq \frac{\epsilon^2}{16}\right\}.$$

Then
$$\mathcal{H}_\epsilon(\Theta_f, d) \leq N_\epsilon \ln(1/\epsilon) + 3N_\epsilon \ln 2 + A(N_\epsilon) + B(\epsilon),$$
where
$$A(N) = -\ln \mathrm{Vol}\left(B_{\mathbb{R}^N}(0,1)\right) = \ln \frac{\Gamma(\frac{N}{2}+1)}{\pi^{\frac{N}{2}}}$$
and
$$B(\epsilon) = \sum_{k=1}^{N_\epsilon} \ln\left(\sqrt{f(k)} + \frac{\epsilon}{4}\right).$$

Furthermore

$$A(N_\epsilon) \underset{\epsilon\to 0}{\sim} \frac{N_\epsilon}{2}\ln N_\epsilon.$$

Note that

$$-N_\epsilon \ln(1/\epsilon) - 2N_\epsilon \ln 2 \le B(\epsilon) \le \frac{\epsilon}{4}N_\epsilon.$$

These bounds on $B(\epsilon)$ show that $B(\epsilon)$ tends to decrease the upper bound, while $A(N_\epsilon)$ contributes to its growth. If $\ln N_\epsilon$ behaves like $\ln(1/\epsilon)$ up to a constant factor, then the upper bound given in Lemma 2 corresponds to a constant times $N_\epsilon \ln N_\epsilon$, and we are concerned with the point 3 of Theorem 3.

We can apply Lemma 2 to the case of exponentially decreasing envelope classes; these two results are proved in Appendix A.

Corollary 1: Let C and α be positive numbers satisfying $C > e^{2\alpha}$. The metric entropy of the parameter set $\Theta_{C,\alpha}$ defined by (1) verifies

$$\mathcal{H}_\epsilon(\Theta_{C,\alpha}, d) \le (1+o(1))\frac{1}{\alpha}\ln^2(1/\epsilon),$$

where $o(1)$ is a function $g(\epsilon)$ such that $g(\epsilon)\to 0$ as $\epsilon \to 0$.

Now, after we have obtained an upper bound on the metric entropy, we need to get a lower bound. In this case too, we want to truncate some coordinates to bring ourselves to a smaller finite dimensional space. This time we truncate the first coordinates. Let us consider the number

$$l_f = \min\{l \ge 0 : \sum_{k\ge l+1} f(k) \le 1\}.$$

Lemma 3: Let Λ_f be the envelope class defined by an integrable envelope function f, which verifies

$$\sum_{k\ge 1} f(k) \ge 2.$$

Let $\epsilon > 0$ be a positive number, and let $m \ge 1$ be an integer. Then

$$\mathcal{H}_\epsilon(\Theta_f, d) \ge \frac{1}{2}\sum_{k=l_f+1}^{l_f+m} \ln f(k) + m\ln\left(\frac{1}{\epsilon}\right) + A(m),$$

where $A(m)$ is defined as in Lemma 2:

$$A(m) = -\ln \mathrm{Vol}\left(B_{\mathbb{R}^m}(0,1)\right) \underset{m\to\infty}{\sim} \frac{m}{2}\ln m.$$

Note that exponentially decreasing envelopes verify the condition $\sum_{k\ge 1} f(k) \ge 2$. Indeed the envelope of exponentially decreasing envelope classes is

$$f(k) = \min(1, Ce^{-\alpha k}),$$

and the condition $C > e^{2\alpha}$ entails that $f(1) = f(2) = 1$.

From Lemma 3 we can infer the following corollary, to be proved in Appendix A together with Lemma 3.

Corollary 2: Let C and α be positive numbers satisfying $C > e^{2\alpha}$. The metric entropy of the parameters set $\Theta_{C,\alpha}$ verifies

$$\mathcal{H}_\epsilon(\Theta_{C,\alpha}, d) \ge (1+o(1))\frac{1}{\alpha}\ln^2(1/\epsilon),$$

where $o(1)$ is a function $g(\epsilon)$ such that $g(\epsilon)\to 0$ as $\epsilon \to 0$. Note that the bound is the same as in Corollary 2. Therefore this concludes the proof of Proposition 3. ∎

C. What about other envelope classes?

In [36] the redundancy of another type of envelope classes is also studied. The *power-law envelope class* $\Lambda_{C\cdot^{-\alpha}}$ is defined, for $C > 1$ and $\alpha > 1$, by the envelope function $f_{\alpha,C}(x) = \min(1, \frac{C}{x^\alpha})$. The bounds obtained in [36, Theorem 6] are

$$\begin{aligned} A(\alpha) n^{1/\alpha} \log\lfloor C\zeta(\alpha)\rfloor & \\ \le \mathbb{R}_n(\Lambda_{C\cdot^{-\alpha}}) & \\ \le \left(\frac{2Cn}{\alpha-1}\right)^{1/\alpha} (\log n)^{1-1/\alpha} + O(1), & \end{aligned} \tag{3}$$

where

$$A(\alpha) = \frac{1}{\alpha}\int_1^\infty \frac{1-e^{-1/(\zeta(\alpha)u)}}{u^{1-1/\alpha}}\mathrm{d}u,$$

and ζ denotes the classical function $\zeta(\alpha) = \sum_{k\ge 1}\frac{1}{k^\alpha}$, for $\alpha > 1$.

If one adapts the calculus made earlier to the power-law envelope classes, one can get the following upper and lower bounds:
There are two (calculable) constants $K_1, K_2 > 0$ such that, for all $\epsilon > 0$,

$$K_1\left(\frac{1}{\epsilon}\right)^{\frac{2}{\alpha-1}} \le \mathcal{H}_\epsilon \le K_2(1+o(1))\left(\frac{1}{\epsilon}\right)^{\frac{2}{\alpha-1}} \ln\left(\frac{1}{\epsilon}\right).$$

Unfortunately this formula leaves a gap between the lower bound and the upper bound. The application of Theorem 3 makes the gap worse. Indeed the polynomial part $\left(\frac{1}{\epsilon}\right)^{\frac{2}{\alpha-1}}$ of the metric entropy causes an additional gap of $\log^{1/\alpha} n$. In practice the bounds are the following:
There are two (unknown) constants $C, c > 0$ such that, for all $n \ge 1$,

$$c(1+o(1))n^{1/\alpha} \le R_n(\Lambda_{C\cdot^{-\alpha}}) \le C(1+o(1))n^{1/\alpha}\log n. \tag{4}$$

These inequalities improve in no way the result of [36]. May a better calculus of the metric entropy improve either their lower bound or their upper bound? Anyway the metric entropy of power-law envelope classes is "too big" to efficiently apply Theorem 3: it does not leave the hope for an equivalence, as for exponentially decreasing envelope classes. To summarize, the strategy based on the metric entropy and Theorem 3 turns out efficient for "small" classes of sources.

III. Dirichlet's prior

Theorem 1 gives a way to calculate a lower bound of the minimax redundancy of a class of sources. Indeed the minimax redundancy is lower bounded by the Bayes redundancy of any prior. In this context, the choice of an appropriate prior is a relevant matter.

In this section a sequence of priors μ_k is constructed; μ_k is the Dirichlet prior on a finite set of coordinates, supported by the envelope class and normalized (see below the exact definition). With an appropriate choice of k depending on n, priors μ_k are "almost" asymptotically least favorable for the exponentially decreasing envelope classes: their Bayes redundancy is equivalent, as n tends to the infinity, to the minimax redundancy.

Theorem 4: Let C and α be positive numbers satisfying $C > e^{2\alpha}$. Let $k_n = \left\lfloor \frac{1}{\alpha \log e} \log n \right\rfloor$. Then the sequence of priors μ_{k_n} verifies

$$R_{n,\mu_{k_n}}(C,\alpha) \geq (1+o(1)) \frac{1}{4\alpha \log e} \log^2 n,$$

where $o(1)$ is a function $g(n)$ such that $g(n) \to 0$ as $n \to \infty$. Additionally, Theorem 4 enables us to retrieve the lower bound of the minimax redundancy obtained in the section II.

What about other envelope classes? For power-law envelope classes the choice $k_n = \left\lfloor \frac{n^{-1/\alpha}}{e} \right\rfloor$ gives

$$R_n(\Lambda_{C^{\cdot -\alpha}}) \geq R_{n,\mu_{k_n}}(\Lambda_{C^{\cdot -\alpha}}) \geq (1+o(1)) \frac{\alpha}{2e} n^{1/\alpha}.$$

This result is similar to those presented in (3) and (4), and that is good. However it does not permit to fill the gap between this lower bound and the upper bound given in (3).

In the context of coding on a finite alphabet, the Bayes strategy using *Jeffreys' prior* plays a significant role. In the important case of the class of all iid sources on a finite alphabet, the Jeffrey's prior is the Dirichlet$(1/2, 1/2, \ldots, 1/2)$ prior. [12] proves that "Jeffrey's prior is asymptotically least favorable" under some conditions. [6] goes further and shows that Dirichlet's prior is asymptotically maximin but not asymptotically minimax. Then [7] proposes an asymptotically minimax modification of the Dirichlet prior.

First, priors μ_k are constructed in subsection III-A. Then subsection III-B contains the proof of Theorem 4.

A. A modification of Dirichlet's prior

First of all, we need to properly define Dirichlet's prior on the class of all iid sources on a finite alphabet.

An iid source $\boldsymbol{P}$ on the alphabet $\{1, 2, \ldots, k\}$ is characterized by the statistics of its first marginal distribution $P(i)$, $1 \leq i \leq k$. The class of all iid sources on this alphabet can be written $(\boldsymbol{P_\theta})$, where the parameter $\boldsymbol{\theta}$ is an element of the simplex of $\mathbb{R}^k$

$$S_k = \{(\theta_1, \theta_2, \ldots, \theta_k) \in [0,1]^k : \sum_{1 \leq i \leq k} \theta_i = 1\},$$

and $P_{\boldsymbol{\theta}}(i) = \theta_i$.

An equivalent notation of the simplex is obtained by setting $\theta_1 = 1 - \sum_{2 \leq i \leq k} \theta_i$, with $(\theta_2, \ldots, \theta_k)$ an element of the set

$$S'_k = \{(\theta_2, \ldots, \theta_k) \in [0,1]^{k-1} : \sum_{2 \leq i \leq k} \theta_i \leq 1\}.$$

This makes easier to define the Lebesgue measure $\mathrm{d}\boldsymbol{\theta}$ on the simplex of $\mathbb{R}^k$, by restriction of the measure on $[0,1]^{k-1}$.

If we consider the sequence $X_{1:n}$ of the first n symbols produced by a source $\boldsymbol{P_\theta}$, let T_i denote the number of occurrences of symbol i

$$T_i = \sum_{j=1}^{n} \mathbb{1}_{X_j = a_i} \quad \text{for all } 1 \leq i \leq k,$$

where

$$\mathbb{1}_{X_j = a_i} = \begin{cases} 1 & \text{if } X_j = a_i, \\ 0 & \text{otherwise.} \end{cases}$$

The same notation $\mathbb{1}_{\text{"some condition"}}$ will be used in the sequel.

Then, the probability of the sequence $X_{1:n}$ under the distribution $\boldsymbol{P_\theta}$ is

$$P^n_{\boldsymbol{\theta}}(X_{1:n}) = \theta_1^{T_1} \theta_2^{T_2} \cdots \theta_k^{T_k}. \tag{5}$$

Dirichlet's prior has a density proportional to $\theta_1^{-1/2} \theta_2^{-1/2} \cdots \theta_k^{-1/2}$ with respect to the Lebesgue measure. The associated Bayes mixture, also called *Krichevsky-Trofimov* mixture, is

$$\begin{aligned} KT_k(X_{1:n}) &= \frac{\int_{S'_k} \theta_1^{T_1 - 1/2} \theta_2^{T_2 - 1/2} \cdots \theta_k^{T_k - 1/2} \mathrm{d}\boldsymbol{\theta}}{\int_{S'_k} \theta_1^{-1/2} \theta_2^{-1/2} \cdots \theta_k^{-1/2} \mathrm{d}\boldsymbol{\theta}} \\ &= \frac{D_k(T_1 + \frac{1}{2}, \cdots, T_k + \frac{1}{2})}{D_k(\frac{1}{2}, \cdots, \frac{1}{2})}, \end{aligned} \tag{6}$$

where $D_k(\cdot, \ldots, \cdot)$ denotes the Dirichlet integrals

$$\begin{aligned} D_k(\lambda_1, \ldots, \lambda_k) &= \int_{S'_k} \theta_1^{\lambda_1 - 1} \theta_2^{\lambda_2 - 1} \cdots \theta_k^{\lambda_k - 1} \mathrm{d}\boldsymbol{\theta} \\ &= \frac{\Gamma(\lambda_1)\Gamma(\lambda_2) \cdots \Gamma(\lambda_k)}{\Gamma\left(\sum_{i=1}^k \lambda_i\right)}. \end{aligned} \tag{7}$$

Another classical definition of Krichevsky-Trofimov mixtures gives the conditional probabilities: if $x_{1:n}$ is a message on the alphabet $\{1, \ldots, k\}$, then, for all $0 \leq i \leq n-1$ and for all $1 \leq j \leq k$,

$$KT_k(X_{i+1} = j | X_{1:i} = x_{1:i}) = \frac{n_i^j + \frac{1}{2}}{i + \frac{k}{2}}, \tag{8}$$

where n_i^j is the number of occurrences of symbol j in $x_{1:i}$.

Let us now go on in the definition of priors μ_k.

Let f be an integrable envelope function, and Λ_f be the associated envelope class. Choose any fixed $m \in \mathbb{N}_+$ such that

$$\sum_{i \geq m} f(i) < 1.$$

For $k \geq m+1$, let Θ_k denote the subset of Θ_f defined by

$$\Theta_k = \Big\{(\theta_1, 0, \ldots, 0, \theta_m, \ldots, \theta_k, 0, \ldots) : \theta_1 = 1 - \sum_{i=m}^{k} \theta_i \text{ and } \forall m \leq i \leq k,\ 0 \leq \theta_i \leq f(i)\Big\}.$$

Then μ_k denotes the prior on Θ_k proportional to the Dirichlet prior μ on the simplex S_{k-m+2} using the coordinates $(\theta_1, \theta_m, \ldots, \theta_k)$:

$$\begin{aligned} &\mathrm{d}\mu_k(\theta_1, 0, \ldots, 0, \theta_m, \ldots, \theta_k, 0, \ldots) \\ &\quad = \frac{\theta_1^{-1/2} \theta_m^{-1/2} \cdots \theta_k^{-1/2} \, \mathrm{d}\boldsymbol{\theta}}{\int_{\Theta_k} \theta_1^{-1/2} \theta_m^{-1/2} \cdots \theta_k^{-1/2} \, \mathrm{d}\boldsymbol{\theta}} \\ &\quad = \frac{\int_{S_{k-m+2}} \theta_1^{-1/2} \theta_m^{-1/2} \cdots \theta_k^{-1/2} \, \mathrm{d}\boldsymbol{\theta}}{\int_{\Theta_k} \theta_1^{-1/2} \theta_m^{-1/2} \cdots \theta_k^{-1/2} \, \mathrm{d}\boldsymbol{\theta}} \mathrm{d}\mu(\theta_1, \theta_m, \ldots, \theta_k). \end{aligned}$$

In this formula, $\mathrm{d}\boldsymbol{\theta}$ is the Lebesgue measure on the simplex S_{k-m+2} indexed by $(\theta_1, \theta_m, \ldots, \theta_k)$, and Θ_k is identified with its projection on the simplex. Similarly, let $\widetilde{KT}_{k-m+2}$ denote the Bayes mixture associated to the Dirichlet prior μ indexed by $(\theta_1, \theta_m, \ldots, \theta_k)$.

B. Proof of Theorem 4

We shall use the following, which is a modification of Proposition 1 in [6]. We give the proof in Appendix B.

Proposition 4: Let $\boldsymbol{\theta}$ be an element of the simplex S_k, and KT_k the Krichevsky-Trofimov mixture. Then

$$D(P_{\boldsymbol{\theta}}^n; KT_k) \geq \frac{k-1}{2}\log\frac{n}{2\pi} + \log\frac{\Gamma(1/2)^k}{\Gamma(k/2)} - \frac{5k}{3}\log e.$$

To simplify the notations, let us define

$$\begin{aligned} C(k) &= \int_{\Theta_k} \theta_1^{-1/2}\theta_m^{-1/2}\cdots\theta_k^{-1/2}\,\mathrm{d}\boldsymbol{\theta}, \\ D(k) &= \int_{S_{k-m+2}} \theta_1^{-1/2}\theta_m^{-1/2}\cdots\theta_k^{-1/2}\,\mathrm{d}\boldsymbol{\theta} \\ &= \frac{\Gamma(1/2)^{k-m+2}}{\Gamma(\frac{k-m+2}{2})}. \end{aligned}$$

In pratice, prior μ_k is supported by the alphabet $A_k = \{1, m, m+1, \ldots, k\}$. The corresponding Bayes strategy doesn't encode messages with other symbols. As a consequence, it is far from being asymptotically minimax! Let $x_{1:n}$ be an element of A_k^n. Then

$$\begin{aligned} M_{n,\mu_k}(x_{1:n}) &= \frac{D(k)}{C(k)}\int_{\Theta_k} P_{\boldsymbol{\theta}}(x_{1:n})\,\mathrm{d}\mu(\theta_1, \theta_m, \ldots, \theta_k) \\ &\leq \frac{D(k)}{C(k)}\widetilde{KT}_{k-m+2}(x_{1:n}). \end{aligned}$$

Therefore, if $\boldsymbol{\theta}$ is an element of Θ_k, Proposition 4 entails

$$\begin{aligned} D(P_{\boldsymbol{\theta}}^n; M_{n,\mu_k}) &\geq D(P_{\boldsymbol{\theta}}^n; \widetilde{KT}_{k-m+2}) + \log C(k) - \log D(k) \\ &\geq \frac{k-m+1}{2}\log\frac{n}{2\pi} + \log C(k) \\ &\quad - \frac{5(k-m+2)}{3}\log e. \end{aligned}$$

Consequently,

$$\begin{aligned} R_{n,\mu_k}(\Lambda_f) &= \int_{\Theta_k} D(P_{\boldsymbol{\theta}}^n; M_{n,\mu_k})\,\mathrm{d}\mu_k(\boldsymbol{\theta}) \\ &\geq \log C(k) + \frac{k}{2}\log n \\ &\quad - \frac{10\log e + 3\log 2\pi}{6}k - \frac{m-1}{2}\log n \\ &\quad + \left(\frac{m-1}{2}\log 2\pi + \frac{5(m-2)}{3}\log e\right). \end{aligned} \tag{9}$$

Now, let us calculate $C(k)$. First note that the choice of m made before is such that all values of $(\theta_m, \ldots, \theta_k)$ in the rectangle $[0, f(m)] \times \cdots \times [0, f(k)]$ are possible. It allows us to write the integrals over Θ_k as integrals over that rectangle.

$$\begin{aligned} C(k) &= \int_{\Theta_k} \frac{\mathrm{d}\theta_m \cdots \mathrm{d}\theta_k}{\sqrt{\theta_1\theta_m\cdots\theta_k}} \\ &\geq \prod_{i=m}^{k}\int_0^{f(i)} \frac{\mathrm{d}\theta_i}{\sqrt{\theta_i}} \\ &= \prod_{i=m}^{k} 2\sqrt{f(i)}. \end{aligned}$$

At this point we need to specify f. In the case of the exponentially decreasing envelope class $\Lambda_{Ce^{-\alpha\cdot}}$, $f(i) = \min(1, Ce^{-\alpha i})$. Since $\sum_{i\geq m} f(i) < 1$, $f(i) = Ce^{-\alpha i} < 1$ for all $i \geq m$. Thus

$$C(k) \geq \prod_{i=m}^{k} 2\sqrt{C}e^{-\frac{\alpha}{2}i},$$

and

$$\log C(k) \geq (k-m+1)\left(1 + \frac{\log C}{2}\right) - \frac{\alpha\log e}{2}\sum_{i=m}^{k} i.$$

If we plug it in (9), we get $R_{n,\mu_k}(C,\alpha) \geq g(n,k)$, where

$$\begin{aligned} g(n,k) = &-\frac{\alpha\log e}{4}k^2 + \frac{k}{2}\log n + \left[1 + \frac{\log C}{2} - \frac{\alpha\log e}{4}\right. \\ &\left. - \frac{5\log e}{3} - \frac{\log 2\pi}{2}\right]k - \frac{m-1}{2}\log n \\ &+ \left[\frac{\alpha(m^2+m-1)\log e}{4} - (m-1)\left(1 + \frac{\log C}{2}\right)\right. \\ &\left. + \frac{m-1}{2}\log 2\pi + \frac{5(m-2)\log e}{3}\right]. \end{aligned}$$

With the choice $k_n = \left\lfloor \frac{1}{\alpha\log e}\log n \right\rfloor$, only the first two terms matter, and

$$g(n, k_n) \sim \frac{1}{4\alpha\log e}\log^2 n.$$

That achieves the proof of Theorem 4.

IV. AutoCensuring Code

This section presents a new algorithm called AutoCensuring Code (ACcode). With respect to the exponentially decreasing envelope class $\Lambda_{Ce^{-\alpha\cdot}}$, its maximum redundancy is equivalent to the minimax redundancy of this class of sources. Furthermore ACcode is adaptive, as the same algorithm verifies that property with all exponentially decreasing envelope classes. This is formulated in the following theorem.

Let $\texttt{ACcode}(x_{1:n})$ denote the binary string produced by ACcode when it encodes the message $x_{1:n}$, and let $l(\cdot)$ denote the length of a string.

Theorem 5: For any positive numbers C and α satisfying $C > e^{2\alpha}$,

$$\sup_{P\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_{P^n}[l(\texttt{ACcode}(X_{1:n})) - H(P^n)] \underset{n\to\infty}{\sim} R_n(C,\alpha).$$

Theorem 5 does not say that the difference between the redundancy of ACcode and the minimax redundancy tends to zero, or even that this difference is bounded. Therefore, there may exist codes whose redundancy is smaller than the redundancy of ACcode, but with a benefit negligible with respect to $\log^2 n$.

The ACcode algorithm does not code entirely on line, but this limitation can be taken off. In subsection IV-B a modified version of the algorithm is proposed, which codes entirely on line.

ACcode is in fact a modification of the Censuring Code proposed by Boucheron, Garivier and Gassiat in [36]. Here is their description:

> [The `CensoringCode` algorithm] is parametrised by a sequence of cutoffs $(K_i)_{i\in\mathbb{N}}$ and handles the i^{th} symbol of the sequence to be compressed differently according to whether it is smaller or larger than cutoff K_i, in the latter situation, the symbol is said to be censored. The `CensoringCode` algorithm uses Elias penultimate code to encode censored symbols and Krichevsky-Trofimov mixtures to encode the sequence of non-censored symbols padded with markers (zeros) to witness acts of censorship.

The main issue that arises with CensuringCode is the choice of cutoffs $(K_i)_{i\in\mathbb{N}}$. Cutoffs must be adapted to the parameters of the envelope. In our case, exponentially decreasing envelopes make great symbols to be very few. So a simple and rather natural way to choose cutoffs is to take the greatest element M_n actually appeared in the sequence $(X_1,\dots,X_n)$:

$$\text{For } 1\le i\le n,\ M_i=\sup_{1\le j\le i} X_j.$$

This idea turns out efficient to get a small redundancy, as we will see. But this choice has another advantage: we do not need to known the actual parameters of the exponentially decreasing envelope.

In order to code on line, we want to be able to deal with each symbol of the message without previous knowledge of subsequent symbols nor of the total length of the message. None of these points is possible if we choose M_n as a constant cutoff for all symbols. So we will code X_{i+1} using as cutoff M_i, the maximum of previous symbols.
Additionally, doing that improves the redundancy of the algorithm, as we will see later.

Let us describe the algorithm in practice.

A. Definition of the `ACcode` algorithm

Let $n\ge 1$ be some positive integer, and let $x_{1:n}=x_1x_2\dots x_n$ be a string from $\mathbb{N}_+^n$ to be encoded.

In this setting, we can define the sequence of maxima

$$m_0=0 \text{ and } m_i=\sup_{1\le k\le i} x_k,\ \text{for all } 1\le i\le n.$$

Let string $\widetilde{\boldsymbol{m}}$ be the sequence of positive (m_i-m_{i-1}), that is to say $(m_i-m_{i-1})_{m_i>m_{i-1},1\le i\le n}$.
Eventually, let $\widetilde{m}_k$ be the k^{th} element of $\widetilde{\boldsymbol{m}}$. In order to link m_i and $\widetilde{m}_k$, let us define $n_i^0=\sum_{j=1}^{i}\mathbb{1}_{m_i>m_{i-1}}$ the number of "new" maxima among the first i symbols. Then

$$m_i=\sum_{k=1}^{n_i^0}\widetilde{m}_k. \tag{10}$$

String $\widetilde{\boldsymbol{m}}$ is encoded into a binary string `C2` by applying Elias penultimate code (see [33]) to each symbol in $\widetilde{\boldsymbol{m}}$. Elias penultimate code is a prefix code which uses $l_E(x)$ bits to encode a positive integer x, with

$$\begin{aligned} &l_E(1)=1,\\ &l_E(x)=1+\lfloor\log x\rfloor+2\lfloor\log\lfloor\log x\rfloor+1\rfloor \quad \text{if } x\ge 2. \end{aligned} \tag{11}$$

Meanwhile the arithmetic coding is applied to $x_{1:n}$ using side information from $\widetilde{\boldsymbol{m}}$. This produces the binary string `C4`.

In the string `C4` we do not really encode $x_{1:n}$, but a modification $\widetilde{x}_{1:n}=\widetilde{x}_1\widetilde{x}_2\dots\widetilde{x}_n$ of $x_{1:n}$ defined by

$$\widetilde{x}_i=x_i\mathbb{1}_{x_i\le m_{i-1}}=\begin{cases} x_i & \text{if } x_i\le m_{i-1},\\ 0 & \text{otherwise.}\end{cases}$$

All symbols greater than m_{i-1} are encoded together: they are replaced by the extra symbol 0, and this extra symbol is encoded instead.

Then we apply the arithmetic coding to $\widetilde{x}_{1:n}$, using Krichevsky-Trofimov mixtures. To further describe it we need to set some counters.
For $j\ge 1$ and $i\ge 0$, let n_i^j be the number of occurrences of symbol j in $x_{1:i}$ (with convention $n_0^j=0$ for all $j\ge 1$). The conditional coding probability is then defined in the following way:

$$Q_{i+1}(\widetilde{X}_{i+1}=j|X_{1:i}=x_{1:i})=\frac{n_i^j+\frac12}{i+\frac{m_i+1}{2}} \quad \text{if} \quad 1\le j\le m_i,$$
$$Q_{i+1}(\widetilde{X}_{i+1}=0|X_{1:i}=x_{1:i})=\frac{1/2}{i+\frac{m_i+1}{2}}.$$

The event $\{\widetilde{X}_{i+1}=0\}$ is equal to $\{X_{i+1}>M_i\}$. If $x_{i+1}=j>m_i$, then $n_i^j=0$, and we still have

$$Q_{i+1}(\widetilde{X}_{i+1}=0|X_{1:i}=x_{1:i})=\frac{n_i^j+\frac12}{i+\frac{m_i+1}{2}}.$$

In the sequel we will use the notation

$$Q^n(\widetilde{x}_{1:n})=\prod_{i=0}^{n-1}Q_{i+1}(\widetilde{x}_{i+1}|x_{1:i})$$

to denote the coding probability used to encode the entire string $\widetilde{x}_{1:n}$.

One could imagine to take m_{i+1} as cutoff, instead of m_i, to code x_{i+1}. It should no more be really a cutoff, as x_{i+1} cannot be greater than m_{i+1}, and we shouldn't need any more the extra symbol 0. However this extra symbol has a special use in our setting: it makes the decoder to know when m_i changes. Indeed the string `C2` permits us to know *what* are the different m_i's, but not *when* each of them must be used.

Another remark we can do is that the extra symbol 0 is always considered as new: when $x_{i+1}>m_i$, we encode the symbol 0 but we increment the counter $n_i^{x_{i+1}}$.

Eventually, when the coder reaches the end of the string $x_{1:n}$, he reads a termination signal on the input, and then he knows n[4]. n is encoded by Elias penultimate code into the string `C3`. Once n is known, the coder knows also n_n^0, the number of elements in $\widetilde{\boldsymbol{m}}$. n_n^0 is encoded with Elias penultimate code into the string `C1`.

The last step is to concatenate the strings `C1` to `C4`, to constitute the final encoded message $\texttt{C1}\cdot\texttt{C2}\cdot\texttt{C3}\cdot\texttt{C4}$.

Fig. 1 gives a technical description of the `ACcode` algorithm in pseudo-code.

[4]In the description in pseudo-code of the `ACcode` algorithm given below, we suppose that the coder knows from the beginning the entire message $x_{1:n}$. In this case, n is simply the length of $x_{1:n}$.

$n \leftarrow$ `length`(x)
$M \leftarrow 0$
$n_0 \leftarrow 0$
`C2` $\leftarrow$ `""`
$counts \leftarrow [1/2, 1/2, \ldots]$
for $i = 1$ to n **do**
  if $x[i] \leq M$ **then**
    `ArithCode`$(x[i], counts[0 : M])$
  else
    `ArithCode`$(0, counts[0 : M])$
    `C2` $\leftarrow$ `C2` $\cdot$ `EliasCode`$(x[i] - M)$
    $M \leftarrow x[i]$
    $n_0 \leftarrow n_0 + 1$
  end if
  $counts[x[i]] \leftarrow counts[x[i]] + 1$
end for
`C1` $\leftarrow$ `EliasCode`(n_0)
`C3` $\leftarrow$ `EliasCode`(n)
`C4` $\leftarrow$ `ArithCode`$()$
return `C1` $\cdot$ `C2` $\cdot$ `C3` $\cdot$ `C4`

Fig. 1. `ACcode` algorithm

Decoding procedure: To correctly decode, the decoder needs to know where are the boundaries of the strings `C1` to `C4`. Since Elias penultimate code is a prefix code, it is easy to know where `C1` and `C3` end. The only use of `C1` is to transmit n_n^0, which makes the decoder to know where `C2` ends.

While decoding the i^{th} element in `C4`, one needs to know the counters $(n_{i-1}^j)_{j\geq 1}$ and the present maximum m_{i-1}. How to obtain these numbers? The counters n_{i-1}^j can be computed from the first $i-1$ decoded symbols. Another useful counter is n_{i-1}^0, which permits us to retrieve m_{i-1} from string $\widetilde{\boldsymbol{m}}$, thanks to (10). n_{i-1}^0 too can be computed from yet decoded symbols: it is the number of times the extra symbol 0 appears in the $i-1$ first symbols. Suppose that $n_{i-1}^0 = k$. If the i^{th} decoded symbol is the extra symbol 0, then $n_i^0 = k+1$ and $x_i = m_i = m_{i-1} + \widetilde{m}_{k+1}$.

Eventually, the decoder knows where `C4` ends: as soon as he has received enough bits to decode the n^{th} symbol. This property makes `ACcode` to be a prefix code (on the alphabet $\mathbb{N}_+^* = \bigcup_{n\geq 1} \mathbb{N}_+^n$ of all finite messages from $\mathbb{N}_+$).

B. Improvements

We describe here a modification `ACbis` of the `ACcode` algorithm. The `ACbis` algorithm is entirely on line.

What prevents `ACcode` to be entirely on line is that the strings `C1` and `C3` cannot be encoded before the coder finishes reading the original message. `C1` is transmitted at the very beginning, and `C3` is transmitted before `C4`, which contains the main part of the information. Furthermore, the transmission of the main part `C4` does not start until `C2` is entirely transmitted.

It is quite easy to take off the string `C3` in the encoded message. `C3` encodes n, which makes the decoder to know when the message ends. Instead of n, we can transmit a termination signal in the following way.
At the end of $\widetilde{X}_{1:n} = \widetilde{X}_1 \ldots \widetilde{X}_n$ we add an additional extra symbol 0, and the new string $\widetilde{X}_1 \ldots \widetilde{X}_n 0$ is encoded by the arithmetic code into the string `C4`. The coding probability is the same as the one of `ACcode`, in this case Q^{n+1}. For simplicity, we use the same notation `C4` to denote the new version of the string `C4`.
When the decoder reads a 0 in `C4`, he looks for the next element of $\widetilde{\boldsymbol{m}} = \widetilde{m}_1 \ldots \widetilde{m}_{n_n^0}$ to know the new maximum m_i. So, we have to encode the true termination signal into `C2` together with the code of $\widetilde{\boldsymbol{m}}$. In practice, we add the number 0 at the end of $\widetilde{\boldsymbol{m}}$, and the new string $\widetilde{\boldsymbol{m}}0 = \widetilde{m}_1 \ldots \widetilde{m}_{n_n^0} 0$ is encoded into `C2`. Since Elias penultimate code does not encode 0, we shift forward all numbers: the integer $x \geq 0$ is encoded by the Elias code of $x+1$.

What about the cost in bits of these modifications? Since `C3` is not transmitted any more, we recover $l_E(n)$ bits. In `C2` we add $l_E(1) = 1$ bit for the final 0, and the cost of transmitting $\widetilde{m}_k$ becomes $l_E(\widetilde{m}_k + 1)$ (instead of $l_E(\widetilde{m}_k)$). However the proof given in the subsection V-B remains the same: the length of `C2` is still negligible with respect to $\log^2 n$. Eventually, the 0 at the end of `C4` increases the redundancy by

$$\begin{aligned}
&\mathbb{E}_{P^n}[-\log Q_{n+1}(0|X_{1:n})] \\
&\quad = \mathbb{E}_{P^n}[\log(M_n + 1 + 2n)] \\
&\quad \leq \log(2n) + \frac{\mathbb{E}_{P^n}[M_n + 1]}{2n} \\
&\quad = O(\log n).
\end{aligned}$$

The last line is obtained through Lemma 7 (see Appendix C-A).

Once we have removed `C3`, we still have to deal with two issues: `C1` is still present, and the transmission of the main part `C4` does not start until `C2` is entirely transmitted. The only use of `C1` is to make the decoder to know where `C2` ends and where `C4` begins. If we find a way to transmit `C2` and `C4` together, we can do away with `C1`. The `ACbis` algorithm offers such a way to overlap `C2` and `C4`.

To decode the $(i+1)^{\text{th}}$ symbol in `C4`, the decoder needs the knewledge of m_i, which is obtained from the beginning of the string `C2`. When the decoder meets a 0 at the $(i+1)^{\text{th}}$ position, then he needs to read the next $\widetilde{m}_k$ in $\widetilde{\boldsymbol{m}}$ to know m_{i+1}. The position of the new $\widetilde{m}_k$ to be read is $k = n_{i+1}^0$; this relation corresponds to the fact that $\widetilde{x}_{i+1}$ is the k^{th} 0 in $\widetilde{x}_{1:n}$.
The arithmetic code uses $\lceil -\log Q^{i+1}(\widetilde{x}_{1:i+1}) \rceil + 1$ bits to send the first $(i+1)$ symbol of $x_{1:n}$. Suppose that $\widetilde{x}_{i+1} = 0$ and $n_{i+1}^0 = k$. As soon as $\lceil -\log Q^{i+1}(\widetilde{x}_{1:i+1}) \rceil + 1$ bits of `C4` have been sent, the k^{th} zero of $\widetilde{x}_{1:n}$ can be decoded, and the `ACbis` algorithm sends the Elias code of $\widetilde{m}_k + 1$. Then `ACbis` transmit `C4` again, from the $(\lceil -\log Q^{i+1}(\widetilde{x}_{1:i+1}) \rceil + 2)^{\text{th}}$ bit.

Fig. 2 shows an illustration of the transmission process of `ACbis`. In this example, the initial message is $x_{1:4} = 5, 3, 2, 7$. Then the message encoded in `C4` is $\widetilde{x}_{1:4}0 = 0, 3, 2, 0, 0$. 10 bits are needed to transmit the second 0, and 14 bits for the last one. In `C2` we transmit $\widetilde{\boldsymbol{m}}0 = 5, 2, 0$.

The first alone bit of `C4` can seem curious: indeed no bit is needed to know that the first symbol of $\widetilde{x}_{1:n}$ is a zero, and we could begin by the transmission of $\widetilde{m}_1$. That point illustrates the fact that sometimes $\lceil -\log Q^{i+1}(\widetilde{x}_{1:i+1}) \rceil$ are enough to

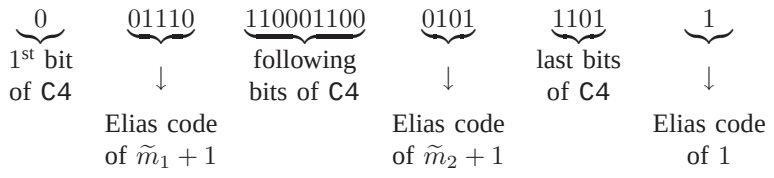

Fig. 2. Example of ACbis

decode the $(i+1)^{\text{th}}$ symbol. To simplify the procedure, we allways transmit $\widetilde{m}_k$ after $\lceil -\log Q^{i+1}(\widetilde{x}_{1:i+1})\rceil + 1$ bits from C4.

V. Proof of Theorem 5

The proof of Theorem 5 depends on two propositions which are respectively proved in subsections V-A and V-B.

Proposition 5: Let C and α be positive numbers satisfying $C > e^{2\alpha}$. Then

$$\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_{P^n}[-\log Q^n(\widetilde{X}_{1:n}) - H(P^n)] \leq (1+o(1))\frac{1}{4\alpha\log e}\log^2 n.$$

Proposition 6: Let C and α be positive numbers satisfying $C > e^{2\alpha}$. Then

$$\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_{P^n}[l(\texttt{C2})] \leq o(\log^2 n).$$

Thanks to these propositions, we can go forward in the proof of Theorem 5.

Proof of Theorem 5: The message sent by the ACcode algorithm is compound of four strings C1 to C4. C3 is the Elias code of n, and is compound of $l(\texttt{C3}) = l_E(n)$ bits, with l_E as in (11).

In the same way, C1 and is compound of $l_E(n_n^0)$ bits. l_E is a nondecreasing function, and $n_n^0 \leq n$ entails $l_E(n_n^0) \leq l_E(n)$. C4 corresponds to the part of the message encoded by the arithmetic code, with coding probability Q^n. The arithmetic code encodes a message $x_{1:n}$ with $\lceil -\log Q^n(x_{1:n})\rceil + 1$ bits. Therefore the redundancy of ACcode can be upper bounded, for all $n \geq 2$, by

$$\begin{aligned}\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} &\mathbb{E}_{P^n}[l(\texttt{C1}\cdot\texttt{C2}\cdot\texttt{C3}\cdot\texttt{C4})] - H(P^n)] \\ &\leq 2 + 2\log n + 4\log\log n \\ &\quad + \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_{P^n}[-\log Q^n(\widetilde{X}_{1:n}) - H(P^n)] \\ &\quad + 2 + \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_{P^n}[l(\texttt{C2})].\end{aligned}$$

Using Propositions 5 and 6, we get

$$\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_{P^n}[l(\texttt{C1}\cdot\texttt{C2}\cdot\texttt{C3}\cdot\texttt{C4})] - H(P^n)] \leq (1+o(1))\frac{1}{4\alpha\log e}\log^2 n.$$

We recognize the quantity given in Theorem 2, equivalent to the minimax redundancy. Since the minimax redundancy is the infimum of the redundancy of all algorithms, this concludes the proof. ∎

Let us now prove Proposition 5 and Proposition 6.

A. Contribution of C4

In this subsection we prove Proposition 5. We give here the sketch of the proof, and we delay to Appendix C-B the proofs of (13), (14), (15), and (16).

Here we deal with the quantity

$$(\mathrm{A}) = \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_{P^n}[-\log Q^n(\widetilde{X}_{1:n}) - H(P^n)].$$

that corresponds to the contribution of C4. As we saw in Section IV, the coding probability Q^n is based on Krichevsky-Trofimov mixtures. But Krichevsky-Trofimov mixtures have been studied by many authors. Can we reuse their work?

For $k \geq 1$, let KT_k denote the Krichevsky-Trofimov mixture as in (8). Let us choose $k = m_n$. In this case, there is a simple relation between KT_{m_n} and Q^n. For any sequence of n positive integers $x_{1:n} \in \mathbb{N}_+^n$,

$$\begin{aligned}&Q_{i+1}(\widetilde{X}_{i+1} = x_{i+1}|X_{1:i} = x_{1:i}) \\ &\quad = \frac{2i+m_n}{2i+1+m_i} KT_{m_n}(X_{i+1} = x_{i+1}|X_{1:i} = x_{1:i}).\end{aligned}$$

As a consequence, we can link the redundancy of Q^n to the redundancy of KT_{m_n}:

$$\begin{aligned}-\log Q^n(\widetilde{X}_{1:n}) = &-\sum_{i=0}^{n-1}\log KT_{M_n}(X_{i+1}|X_{1:i}) \\ &-\sum_{i=0}^{n-1}\log\frac{2i+M_n}{2i+1+M_i}\end{aligned}$$

and therefore

$$(\mathrm{A}) = \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \Bigg(\overbrace{\mathbb{E}_{P^n}\big[-\log KT_{M_n}(X_{1:n}) - H(P^n)\big]}^{(\mathrm{A}_1)} - \overbrace{\mathbb{E}_{P^n}\Big[\sum_{i=0}^{n-1}\log\frac{2i+M_n}{2i+1+M_i}\Big]}^{(\mathrm{A}_2)}\Bigg). \tag{12}$$

Note that (A_2) corresponds to the gain in redundancy of Q^n with respect to KT_{M_n}. It illustrates the benefit of taking M_i instead of M_n as cutoff to encode X_{i+1}.

On the one hand, we have

$$(\mathrm{A}_1) \leq \frac{\mathbb{E}[M_n]-1}{2}\log n + \mathbb{E}[\log M_n]. \tag{13}$$

Since $\mathbb{E}[\log M_n] \leq \mathbb{E}[M_n]$, Lemma 7 (see Appendix C-A) entails

$$\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}[\log M_n] = o(\log^2 n).$$

Applying Lemma 7 again, we see that (A_1) produces a redundancy equivalent to $\frac{1}{2\alpha \log e} \log^2 n$, which is twice bigger than the minimax redundancy obtained in Theorem 2. So, we will hope the corrective term (A_2) to be about $\frac{1}{4\alpha \log e} \log^2 n$. To deal with (A_2), we use the concavity of the log function, and we group the terms in the sum, M_n by M_n. Let $m = \left\lfloor \frac{n-1}{M_n} \right\rfloor$ be the number of bundles.
To simplify the expression, we also neglect few terms at the beginning of the sum. Let $(h_n)_{n\geq 1}$ be a nondecreasing sequence of positive integers, such that $h_n \to \infty$ as $n \to \infty$, and let us define $\lambda_n = 2h_n \log\left(1 + \frac{1}{2h_n}\right)$ .Then

$$(\mathrm{A}_2) \geq \lambda_n \mathbb{E}_{P^n}\left[\sum_{k=h_n+1}^{m} \frac{M_n - M_{kM_n}}{2k+1}\right] - \frac{1+\ln n}{2}. \quad (14)$$

It is easy to verify that the function $x \mapsto x \log\left(1 + \frac{1}{x}\right)$ is nondecreasing, and tends to $\log e$ when x tends to the infinity; therefore (λ_n) tends to $\log e$.

At this point, we can do heuristic considerations. Let us substitute $\frac{1}{\alpha} \ln k$ for M_k (this comes from Lemma 7). Neglecting the term $\frac{1+\ln n}{2}$, (A_2) becomes

$$\begin{aligned}(\mathrm{A}_2') &= \lambda_n \sum_{k=h_n+1}^{m'} \frac{\ln n - \ln\left(\frac{k}{2\alpha}\ln n\right)}{\alpha\,(2k+1)} \\ &\geq \frac{\lambda_n}{2\alpha}\left[\left(\ln \frac{2\alpha n}{\ln n}\right)\int_{h_n}^{m'} \frac{\mathrm{d}x}{x+\frac{1}{2}} - \int_{h_n}^{m'} \frac{\ln x}{x-\frac{1}{2}}\,\mathrm{d}x\right]\end{aligned}$$

with $m' = \left\lfloor \frac{\alpha\,(n-1)}{\ln n} \right\rfloor$. A further calculation gives what we hoped:

$$(\mathrm{A}_2') \geq \frac{1+o(1)}{2\alpha \log e} \log^2 n.$$

When we do the exact calculation, the presence of the supremum leads us to act a little differently:

$$\begin{aligned}(\mathrm{A}) \leq \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} &\left(\frac{\mathbb{E}[M_n]}{2}\log n \right.\\ &\left. - \lambda_n \mathbb{E}\left[\sum_{k=h_n+1}^{m} \frac{M_n - M_{kM_n}}{2k+1}\right]\right) + o(\log^2 n)\end{aligned}$$

$$\leq \frac{1}{2}\overbrace{\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_{P^n}\left[M_n \log n - \lambda_n M_n \sum_{k=h_n+1}^{m} \frac{1}{k+\frac{1}{2}}\right]}^{(\mathrm{A}_3)}$$

$$+ \frac{\lambda_n}{2}\overbrace{\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_{P^n}\left[\sum_{k=h_n+1}^{m} \frac{M_{kM_n}}{k+\frac{1}{2}}\right]}^{(\mathrm{A}_4)} + o(\log^2 n)$$

Let us choose $h_n = \max\{1, \lfloor \ln n - 3/2 \rfloor\}$. Then

$$(\mathrm{A}_3) = o(\log^2 n) \quad (15)$$

$$(\mathrm{A}_4) \leq \frac{\log^2 n}{2\alpha \log^2 e} + o(\log^2 n). \quad (16)$$

Therefore we have

$$(\mathrm{A}) \leq (1+o(1))\frac{1}{4\alpha \log e} \log^2 n$$

which concludes the proof of Proposition 5.

B. Contribution of C2

Like in the previous subsection, we give first the sketch of the proof, and we delay to Appendix C-C several technical lemmas.

Here we want to calculate the contribution of the quantity

$$(\mathrm{B}) = \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_{P^n}[l(\mathtt{C2})].$$

For a message $X_{1:n}$, the string C2 is the concatenation of the Elias codes of the sequence of positive $(M_i - M_{i-1})$. Elias code uses $l_E(x) = 1 + \lfloor \log x \rfloor + 2\lfloor \log \lfloor \log x + 1 \rfloor \rfloor$ bits to encode a integer $x \geq 2$, and $l_E(1) = 1$ bit to encode 1. Consequently,

$$(\mathrm{B}) \leq \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \sum_{i=1}^{n} \mathbb{E}_{P^n}\left[\mathbb{1}_{X_i > M_{i-1}} l_E(X_i)\right].$$

At this point we use the following lemma, which is proved later.

Lemma 4: Let g be a positive and nondecreasing function on $[1,\infty)$. Let $(K_n)_{n\geq 1}$ be a nondecreasing sequence of positive integers. Then, for all $n \geq 1$,

$$\begin{aligned}&\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \sum_{i=1}^{n} \mathbb{E}_{P^n}\left[\mathbb{1}_{X_i > M_{i-1}} g(X_i)\right] \\ &\leq (1+o(1))g(K_n)\frac{1}{\alpha}\ln n + C\,n\int_{K_n}^{\infty} g(x+1)e^{-\alpha x}\,\mathrm{d}x.\end{aligned}$$

To apply Lemma 4, we extend the definition of l_E on $[1,\infty)$ by

$$l_E(x) = \begin{cases} 1 & \text{if } x \in [1,2), \\ 1 + \lfloor \log x \rfloor + 2\lfloor \log \lfloor \log x \rfloor + 1 \rfloor & \text{if } x \geq 2. \end{cases}$$

We get

$$(\mathrm{B}) \leq (1+o(1))\frac{1}{\alpha} l_E(K_n)\ln n + C\,n\int_{K_n}^{\infty} l_E(x+1)e^{-\alpha x}\,\mathrm{d}x$$

Then we can choose $K_n = \max\{1, \lfloor \frac{1}{\alpha}\ln n \rfloor\}$. This entails

$$l_E(K_n) \underset{n\to\infty}{\sim} \log K_n \sim \log\log n = o(\log n),$$

and therefore

$$\frac{1}{\alpha} l_E(K_n)\ln n = o(\log^2 n).$$

The remaining term is treated by Lemma 5, which achieves the proof of Proposition 6:

Lemma 5: Let $\alpha > 0$ be a real number, and let $K_n = \max\{1, \lfloor \frac{1}{\alpha}\ln n \rfloor\}$. Then

$$n\int_{K_n}^{\infty} l_E(x+1)e^{-\alpha x}\,\mathrm{d}x = o(\log n).$$

APPENDIX A
METRIC ENTROPY OF EXPONENTIALLY DECREASING ENVELOPE CLASSES

We prove here several lemmas from subsection II-B.

Proof of Lemma 2: N_ϵ denotes the threshold from which we want to truncate the coordinates. If $y = (y_n)_{n\geq 1}$ is an element of A_f, its truncated version is $\widetilde{y} = (y_n \mathbb{1}_{n\leq N_\epsilon})_{n\geq 1}$. Then

$$\begin{aligned}\|y-\widetilde{y}\| &= \sqrt{\sum_{n\geq N_\epsilon+1} y_n^2} \\ &\leq \sqrt{\sum_{n\geq N_\epsilon+1} \sqrt{f(n)}^2} \\ &= \sqrt{\sum_{k\geq N_\epsilon+1} f(k)} \\ &\leq \frac{\epsilon}{4}.\end{aligned}$$

Suppose now that S is an $\epsilon/4$-cover of $\{y \in A_f : \ \forall n \geq N_\epsilon, y_n = 0\}$. Let z denote an element of A_f. Then it exists some $y \in S$ such that $\|\widetilde{z} - y\| \leq \epsilon/4$. Thus $\|z-y\| \leq \epsilon/2$, and S is an $\epsilon/2$-cover of A_f.

Using the fact that

$$\begin{aligned}&\{y \in A_f : \ \forall k \geq N_\epsilon, y_k = 0\} \\ &\qquad = \prod_{1\leq k\leq N_\epsilon} \left[0, \sqrt{f(k)}\right] \times \{0\}^{\{k\geq N_\epsilon+1\}},\end{aligned}$$

we can then write

$$\begin{aligned}\mathcal{D}_\epsilon(\Theta_f, d) &= \mathcal{D}_\epsilon(A_f \cap \{\|x\| = 1\}, \|\cdot\|_{\ell^2}) \\ &\leq \mathcal{N}_{\epsilon/2}(A_f, \|\cdot\|_{\ell^2}) \\ &\leq \mathcal{N}_{\epsilon/4}\left(\prod_{1\leq k\leq N_\epsilon} \left[0, \sqrt{f(k)}\right], \|\cdot\|_{\mathbb{R}^{N_\epsilon}}\right) \\ &\leq \mathcal{M}_{\epsilon/4}\left(\prod_{1\leq k\leq N_\epsilon} \left[0, \sqrt{f(k)}\right], \|\cdot\|_{\mathbb{R}^{N_\epsilon}}\right) \\ &\leq \frac{\mathrm{Vol}\left(\prod_{1\leq k\leq N_\epsilon}\left[-\frac{\epsilon}{8}, \sqrt{f(k)}+\frac{\epsilon}{8}\right]\right)}{\mathrm{Vol}\left(B_{\mathbb{R}^{N_\epsilon}}(0, \frac{\epsilon}{8})\right)} \\ &\leq \left(\frac{\epsilon}{8}\right)^{-N_\epsilon} \frac{1}{\mathrm{Vol}\left(B_{\mathbb{R}^{N_\epsilon}}(0,1)\right)} \prod_{k=1}^{N_\epsilon}\left(\sqrt{f(k)}+\frac{\epsilon}{4}\right).\end{aligned}$$

The second line comes from Lemma 1. The third from the argument given before. The next line comes from Lemma 1 again. The passage from $\mathcal{M}_{\epsilon/4}$ to the volumes is justified by the following reasoning: the cardinality of an $\epsilon/4$-separated subset of a set S is smaller than the number of disjoint balls of radius $\epsilon/8$ you can put in S enlarged by $\epsilon/8$; the later number is then upper bounded by the ratio in volume of these sets. The last line does not raise any issue.

A first consequence of that calculus is that $\mathcal{D}_\epsilon(\Theta_f, d)$ is finite for all $\epsilon > 0$. The first assertion of Lemma 2 is then obtained by applying the logarithm function.

We have also to justify the assertions about $A(N)$. First, the volume of the unit ball of $\mathbb{R}^N$ is a classical result that can be found for example in [40, p. 11]. Then we use the Feller bounds in the version proposed by [41, ch. XII]:

$$\Gamma(x) = \sqrt{2\pi}\, x^{x-1/2} e^{-x} e^{\frac{\beta}{12x}}, \quad \text{with } \beta \in [0,1]. \tag{17}$$

Therefore we can write

$$\begin{aligned}A(N) &= -\frac{N}{2}\ln\pi + \frac{N+1}{2}\ln\left(\frac{N}{2}+1\right) - \left(\frac{N}{2}+1\right) \\ &\quad + \frac{\ln(2\pi)}{2} + \frac{\beta}{12\left(\frac{N}{2}+1\right)} \\ &\sim \frac{N}{2}\ln N.\end{aligned}$$

As $N_\epsilon \to \infty$, the announced equivalence is true. ∎

Proof of Corollary 1: In this case we have

$$\begin{aligned}N_\epsilon &\leq \inf\left\{n \geq 1 : \sum_{k\geq n+1} Ce^{-\alpha k} \leq \frac{\epsilon^2}{16}\right\} \\ &\leq \left\lceil \frac{2}{\alpha}\ln(1/\epsilon) + \frac{1}{\alpha}\ln\frac{16C}{1-e^{-\alpha}} - 1\right\rceil,\end{aligned} \tag{18}$$

and then

$$N_\epsilon \leq \frac{2}{\alpha}\ln(1/\epsilon) + \frac{1}{\alpha}\ln\frac{16C}{1-e^{-\alpha}} \underset{\epsilon\to 0}{\sim} \frac{2}{\alpha}\ln(1/\epsilon).$$

Therefore we can then upper bound $B(\epsilon)$:

$$\begin{aligned}B(\epsilon) &\leq \sum_{k=1}^{N_\epsilon} \ln\left(\frac{\epsilon}{4} + \sqrt{C}e^{-\frac{\alpha}{2}k}\right) \\ &\leq \int_0^{N_\epsilon}\left[\frac{\ln C}{2} - \frac{\alpha}{2}x + \ln\left(1 + \frac{\epsilon\, e^{\frac{\alpha}{2}x}}{4\sqrt{C}}\right)\right]\mathrm{d}x.\end{aligned}$$

On the other hand (18) gives

$$\frac{\epsilon\, e^{\frac{\alpha}{2}N_\epsilon}}{4\sqrt{C}} \leq \frac{1}{\sqrt{1-e^{-\alpha}}},$$

and therefore

$$\begin{aligned}B(\epsilon) &\leq -\frac{\alpha}{4}N_\epsilon^2 + \left(\frac{\ln C}{2} + \frac{1}{\sqrt{1-e^{-\alpha}}}\right)N_\epsilon \\ &\sim -\frac{\alpha}{4}N_\epsilon^2 \sim -\frac{1}{\alpha}\ln^2(1/\epsilon).\end{aligned}$$

We have also

$$A(N_\epsilon) \sim \frac{2}{\alpha}\ln(1/\epsilon) \times \ln\ln(1/\epsilon) = o(\ln^2(1/\epsilon)),$$

and, gathering all these results, we get

$$\mathcal{H}_\epsilon \leq g(\epsilon)$$

with $g(\epsilon) \underset{\epsilon\to 0}{\sim} \frac{1}{\alpha}\ln^2(1/\epsilon)$. ∎

Before proving Lemma 3, we need a simple tool which permits us to truncate coordinates:

Lemma 6: Let $(E, \|\cdot\|)$ be a Hilbert space. Let F be a closed subspace of E, and $\pi : E \mapsto F$ be the orthogonal projection over F. Let S be any closed and totally bounded subset of E. Then, for all $\epsilon > 0$,

$$\mathcal{D}_\epsilon(S, \|\cdot\|) \geq \mathcal{D}_\epsilon(\pi(S), \|\cdot\|).$$

Proof of Lemma 6: Let $\Pi = \{B_i\}_{1\leq i\leq \mathcal{D}_\epsilon(S,\|\cdot\|)}$ be a finite partition of S of diameter smaller than, or equal to ϵ.

From Π we can construct a partition of $\pi(S)$ of diameter at most ϵ. Let us define

$$B'_1 = \pi(B_1),$$
$$B'_i = \pi(B_i) - \bigcup_{1\le k\le i-1} B'_k, \quad \text{for all } 2 \le i \le \mathcal{D}_\epsilon(S, \|\cdot\|).$$

The collection $\{B'_i\}_{\{i:B'_i\neq\varnothing\}}$ is a partition of $\pi(S)$ of diameter at most ϵ, and its cardinality is not more than $\mathcal{D}_\epsilon(S, \|\cdot\|)$. ∎

Proof of Lemma 3: Let $m \ge 1$ be an integer. We want to project the set $A_f \cap \{\|x\| = 1\}$ over the m-dimensional space

$$E_m = \{0\}^{l_f} \times \mathbb{R}^m \times \{0\}^{\{k:k\ge l_f+m+1\}}$$

generated by the coordinates from $l_f + 1$ to $l_f + m$. It can be shown that the resulting set is

$$A_{f,m} = \{\boldsymbol{\theta} = (\theta_k)_{k\ge1} : \forall\ 1 \le k \le l_f,\ \theta_k = 0,$$
$$\text{and } \forall\ l_f + 1 \le k \le l_f + m,\ 0 \le \theta_k \le \sqrt{f(k)},$$
$$\text{and } \forall\ k \ge l_f + m + 1,\ \theta_k = 0\}.$$

The same type of reasoning as in the proof of Lemma 2 leads us to

$$\begin{aligned}\mathcal{D}_\epsilon(\Theta_f, d) &\ge \mathcal{D}_\epsilon(A_{f,m}, \|\cdot\|_{\ell^2}) \\ &\ge \mathcal{N}_\epsilon\left(\prod_{k=l_f+1}^{l_f+m}\left[0, \sqrt{f(k)}\right], \|\cdot\|_{\mathbb{R}^m}\right) \\ &\ge \frac{\mathrm{Vol}\left(\prod_{k=l_f+1}^{l_f+m}\left[0, \sqrt{f(k)}\right]\right)}{\mathrm{Vol}\left(B_{\mathbb{R}^m}(0,\epsilon)\right)} \\ &\ge \left(\frac{1}{\epsilon}\right)^m \frac{1}{\mathrm{Vol}\left(B_{\mathbb{R}^m}(0,1)\right)} \prod_{k=l_f+1}^{l_f+m}\sqrt{f(k)}.\end{aligned}$$

It only remains to apply the logarithm function. ∎

Proof of Corollary 2: Here we have $f(k) = \min(1, Ce^{-\alpha k})$. Since $\sum_{k\ge l_f+1} f(k) \le 1$, then $Ce^{-\alpha k} \le 1$ for all $k \ge l_f + 1$. Therefore

$$\ln f(k) = -\alpha k + \ln C$$

for all $k \ge l_f + 1$, and

$$\begin{aligned}\frac{1}{2}\sum_{k=l_f+1}^{l_f+m} \ln f(k) &= \frac{m}{2}\ln C - \frac{\alpha}{2}\sum_{k=l_f+1}^{l_f+m} k \\ &= -(1+o(1))\frac{\alpha}{4}m^2,\end{aligned}$$

where $o(1)$ is a function $g(m)$ such that $g(m) \to 0$ as $m \to \infty$. From Lemma 3 we obtain

$$\mathcal{H}_\epsilon(\Theta_f, d) \ge m\ln\left(\frac{1}{\epsilon}\right) + (1+o(1))\frac{m}{2}\ln m - (1+o(1))\frac{\alpha}{4}m^2.$$

Let us choose

$$m(\epsilon) = \left\lfloor \frac{2}{\alpha}\ln\left(\frac{1}{\epsilon}\right)\right\rfloor.$$

Then the term $\frac{m}{2}\ln m$ becomes negligible and we get the desired result. ∎

Appendix B
Proof of Proposition 4

Proof: Let $\mathbb{E}_{\boldsymbol{\theta}}$ denote the expected value under the distribution $\boldsymbol{P}_{\boldsymbol{\theta}}$. The calculus is made using natural logarithm:

$$\begin{aligned}&(\ln 2)\, D(P^n_{\boldsymbol{\theta}}; KT_k) \\ &= \mathbb{E}_{\boldsymbol{\theta}} \ln \frac{P^n_{\boldsymbol{\theta}}(X_{1:n})}{KT_k(X_{1:n})} \\ &= \mathbb{E}_{\boldsymbol{\theta}} \sum_{i=1}^k T_i \ln\theta_i - \mathbb{E}_{\boldsymbol{\theta}} \ln \frac{D_k(T_1+\frac12, \cdots, T_k+\frac12)}{D_k(\frac12, \cdots, \frac12)} \\ &= \ln\frac{\Gamma(1/2)^k}{\Gamma(k/2)} + \overbrace{\sum_{i=1}^k n\theta_i\ln\theta_i - \mathbb{E}_{\boldsymbol{\theta}}\ln\frac{\prod_{i=1}^k\Gamma(T_i+\frac12)}{\Gamma(n+\frac{k}{2})}}^{\text{(A)}}.\end{aligned} \tag{19}$$

The second line comes from (5) and (6), and the third from (7). We use now the Feller bounds given in (17), with β_0 denoting the coefficient corresponding to the formula of $\Gamma(n+\frac{k}{2})$, and β_i being the coefficient in the formula of $\Gamma(T_i+\frac12)$.

$$\begin{aligned}\text{(A)} &= \sum_{i=1}^k n\theta_i\ln\theta_i - \mathbb{E}_{\boldsymbol{\theta}}\ln\frac{\prod_{i=1}^k\left(\sqrt{2\pi}(T_i+\frac12)^{T_i}\right)}{\sqrt{2\pi}(n+\frac{k}{2})^{n+(k-1)/2}} \\ &\quad - \sum_{i=1}^k \mathbb{E}_{\boldsymbol{\theta}}\frac{\beta_i}{12(T_i+\frac12)} + \frac{\beta_0}{12(n+\frac{k}{2})} \\ &\ge -\frac{k-1}{2}\ln 2\pi - \frac{k}{6} \\ &\quad + \overbrace{\sum_{i=1}^k\left(n\theta_i\ln\theta_i - \mathbb{E}_{\theta_i}T_i\ln\left(T_i+\frac12\right)\right)}^{\text{(B)}} \\ &\quad + \overbrace{\left(n+\frac{k-1}{2}\right)\ln\left(n+\frac{k}{2}\right)}^{\text{(C)}}.\end{aligned}$$

Now we lower bound separately (B) and (C). For the later,

$$\begin{aligned}\text{(C)} &= \left(n+\frac{k-1}{2}\right)\ln n + \left(n+\frac{k-1}{2}\right)\ln\left(1+\frac{k}{2n}\right) \\ &\ge n\ln n + \frac{k-1}{2}\ln n.\end{aligned}$$

On the other hand,

$$\begin{aligned}\text{(B)} &= -n\ln n + \sum_{i=1}^k \overbrace{\left(n\theta_i\ln n\theta_i - \mathbb{E}_{\theta_i}T_i\ln T_i\right)}^{\text{(B}_i\text{)}} \\ &\quad - \sum_{i=1}^k \overbrace{\mathbb{E}_{\theta_i}T_i\ln\left(1+\frac{1}{2T_i}\right)}^{\le 1/2}.\end{aligned}$$

From the relation $\ln t \le t - 1$, with $t = \frac{T_i}{n\theta_i}$, we get

$$\ln T_i - \ln n\theta_i \le \frac{T_i - n\theta_i}{n\theta_i}$$
$$T_i\ln T_i - T_i\ln n\theta_i \le \frac{(T_i-n\theta_i)^2}{n\theta_i} + (T_i - n\theta_i)$$
$$\mathbb{E}_{\theta_i}T_i\ln T_i - n\theta_i\ln n\theta_i \le \frac{\mathrm{Var}\,T_i}{n\theta_i} = 1 - \theta_i.$$

As a consequence,

$$(\mathrm{B}_i) \geq -(1-\theta_i) \geq -1,$$

$$(\mathrm{B}) \geq -n\ln n - \frac{3k}{2},$$

and

$$(\mathrm{A}) \geq \frac{k-1}{2}\ln\frac{n}{2\pi} - \frac{5k}{3}.$$

All that remains is to collect the different elements of (19) to get the announced result. ∎

Appendix C
Redundancy of ACcode

A. Moments of M_n

Before completing the proofs announced in Section V, we need a lemma which contains several useful results about the moments of M_n.

Lemma 7: Let C and α be positive numbers satisfying $C > e^{2\alpha}$. Then, for all $n \geq 1$,

1)

$$\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_P[M_n] \leq \frac{1}{\alpha}\ln n + \frac{1}{\alpha} + \ln\frac{C}{1-e^{-\alpha}}.$$

2)

$$\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_P[M_n \mathbb{1}_{M_n > \frac{2}{\alpha}\ln n + \ln\frac{C}{1-e^{-\alpha}}}] = O\left(\frac{\ln n}{n}\right).$$

3)

$$\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_P[M_n \ln M_n] = o(\ln^2 n).$$

Proof: Let F denote the distribution function associated with P. For $t \geq 0$, we have

$$\begin{aligned}\mathbb{P}(X_1 > t) &= \sum_{k\geq\lfloor t\rfloor+1} P(k)\\ &\leq \frac{C}{1-e^{-\alpha}} e^{-\alpha(\lfloor t\rfloor+1)}\\ &\leq e^{-\alpha(t-\beta)},\end{aligned}$$

where $\beta = \ln\frac{C}{1-e^{-\alpha}}$. Therefore $F(t) \geq G(t)$ for all $t \in \mathbb{R}$, where

$$G(t) = \mathbb{1}_{t\geq\beta}\left(1 - e^{-\alpha(t-\beta)}\right).$$

We can identify in G the distribution function of the random variable $\beta + Y$, where Y follows the exponential distribution with parameter α.

Let $U_1, \ldots, U_n$ be n iid random variables following the uniform distribution on $[0,1]$. For $1 \leq i \leq n$, let us define

$$\begin{aligned}X_i' &= F^{-1}(U_i)\\ Y_i &= G^{-1}(U_i) - \beta,\end{aligned}$$

where F^{-1} and G^{-1} denote the pseudo-inverses of F and G:

$$\forall t \in [0,1], \quad F^{-1}(t) = \inf\{x \in \mathbb{R} : F(x) \geq t\}.$$

Then the n-dimensional vector $X'_{1:n} = (X'_1, \ldots, X'_n)$ has the same distribution as $X_{1:n}$, and the maxima $M'_n = \sup_{1\leq i\leq n} X'_i$ and M_n follow the same distribution.

On the other hand, the relation $F \geq G$ entails $X'_i \leq \beta + Y_i$, for all $1 \leq i \leq n$. As the consequence, if $M''_n = \sup_{1\leq i\leq n} Y_i$ denotes the maximum of all Y_i, we have $M'_n \leq \beta + M''_n$. Since the random variables Y_i are independent, the probability distribution of M''_n is easy to calculate. Indeed for all $t > 0$,

$$\begin{aligned}\mathbb{P}(M''_n \leq t) &= \mathbb{P}(\forall\ 1 \leq i \leq n,\ Y_i \leq t)\\ &= \left(1 - e^{-\alpha t}\right)^n.\end{aligned}$$

We can write down the density function of M''_n:

$$f(t) = \begin{cases} n\,\alpha\, e^{-\alpha t}(1-e^{-\alpha t})^{n-1} & \text{if } t > 0,\\ 0 & \text{otherwise.}\end{cases}$$

Now we look for an upper bound of $\mathbb{E}[M_n]$ by taking advantage of the knowledge of that distribution:

$$\begin{aligned}\mathbb{E}[M_n] &= \mathbb{E}[M'_n]\\ &\leq \mathbb{E}[\beta + M''_n]\\ &= \beta + \int_0^\infty t\, n\,\alpha\, e^{-\alpha t}(1-e^{-\alpha t})^{n-1}\,\mathrm{d}t\\ &= \beta + \int_0^\infty \left(1 - (1-e^{-\alpha t})^n\right)\,\mathrm{d}t\end{aligned}$$

integrating by parts. Use now the change of variables

$$\begin{cases} u = 1 - e^{-\alpha t}\\ t = \frac{-\ln(1-u)}{\alpha}\end{cases}$$

$$\begin{aligned}\mathbb{E}[M_n] &\leq \beta + \frac{1}{\alpha}\int_0^1 \frac{1-u^n}{1-u}\mathrm{d}u\\ &\leq \frac{1}{\alpha}\ln n + \frac{1}{\alpha} + \ln\frac{C}{1-e^{-\alpha}}.\end{aligned}$$

Since the upper bound does not depend on P, that achieves the proof of the point 1. We can handle the point 2 in the same way. For all $t > 0$, we have

$$\begin{aligned}\mathbb{E}\left[M_n \mathbb{1}_{M_n > \beta + t}\right] &\leq \mathbb{E}\left[(\beta + M''_n)\mathbb{1}_{M''_n > t}\right]\\ &\leq \int_t^\infty (\beta + u) n\,\alpha\, e^{-\alpha u}\,\mathrm{d}u\\ &= n e^{-\alpha t}\left(t + \frac{1}{\alpha} + \beta\right).\end{aligned}$$

With $t = \frac{2}{\alpha}\ln n$, we get the second point of Lemma 7.

The third item is not very different. Since the function $x \mapsto x\ln x$ is increasing on $[1, +\infty)$ and $1 \leq M'_n \leq \beta + M''_n$, we have

$$\begin{aligned}&\mathbb{E}\left[M_n \ln M_n\right]\\ &\quad\leq \mathbb{E}\left[(\beta + M''_n)\ln(\beta + M''_n)\right]\\ &\quad= \mathbb{E}\left[\mathbb{1}_{M''_n \leq \beta}(\beta + M''_n)\ln(\beta + M''_n)\right]\\ &\qquad+ \mathbb{E}\left[\mathbb{1}_{M''_n > \beta}\mathbb{1}_{M''_n \leq \frac{2}{\alpha}\ln n}(\beta + M''_n)\ln(\beta + M''_n)\right]\\ &\qquad+ \mathbb{E}\left[\mathbb{1}_{M''_n > \beta}\mathbb{1}_{M''_n > \frac{2}{\alpha}\ln n}(\beta + M''_n)\ln(\beta + M''_n)\right]\\ &\quad\leq 2\beta\ln(2\beta) + \frac{4}{\alpha}(\ln n)\ln\left(\frac{4}{\alpha}\ln n\right)\\ &\qquad+ \mathbb{E}\left[2M''_n\ln(2M''_n)\mathbb{1}_{M''_n > \frac{2}{\alpha}\ln n}\right]\\ &\quad\leq 2\beta\ln(2\beta) + \left(\frac{4}{\alpha}\ln\frac{4}{\alpha}\right)\ln n + \frac{4}{\alpha}(\ln n)(\ln\ln n)\\ &\qquad+ \mathbb{E}\left[4{M''_n}^2\mathbb{1}_{M''_n > \frac{2}{\alpha}\ln n}\right].\end{aligned}$$

Let us define

$$\gamma(n) = 2\beta\ln(2\beta) + \left(\frac{4}{\alpha}\ln\frac{4}{\alpha}\right)\ln n + \frac{4}{\alpha}(\ln n)(\ln\ln n).$$

Note that $\gamma(n) = o(\ln^2 n)$. Then

$$\begin{aligned}\mathbb{E}\left[M_n \ln M_n\right] &\le \gamma(n) + \int_{\frac{2}{\alpha}\ln n}^{\infty} 4u^2 n\,\alpha\, e^{-\alpha u}\,\mathrm{d}u \\ &= \gamma(n) + \frac{4ne^{-2\ln n}}{\alpha^2}(4\ln^2 n + 4\ln n + 2).\end{aligned}$$

Taking the supremum over P, we get

$$\begin{aligned}\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_P\left[M_n \ln M_n\right] &\le \gamma(n) + \frac{16\ln^2 n + 16\ln n + 8}{\alpha^2\, n} \\ &= o(\ln^2 n).\end{aligned}$$

■

B. Contribution of C4

We make here several calculus we announced in subsection V-A.

Proof of (13): Let

$$\hat{P}^n(x_{1:n}) = \sup_{P^n} P^n(x_{1:n}) = \prod_{j\in\{x_1,\dots,x_n\}} \left(\frac{n_j^n}{n}\right)^{n_j^n}$$

be the maximum likelihood of the string $x_{1:n}$ over all iid distribution on $\mathbb{N}^n$. Then

$$\begin{aligned}(\mathrm{A}_1) &\le \mathbb{E}_{P^n}\left[\log\frac{\hat{P}^n(X_{1:n})}{KT_{M_n}(X_{1:n})}\right] \\ &\le \mathbb{E}_{P^n}\left[\sup_{x_{1:n}\in\{1,\dots,M_n\}} \log\frac{\hat{P}^n(x_{1:n})}{KT_{M_n}(x_{1:n})}\right]\end{aligned}$$

Now we can apply a result from Catoni ([9, prop 1.4.1]):

Lemma 8: For all $k \ge 1$ and for all $x_{1:n} \in \{1,\dots,k\}^n$,

$$-\log KT_k(x_{1:n}) + \log\hat{P}^n(x_{1:n}) \le \frac{k-1}{2}\log n + \log k.$$

Therefore

$$(\mathrm{A}_1) \le \frac{\mathbb{E}[M_n]-1}{2}\log n + \mathbb{E}[\log M_n].$$

■

Proof of (14): We group the terms in (A_2), M_n by M_n:

$$\begin{aligned}(\mathrm{A}_2) &\ge \mathbb{E}_{P^n}\left[\sum_{i=1}^{n-1}\log\frac{2i+M_n}{2i+M_i} + \sum_{i=1}^{n-1}\log\frac{2i+M_i}{2i+1+M_i}\right] \\ &\ge \mathbb{E}_{P^n}\left[\sum_{k=1}^{m-1}\sum_{i=kM_n+1}^{(k+1)M_n}\log\left(1+\frac{M_n-M_i}{2i+M_i}\right)\right] \\ &\quad - \mathbb{E}_{P^n}\left[\sum_{i=1}^{n-1}\log\left(1+\frac{1}{2i+M_i}\right)\right].\end{aligned}$$

We have

$$\sum_{i=1}^{n-1}\log\left(1+\frac{1}{2i+M_i}\right) \le \sum_{i=1}^{n-1}\frac{1}{2i} \le \frac{1+\ln(n-1)}{2}.$$

From the relation $M_k \le M_{k'}$ for all $k' \ge k \ge 1$, we can infer, for all $i \ge kM_n$,

$$\frac{M_n - M_i}{2i+M_i} \le \frac{M_n}{2kM_n} = \frac{1}{2k}$$

Since $\log$ is a concave function, we have $\log(1+x) \ge \frac{x\log(1+a)}{a}$ for all $a > 0$ and $0 \le x \le a$. Consequently, if we choose $a = \frac{1}{2k}$,

$$\begin{aligned}(\mathrm{A}_2) &\ge \mathbb{E}_{P^n}\left[\sum_{k=1}^{m-1}\sum_{i=kM_n+1}^{(k+1)M_n} 2k\log\left(1+\frac{1}{2k}\right)\frac{M_n-M_i}{2i+M_i}\right] \\ &\quad - \frac{1+\ln n}{2} \\ &\ge \mathbb{E}_{P^n}\left[\sum_{k=1}^{m-1} 2k\log\left(1+\frac{1}{2k}\right) M_n\frac{M_n - M_{(k+1)M_n}}{2(k+1)M_n + M_n}\right] \\ &\quad - \frac{1+\ln n}{2} \\ &\ge \mathbb{E}_{P^n}\left[\sum_{k=h_n+1}^{m}\lambda_n\frac{M_n - M_{kM_n}}{2k+1}\right] - \frac{1+\ln n}{2}.\end{aligned}$$

■

Proof of (15): We have

$$\begin{aligned}(\mathrm{A}_3) = \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \Bigg[\sum_{j\ge1} P^n(M_n = j) \\ \times\left(j\log n - \lambda_n j\sum_{k=h_n+1}^{\lfloor\frac{n-1}{j}\rfloor}\frac{1}{k+\frac{1}{2}}\right)\Bigg].\end{aligned}$$

Then we plug in $h_n = \lfloor \ln n - 3/2\rfloor$. For n big enough, $h_n \ge 1$, and we have

$$\begin{aligned}j\sum_{k=h_n+1}^{\lfloor\frac{n-1}{j}\rfloor}\frac{1}{k+\frac{1}{2}} &\ge j\int_{\ln n-1/2}^{\lfloor\frac{n-1}{j}\rfloor+1}\frac{\mathrm{d}x}{x+\frac{1}{2}} \\ &= j\left(\ln\left(\left\lfloor\frac{n-1}{j}\right\rfloor+\frac{3}{2}\right) - \ln(\ln n)\right) \\ &\ge j\ln(n-1) - j\ln j - j\ln(\ln n),\end{aligned}$$

and therefore

$$\begin{aligned}(\mathrm{A}_3) \le \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \big[&(\log e - \lambda_n)\mathbb{E}[M_n]\ln n \\ &+ \lambda_n\mathbb{E}[M_n]\ln\frac{n}{n-1} \\ &+ \lambda_n\mathbb{E}[M_n\ln M_n] + \lambda_n\mathbb{E}[M_n]\ln(\ln n)\big].\end{aligned}$$

Then, if we use Lemma 7 and the fact that λ_n tends to $\log e$, we get (15). ■

Proof of (16): We want to commute the expected value and the sum in (A_4). To do it, we need to get rid of the

m. We can note that the condition $k \leq m = \lfloor \frac{n-1}{M_n} \rfloor$ entails $kM_n \leq n-1$. Consequently, for n big enough,

$$\begin{aligned}(\mathrm{A}_4) &\leq \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_{P^n}\left[\sum_{k=3}^{m} \frac{M_{kM_n}}{k+\frac{1}{2}}\right] \\ &\leq \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_{P^n}\left[\sum_{k=3}^{n-1} \frac{M_{kM_n}\mathbb{1}_{kM_n\leq n-1}}{k+\frac{1}{2}}\right] \\ &\leq \sum_{k=3}^{n-1} \frac{\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_{P^n}[M_{kM_n}\mathbb{1}_{M_n\leq l_n}]}{k+\frac{1}{2}} \\ &\qquad + \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}_{P^n}[M_n\mathbb{1}_{M_n>l_n}] \sum_{k=3}^{n-1} \frac{1}{k+\frac{1}{2}},\end{aligned}$$

where $l_n = \left\lfloor \frac{2}{\alpha}\ln n + \ln\frac{C}{1-e^{-\alpha}} \right\rfloor$. We can now plug in the results of Lemma 7:

$$\begin{aligned}(\mathrm{A}_4) &\leq \sum_{k=3}^{n-1} \frac{\frac{1}{\alpha}\ln(kl_n) + \frac{1}{\alpha} + \ln\frac{C}{1-e^{-\alpha}}}{k+\frac{1}{2}} + o(1)\sum_{k=3}^{n-1}\frac{1}{k+\frac{1}{2}} \\ &\leq \frac{1}{\alpha}\sum_{k=3}^{n-1}\frac{\ln k}{k+\frac{1}{2}} + \frac{1}{\alpha}(\ln l_n + O(1))\sum_{k=3}^{n-1}\frac{1}{k+\frac{1}{2}}.\end{aligned}$$

Note that $l_n = O(\ln n)$, and consequently $\ln l_n = O(\ln\ln n)$. So

$$\begin{aligned}(\mathrm{A}_4) &\leq \frac{1}{\alpha}\int_3^n \frac{\ln x}{x-\frac{1}{2}}\,\mathrm{d}x + O(\ln\ln n)\int_3^n \frac{\mathrm{d}x}{x-\frac{1}{2}} \\ &\leq \frac{1}{2\alpha}\ln^2 n + o(\ln^2 n).\end{aligned}$$

■

C. Contribution of C2

We prove here several lemmas used in subsection V-B to prove Proposition 6.

Proof of Lemma 4: Let $\boldsymbol{P}$ be an element of $\Lambda_{Ce^{-\alpha\cdot}}$. Let us define, for all $k \geq 0$,

$$\bar{p}(k) = P(X_1 > k) = \sum_{j\geq k+1} P(k),$$

and

$$(\mathrm{B}_1) = \sum_{i=1}^{n} \mathbb{E}_{P^n}\left[\mathbb{1}_{X_i>M_{i-1}} g(X_i)\right].$$

Note that, for all $1 \leq i \leq n$, X_i and M_{i-1} are independent random variables, and

$$\begin{aligned}P^n(M_i \leq k) &= P^n(\forall\ 1\leq j\leq i,\ X_j \leq k) \\ &= (1-\bar{p}(k))^i.\end{aligned}$$

Then we can write

$$\begin{aligned}(\mathrm{B}_1) &= \sum_{i=1}^{n}\sum_{k\geq 0} P^n(M_{i-1}=k)\sum_{m\geq k+1} P(m)g(m) \\ &= \sum_{m\geq 1} P(m)g(m)\sum_{i=1}^{n}\sum_{k=0}^{m-1}\mathbb{P}(M_{i-1}=k) \\ &= P(1)g(1) + \sum_{m\geq 2} P(m)g(m)\sum_{i=1}^{n}(1-\bar{p}(m-1))^{i-1} \\ &= \sum_{m\geq 1} P(m)g(m)\frac{1-(1-\bar{p}(m-1))^n}{\bar{p}(m-1)}.\end{aligned}$$

If we take $g(x) = 1$ for all x, we get

$$\begin{aligned}\sum_{m\geq 1} P(m)\frac{1-(1-\bar{p}(m-1))^n}{\bar{p}(m-1)} &= \mathbb{E}\left[\sum_{i=1}^{n}\mathbb{1}_{X_i>M_{i-1}}\right] \\ &\leq \mathbb{E}[M_n].\end{aligned}$$

In the general case, we can split the sum at K_n, and we get

$$\begin{aligned}(\mathrm{B}_1) &= \sum_{m=1}^{K_n} P(m)g(m)\frac{1-(1-\bar{p}(m-1))^n}{\bar{p}(m-1)} \\ &\qquad + \sum_{m\geq K_n+1} P(m)g(m)\frac{1-(1-\bar{p}(m-1))^n}{\bar{p}(m-1)} \\ &\leq g(K_n)\sum_{m\geq 1} P(m)\frac{1-(1-\bar{p}(m-1))^n}{\bar{p}(m-1)} \\ &\qquad + \sum_{m\geq K_n+1} nP(m)g(m) \\ &\leq g(K_n)\mathbb{E}[M_n] + C\,n\sum_{m\geq K_n+1} g(m)e^{-\alpha m}.\end{aligned}$$

At this point, we can take the supremum over all sources $\boldsymbol{P}$ in $\Lambda_{Ce^{-\alpha\cdot}}$:

$$\begin{aligned}&\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \sum_{i=1}^{n}\mathbb{E}_{P^n}\left[\mathbb{1}_{X_i>M_{i-1}}g(X_i)\right] \\ &\quad\leq (1+o(1))g(K_n)\frac{1}{\alpha}\ln n + C\,n\int_{K_n}^{\infty} g(x+1)e^{-\alpha x}\,\mathrm{d}x.\end{aligned}$$

■

Proof of Lemma 5:

$$\begin{aligned}&n\int_{K_n}^{\infty} l_E(x+1)e^{-\alpha x}\,\mathrm{d}x \\ &\quad\leq n\int_{K_n}^{\infty}(\log(x+1) + 2\log\log(x+2) + 1)\,e^{-\alpha x}\,\mathrm{d}x \\ &\quad\leq ne^{-\alpha K_n}\log(K_n+2) \\ &\qquad \int_{K_n+2}^{\infty}\frac{\log x + 2\log\log x + 1}{\log(K_n+2)}e^{-\alpha(x-K_n-2)}\,\mathrm{d}x \\ &\quad\leq e^{\alpha}\log(K_n+2)\left(\sup_{x\geq K_n+2}\frac{\log x + 2\log\log x + 1}{\log x}\right) \\ &\qquad \int_{K_n+2}^{\infty}\frac{\log x}{\log(K_n+2)}e^{-\alpha(x-K_n-2)}\,\mathrm{d}x \\ &\quad= O(\log K_n)\int_0^{\infty}\left(1+\frac{\log\left(1+\frac{x}{K_n+2}\right)}{\log(K_n+2)}\right)e^{-\alpha x}\,\mathrm{d}x \\ &\quad= o(\log n).\end{aligned}$$

The supremum is correctly defined and bounded, because the function

$$x \mapsto \frac{\log x + 2\log\log x + 1}{\log x}$$

is continuous and tends to 1 as x tends to the infinity. ■

ACKNOWLEDGMENT

The author wishes to thank Prof. E. Gassiat for her helpful advice, for the many ideas in this paper she suggested him, and for her constant availability to his questions.

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, ser. Wiley Series in Telecommunications. New York: Wiley, 1991.

[2] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[3] L. D. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inf. Theory*, vol. 26, pp. 166–174, 1980.

[4] D. Haussler, "A general minimax result for relative entropy," University of California, UC Santa Cruz, CA 96064, Tech. Rep. UCSC-CRL-96-26, 1996. [Online]. Available: http://citeseer.ist.psu.edu/haussler96general.html

[5] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inf. Theory*, vol. 27, no. 2, pp. 199–207, 1981.

[6] Q. Xie and A. R. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 646–657, 1997.

[7] ——, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 431–445, 2000.

[8] A. R. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2743–2760, 1998.

[9] O. Catoni, *Statistical Learning Theory and Stochastic Optimization*, ser. Lecture Notes in Mathematics. Springer-Verlag, 2001, vol. 1851, École d'Été de Probabilités de Saint-Flour XXXI.

[10] M. Drmota and W. Szpankowski, "Precise minimax redundancy and regret," *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2686–2707, 2004.

[11] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of bayes methods," *IEEE Trans. Inf. Theory*, vol. 36, no. 3, pp. 453–471, 1990.

[12] ——, "Jeffrey's prior is asymptotically least favorable under entropy risk," *J. Statist. Planning Inference*, vol. 41, pp. 37–60, 1994.

[13] A. R. Barron, "Information-theoretic characterization of bayes performance and the choice of priors in parametric and nonparametric problems," in *Bayesian Statistics*, J. M. Bernardo, J. O. Berger, D. A. P., and S. A. F. M., Eds. Oxford Univ. Press, 1998, vol. 6, pp. 27–52.

[14] L. D. Davisson, "Minimax noiseless universal coding for markov sources," *IEEE Trans. Inf. Theory*, vol. 29, no. 2, pp. 211–214, 1983.

[15] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, 1995.

[16] K. Atteson, "The asymptotic redundancy of bayes rules for markov chains," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 2104–2109, 1999.

[17] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inf. Theory*, vol. 30, no. 4, pp. 629–636, 1984.

[18] I. Csiszár and P. C. Shields, "Redundancy rates for renewal and other processes," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 2065–2072, 1996.

[19] P. Flajolet and W. Szpankowski, "Analytic variations on redundancy rates of renewal processes," *IEEE Trans. Inf. Theory*, vol. 48, no. 11, pp. 2911–2921, 2002.

[20] P. C. Shields, "Universal redundancy rates do not exist," *IEEE Trans. Inf. Theory*, vol. 39, no. 2, pp. 520–524, 1993.

[21] J. C. Kieffer, "A unified approach to weak universal source coding," *IEEE Trans. Inf. Theory*, vol. 24, no. 6, pp. 674–682, 1978.

[22] L. Györfi, I. Páli, and E. C. van der Meulen, "There is no universal source code for an infinite source alphabet," *IEEE Trans. Inf. Theory*, vol. 40, no. 1, pp. 267–271, 1994.

[23] ——, "On universal noiseless source coding for infinite source alphabets," *Eur. Trans. Telecom.*, vol. 4, pp. 4–16, 1993.

[24] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, pp. 1469–1481, 2004.

[25] ——, "Speaking of infinity [i.i.d. strings]," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2215–2230, 2004.

[26] N. Jevtic, A. Orlitsky, and N. P. Santhanam, "A lower bound on compression of unknown alphabets," *Theor. Comput. Sci.*, vol. 332, no. 1-3, pp. 293–311, 2005.

[27] A. Orlitsky, N. P. Santhanam, K. Viswanathan, and J. Zhang, "Limit results on pattern entropy," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 2954–2964, 2006.

[28] G. I. Shamir and D. J. J. Costello, "On the entropy rate of pattern processes," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1620–1635, 2004.

[29] G. I. Shamir, "Sequential universal lossless techniques for compression of patterns and their description length," in *Data Compression Conference*, 2004, pp. 419–428. [Online]. Available: http://csdl.computer.org/comp/proceedings/dcc/2004/2082/00/20820419abs.htm

[30] G. M. Gemelos and T. Weissman, "On the entropy rate of pattern processes," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3994–4007, 2006. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TIT.2006.880044

[31] A. Garivier, "A lower bound for the maximin redundancy in pattern coding," *IEEE Trans. Inf. Theory*, submitted for publication. [Online]. Available: http://www.math.u-psud.fr/~garivier/doc/articles/binf.ps

[32] Y. Choi and W. Szpankowski, "Pattern matching in constrained sequences," in *2007 Int. Symp. Information Theory*, Nice, 2007, pp. 2606–2610.

[33] P. Elias, "Universal codeword sets and representations of the integers," *IEEE Trans. Inf. Theory*, vol. 21, no. 2, pp. 194–203, 1975.

[34] D.-k. He and E.-h. Yang, "The universality of grammar-based codes for sources with countably infinite alphabets," *IEEE Trans. Inf. Theory*, vol. 51, no. 11, pp. 3753–3765, 2005.

[35] D. P. Foster, R. A. Stine, and A. J. Wyner, "Universal codes for finite sequences of integers drawn from a monotone distribution," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1713–1720, 2002.

[36] S. Boucheron, A. Garivier, and E. Gassiat, "Coding on countably infinite alphabets," *IEEE Trans. Inf. Theory*, submitted for publication. [Online]. Available: http://www.math.u-psud.fr/~gassiat/nml.pdf

[37] D. Haussler and M. Opper, "Mutual information, metric entropy and cumulative relative entropy risk," *Ann. Statist.*, vol. 25, no. 6, pp. 2451–1492, 1997. [Online]. Available: http://www.cbse.ucsc.edu/staff/haussler_pubs/mutual.pdf

[38] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*, ser. Springer Series in Statistics. Springer-Verlag, 1996.

[39] D. Bontemps, "Redondance bayésienne et minimax, sources stationnaires sans mémoire en alphabet infini," Master's thesis, Paris-XI University, sep 2007. [Online]. Available: http://www.math.u-psud.fr/~bontemps/rapportM2.pdf

[40] G. Pisier, *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge, UK: Cambridge Univ. Press, 1989.

[41] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 4th ed., ser. Cambridge Mathematical Library. Cambridge: Cambridge Univ. Press, 1927, reprinted 1990.