Universal Coding on Infinite Alphabets: Exponentially Decreasing Envelopes

Dominique Bontemps *Member, IEEE*

***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, an adaptive algorithm is proposed, whose maximum redundancy is equivalent to the minimax redundancy.

***Index Terms*—Adaptive compression, Bayes mixture, Data compression, Infinite countable alphabets, Redundancy, Universal coding**

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 data compression on an unknown alphabet. This has applications in areas such as language modeling or lossless multimedia codecs.

First, we present briefly the problematics of data compression. 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. Lossless data compression

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 data compression 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*.

D. Bontemps is PhD fellowship holder at Laboratoire de Mathématiques d'Orsay, Univ Paris-Sud, Orsay, F-91405.

Manuscript received June 03, 2008; revised Mars 24, 2010.

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 Kullback-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 Λ of sources (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

$$M_{n,\mu}(x_{1:n}) = \int_{\Lambda} P^n(x_{1:n})\,\mathrm{d}\mu(\boldsymbol{P}).$$

When Λ is a class of iid sources on the set $\mathcal{X} = \mathbb{N}_* = \mathbb{N}\backslash\{0\}$, 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.

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 on a finite alphabet 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 the envelope function f be defined by $f(x) = \sup_{\boldsymbol{P}\in\Lambda} P(x)$. Then the minimax regret satisfies

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

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

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

We present two 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 knowledge, 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 $\sqrt{n}$ have been obtained.

Once the minimax redundancy of a class of sources is known, we are interested in finding a minimax coding algorithm. Section III proposes a new adaptive coding algorithm, and we show that its maximum redundancy is equivalent to the minimax redundancy of exponentially decreasing envelope classes.

Eventually, the Appendix contains some proofs and some auxiliary results 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 such that $C > e^{2\alpha}$. The minimax redundancy of the exponentially decreasing envelope class $\Lambda_{Ce^{-\alpha\cdot}}$ satisfies

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

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:

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

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 the metric entropy to the 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 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

$$h(P,Q) = \sqrt{\sum_{k\geq 1}\left(\sqrt{P(k)}-\sqrt{Q(k)}\right)^2}.$$

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

$$d(\boldsymbol{\theta},\boldsymbol{\theta}') = h(P_{\boldsymbol{\theta}},P_{\boldsymbol{\theta}'}) = \sqrt{\sum_{k\geq 1}\left(\sqrt{\theta_k}-\sqrt{\theta'_k}\right)^2}.$$

With a metric we can define the *metric entropy*. We need to consider first some numbers.

Definition 3: Let S be a subset of Θ, and ϵ be a positive number.

1) We denote by $\mathcal{D}_\epsilon(S,d)$ the cardinality of the smallest finite partition of S with sets of diameter at most ϵ, or we set $\mathcal{D}_\epsilon(S,d)=\infty$ if no such finite partition exists.
2) The metric entropy of (S,d) is defined by
$$\mathcal{H}_\epsilon(S,d) = \ln\mathcal{D}_\epsilon(S,d).$$ [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 $d(x,y)<\epsilon$. The *covering number* $\mathcal{N}_\epsilon(S,d)$ is the cardinality of the smallest finite ϵ-cover of S, or we define $\mathcal{N}_\epsilon(S,d)=\infty$ 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, $d(x,y)>\epsilon$. The *packing number* $\mathcal{M}_\epsilon(S,d)$ is the cardinality of the largest finite ϵ-separated subset of S, or we define $\mathcal{M}_\epsilon(S,d)=\infty$ 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 be a subset of Θ. For all $\epsilon>0$,

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

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 generalizes the classical notion of dimension. But the metric entropy lets us know in some way how dense the elements are in a set, even infinite dimensional.

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

$$R_\lambda(\Lambda) = \inf_Q \sup_{\boldsymbol{\theta}\in\Theta_\lambda}\sum_{k\geq 1}P_{\boldsymbol{\theta}}(k)^{1+\lambda}Q(k)^{-\lambda},$$

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 $R_\lambda(\Lambda_f)<\infty$ for all $\lambda>0$. Indeed the choice

$$Q(k)=\frac{f(k)}{\sum_{l\geq 1}f(l)}$$

leads to the relation

$$R_\lambda(\Lambda_f)\leq\left(\sum_{k\geq 1}f(k)\right)^\lambda.$$

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

[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.

[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].

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 $R_\lambda(\Lambda) < \infty$. Let $h(x)$ be a continuous, non-decreasing 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 $C(\ln x)^\eta$, with $\eta \geq 0$.

The first case in the theorem is the one we use for exponentially decreasing envelope classes. In this case, the fast decreasing envelope produces a “not too big” metric entropy. Theorem 3 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 $(\ln n)^{\alpha/(\alpha+2)}$ between 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 now develop these applications.

[3]The $(\log e)$ factor comes from the use of the logarithm taken to base 2, in the definition of R_n.

B. The minimax redundancy of exponentially decreasing envelope classes

We want to apply Theorem 3 in order to prove Theorem 2, and therefore we have to calculate the metric entropy of exponentially decreasing envelope classes:

Proposition 3: Let C and α be positive numbers such that $C > e^{2\alpha}$. The metric entropy of the parameter set $\Theta_{C,\alpha}$ satisfies
$$\mathcal{H}_\epsilon(\Theta_{C,\alpha}, d) = (1+o(1))\frac{1}{\alpha}\ln^2(1/\epsilon),$$
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. ■

We make first some general considerations about the entropy of envelope classes, and then prove Proposition 3. 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.

Next we 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 of the metric entropy of (Θ_f, d). The outlines of the proofs of the next lemmas can be found in Appendix A. We start with the upper bound.

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

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

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.

Next we state a lower bound on the metric entropy. 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

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

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

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

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

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

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

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

Let us now apply these two results to the exponentially decreasing envelope classes.

Proof of Proposition 3: If we apply Lemma 2 we obtain the upper bounds

$$N_\epsilon \leq \frac{2}{\alpha} \ln(1/\epsilon) + \frac{1}{\alpha} \ln \frac{16C}{1-e^{-\alpha}}$$
$$B(\epsilon) \leq -\frac{\alpha}{4} N_\epsilon^2 + \left(\frac{\ln C}{2} + \frac{1}{\sqrt{1-e^{-\alpha}}}\right) N_\epsilon.$$

This leads to

$$\mathcal{H}_\epsilon(\Theta_{C,\alpha}, d) \leq (1+o(1)) \frac{1}{\alpha} \ln^2(1/\epsilon). \quad (3)$$

Consider now Lemma 3. Note that the exponentially decreasing envelope classes satisfy the condition $\sum_{k\geq 1} f(k) \geq 2$. Indeed their envelope is

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

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

With the choice $m = \left\lfloor \frac{2}{\alpha} \ln\left(\frac{1}{\varepsilon}\right) \right\rfloor$ we can infer the following lower bound:

$$\mathcal{H}_\epsilon(\Theta_{C,\alpha}, d) \geq (1+o(1)) \frac{1}{\alpha} \ln^2(1/\epsilon). \quad (4)$$

Since the lower bound (4) is the same as the upper bound (3), 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

$$\begin{aligned} &A(\alpha) n^{1/\alpha} \log\lfloor C\zeta(\alpha) \rfloor \\ &\quad \leq \mathbb{R}_n(\Lambda_{C\cdot^{-\alpha}}) \\ &\quad \leq \left(\frac{2Cn}{\alpha-1}\right)^{1/\alpha} (\log n)^{1-1/\alpha} + O(1), \end{aligned} \quad (5)$$

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 $\zeta(\alpha) = \sum_{k\geq 1} \frac{1}{k^\alpha}$, 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 $\epsilon > 0$,

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

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 $n \geq 1$,

$$c(1+o(1)) n^{1/\alpha} \leq R_n(\Lambda_{C\cdot^{-\alpha}}) \leq C(1+o(1)) n^{1/\alpha} \log n. \quad (6)$$

These inequalities improve in no way the result of [36]. May a better calculation 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. AUTOCENSURING CODE

This section presents a new algorithm called AutoCensuring Code (`ACcode`). It is in fact a modification of the Censuring Code proposed by Boucheron, Garivier and Gassiat in [36]. We keep the idea that big symbols are very few, and must be encoded differently, with an Elias code. Smaller symbols are encoded by arithmetic coding based on Krichevsky-Trofimov mixtures, which are known to be effective for finite alphabets. Our innovation is a data-driven cutoff $M_i = \sup_{1\leq k\leq i} X_k$ used to encode X_{i+1}: with this choice we do not need to know the exact parameters of the exponentially decreasing envelope.

`ACcode` is a prefix code on the set of all finite length messages, and it works on line. Its maximum redundancy on an exponentially decreasing envelope class $\Lambda_{Ce^{-\alpha\cdot}}$ is equivalent to the minimax redundancy of this class of sources.

Furthermore ACcode is adaptive, as the same algorithm satisfies this property with all exponentially decreasing envelope classes. This is formulated in the following theorem, proved in Appendix B.

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

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

The difference between the redundancy of ACcode and the minimax redundancy is not necessarily bounded: there may exist codes whose redundancy is smaller than the redundancy of ACcode, but with a benefit asymptotically negligible with respect to $\log^2 n$.

Additionally, Theorem 4 enables us to retrieve the upper bound of the minimax redundancy obtained in the section II.

Let us now define ACcode. Let $n \geq 1$ be some positive integer, and let $x_{1:n} = x_1 x_2 \ldots x_n$ be a string from $\mathbb{N}_*^n$ to be encoded. We define the sequence of maxima

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

The sequence $(m_i)_{1\leq i\leq n}$ is non-decreasing, piecewise constant. For $1 \leq i \leq n$, let $n_i^0 = \sum_{j=1}^{i} \mathbb{1}_{m_j > m_{j-1}}$ be the number of plateaus between 1 and i. For $1 \leq k \leq n_n^0$, let $\widetilde{m}_k$ be the k^{th} new maximum:

$$m_i = \widetilde{m}_{n_i^0}. \tag{7}$$

We define also $\widetilde{m}_0 = 0$. Let string $\widetilde{\boldsymbol{m}}$ be the sequence $(\widetilde{m}_1 - \widetilde{m}_0 + 1), \ldots, (\widetilde{m}_{n_n^0} - \widetilde{m}_{n_n^0 - 1} + 1), 1$. $\widetilde{\boldsymbol{m}}$ is encoded into a binary string C2 by applying Elias penultimate code (see [33]) to each number in $\widetilde{\boldsymbol{m}}$. It is a prefix code which uses $l_E(x)$ bits to encode a positive integer x, with

$$\begin{aligned} &l_E(1) = 1, \\ &l_E(x) = 1 + \lfloor \log x \rfloor + 2\lfloor \log \lfloor \log x \rfloor + 1 \rfloor \quad \text{ if } x \geq 2. \end{aligned} \tag{8}$$

Meanwhile the sequence of censored symbols is encoded using side information from $\widetilde{\boldsymbol{m}}$. Consider the censored sequence $\widetilde{x}_{1:n} = \widetilde{x}_1 \widetilde{x}_2 \ldots \widetilde{x}_n$ defined by

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

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. 0 has a special use in our setting: it makes the decoder to know when m_i changes, and that the new value has to be read in C2. We add at the end of $\widetilde{x}_{1:n}$ an additional 0, which acts as a termination signal together with the last 1 in $\widetilde{\boldsymbol{m}}$. This makes our code to be prefix on the set of all finite length messages (whatever n).

Therefore we produce the binary string C1 by arithmetic coding of $\widetilde{x}_{1:n}0$. The conditional coding probabilities are defined by

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

$$Q_{i+1}(\widetilde{X}_{i+1} = 0 | X_{1:i} = x_{1:i}) = \frac{1/2}{i + \frac{m_i+1}{2}},$$

where for $j \geq 1$ and $i \geq 0$, n_i^j is the number of occurrences of symbol j in $x_{1:i}$ (with convention $n_0^j = 0$ for all $j \geq 1$).

If $i \leq n-1$, 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

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

In the sequel we note the coding probability used to encode the entire string $\widetilde{x}_{1:n}0$ by

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

A remark we can do is that the symbol 0 is always considered as new: when $x_{i+1} > m_i$, we encode 0 but we increment the counter $n_i^{x_{i+1}}$. (This choice has been made to simplify the calculation of the redundancy of ACcode, but we suspect that changing this behavior could improve the performances.)

Now we have defined C1 and C2, we have to describe how they are transmitted. To keep our code on line, we overlap these two strings in the following way.

Arithmetic code needs a certain amount of bits, say l_i, to send the first i symbol of $\widetilde{x}_{1:n}$. Unfortunately, l_i depends on whether $i = n+1$ or not. In the previous case $l_{n+1} = \lceil -\log Q^{n+1}(\widetilde{x}_{1:n}0) \rceil + 1$, and in the later one l_i depends on the following symbols and has to be computed.

ACcode begins with C2, by the transmission of the Elias code of $\widetilde{m}_1 + 1$. Then the transmission of C1 is initiated. Suppose that $\widetilde{x}_i = 0$ and $n_i^0 = k$. As soon as l_i bits of C1 have been sent, the ACcode algorithm sends the Elias code of $\widetilde{m}_k - \widetilde{m}_{k-1} + 1$. Then C1 is transmitted again, from the next bit.

To decode the i^{th} symbol in C1, the knowledge of the current maximum m_{i-1} is needed; it is obtained from the beginning of the string C2. The decoder also needs the counters $(n_{i-1}^j)_{j\geq 1}$, which can be computed from the first $i-1$ decoded symbols. As soon as l_i bits of C1 have been received, $\widetilde{x}_i$ can be decoded. When the decoder meets a 0 at the i^{th} position, he knows that the next bits are the Elias code of the next symbol in $\widetilde{\boldsymbol{m}}$, and deduces m_i via (7). Since the Elias code is prefix, the decoder knows when he receives C1 again. Then the $(i+1)^{\text{th}}$ symbol can be processed.

Fig. 1 shows an illustration of the transmission process. In this example, the initial message is $x_{1:4} = 5, 3, 2, 7$. Then the message encoded in C1 is $\widetilde{x}_{1:4}0 = 0, 3, 2, 0, 0$. 13 bits are needed to transmit the second 0, and 15 bits for the last one. In C2 we transmit $\widetilde{\boldsymbol{m}} = 6, 3, 1$.

In the previous example, exact calculations have been performed, but this is not sensible for a practical implementation of arithmetic coding. Some rule is needed to set the precision

$$\underbrace{01110}_{\substack{\downarrow \\ \text{Elias code} \\ \text{of } \widetilde{m}_1+1}}\quad \underbrace{0110011001101}_{\substack{\text{beginning} \\ \text{of } \texttt{C1}}}\quad \underbrace{0101}_{\substack{\downarrow \\ \text{Elias code} \\ \text{of } \widetilde{m}_2-\widetilde{m}_1+1}}\quad \underbrace{00}_{\substack{\text{last bits} \\ \text{of } \texttt{C1}}}\quad \underbrace{1}_{\substack{\downarrow \\ \text{Elias code} \\ \text{of } 1}}$$

Fig. 1. Example of `ACcode`

in calculus, and it must be used by both coder and decoder. To avoid a too big extra redundancy caused by approximations, precision can grow as n grows. For instance, calculations can be made in memory with a further precision of $2\lceil \log i\rceil$ bits, in addition to the $\lceil -\log Q^i(\widetilde{x}_{1:i})\rceil + 1$ bits needed to encode $x_{1:i}$; this insures that the extra redundancy is bounded.

APPENDIX A
METRIC ENTROPY OF EXPONENTIALLY DECREASING ENVELOPE CLASSES

We give here the outlines of the lemmas we stated in 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 $\widetilde{y} = (y_n \mathbb{1}_{n\leq N_\epsilon})_{n\geq 1}$. One can check that

$$\|y - \widetilde{y}\| \leq \frac{\epsilon}{4}.$$

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. This leads to

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

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.

The rest of Lemma 2 follows from the Feller bounds, in their version proposed by [40, ch. XII]:

$$\sqrt{2\pi}\, x^{x-1/2} e^{-x} \leq \Gamma(x) \leq \sqrt{2\pi}\, x^{x-1/2} e^{-x} e^{\frac{1}{12x}}$$

■

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

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

generated by the coordinates from $l_f + 1$ to $l_f + m$. This leads to

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

It only remains to apply the logarithm function. ■

APPENDIX B
REDUNDANCY OF `ACcode`

A. *Moments of M_n*

First, we present several useful results about the moments of M_n.

Lemma 4: Let C and α be positive numbers satisfying $C > e^{2\alpha}$. Then, for all $n \geq 1$,

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

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

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

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

G is the distribution function of a 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

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

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

Since the upper bound does not depend on P, that achieves the proof of point 1. We can handle 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 4.

The third item is similar. 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

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

Taking the supremum over P, we get

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

■

B. Contribution of C1

Proposition 4: Let C and α be positive numbers satisfying $C > e^{2\alpha}$. Then

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

Proof: We give here the sketch of the proof, and we delay the proofs of (9), (10), (11), and (12).

Here we deal with the quantity

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

which corresponds to the contribution of C1. As we saw in Section III, the coding probability Q^n is based on Krichevsky-Trofimov mixtures. For $k \geq 1$, let KT_k denote the usual Krichevsky-Trofimov mixture on the alphabet $\{1, \ldots, k\}$, whose conditional probabilities are, for all $0 \leq i \leq n-1$ and for all $1 \leq j \leq k$,

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

Let us choose $k = m_n + 1$. In this case, there is a simple relation between KT_{m_n+1} and Q^n. For any sequence of n positive integers $x_{1:n} \in \mathbb{N}_*^n$,

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

As a consequence, we can link the redundancy of Q^n to the redundancy of KT_{m_n+1}:

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

and therefore

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

Note that (A_2) corresponds to the gain in redundancy of Q^n with respect to KT_{M_n+1}. 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

$$(\mathrm{A}_1) \leq \frac{\mathbb{E}[M_n]}{2}\log n + \mathbb{E}[\log(M_n+1)]. \tag{9}$$

Since $\mathbb{E}[\log M_n] \leq \mathbb{E}[M_n]$, Lemma 4 entails

$$\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \mathbb{E}[\log(M_n+1)] = o(\log^2 n).$$

Applying Lemma 4 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 non-decreasing 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

$$(\mathrm{A}_2) \geq \lambda_n \mathbb{E}_{P^n}\left[\sum_{k=h_n+1}^{m}\frac{M_n - M_{kM_n}}{2(k+1)}\right]. \tag{10}$$

It is easy to check that the function $x\mapsto x\log\left(1+\frac{1}{x}\right)$ is non-decreasing, and tends to $\log e$ when x tends to the infinity; therefore (λ_n) tends to $\log e$. We can write now

$$\begin{aligned}
(\mathrm{A}) &\leq \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}}\Bigg(\frac{\mathbb{E}[M_n]}{2}\log n \\
&\qquad - \lambda_n\mathbb{E}\left[\sum_{k=h_n+1}^{m}\frac{M_n-M_{kM_n}}{2(k+1)}\right]\Bigg) + o(\log^2 n)\\
&\leq \frac{1}{2}\overbrace{\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}}\mathbb{E}_{P^n}\left[M_n\log n - \lambda_n M_n\sum_{k=h_n+1}^{m}\frac{1}{k+1}\right]}^{(\mathrm{A}_3)}\\
&\quad + \frac{\lambda_n}{2}\overbrace{\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}}\mathbb{E}_{P^n}\left[\sum_{k=h_n+1}^{m}\frac{M_{kM_n}}{k+1}\right]}^{(\mathrm{A}_4)} + o(\log^2 n)
\end{aligned}$$

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

$$(\mathrm{A}_3) = o(\log^2 n) \tag{11}$$

$$(\mathrm{A}_4) \leq \frac{\log^2 n}{2\alpha\log^2 e} + o(\log^2 n). \tag{12}$$

Therefore we have

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

which concludes the proof of Proposition 4. ∎

Proof of (9): Let

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

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

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

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

Lemma 5: For all $k\geq 1$ and for all $x_{1:n}\in\{1,\ldots,k\}^n$,

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

∎

Proof of (10): We group the terms in (A_2), M_n by M_n:

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

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

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

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

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

∎

Proof of (11): We have

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

Then we plug in $h_n = \lfloor \ln n - 2\rfloor$. For n large enough, $h_n\geq 1$, and we have

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

and therefore

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

Then, if we use Lemma 4 and the fact that λ_n tends to $\log e$, we get (11). ∎

Proof of (12): We want to commute the expected value and the sum in (A_4). To do it, we need to get rid of 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,

$$\begin{aligned}
(A_4) &\leq \sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} \mathbb{E}_{P^n}\left[\sum_{k=3}^{m} \frac{M_{kM_n}}{k+1}\right] \\
&\leq \sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} \mathbb{E}_{P^n}\left[\sum_{k=3}^{n-1} \frac{M_{kM_n}\mathbb{1}_{kM_n \leq n-1}}{k+1}\right] \\
&\leq \sum_{k=3}^{n-1} \frac{\sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} \mathbb{E}_{P^n}[M_{kM_n}\mathbb{1}_{M_n \leq l_n}]}{k+1} \\
&\quad + \sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} \mathbb{E}_{P^n}[M_n \mathbb{1}_{M_n > l_n}] \sum_{k=3}^{n-1} \frac{1}{k+1},
\end{aligned}$$

where $l_n = \left\lfloor \frac{1}{\alpha}\left(2\ln n + \ln \frac{C}{1-e^{-\alpha}}\right)\right\rfloor$. We can now plug in the results of Lemma 4:

$$\begin{aligned}
(A_4) &\leq \sum_{k=3}^{n-1} \frac{\ln(kl_n) + 1 + \ln\frac{C}{1-e^{-\alpha}}}{(k+1)\alpha} + o(1)\sum_{k=3}^{n-1}\frac{1}{k+1} \\
&\leq \frac{1}{\alpha}\sum_{k=3}^{n-1}\frac{\ln k}{k+1} + \frac{1}{\alpha}(\ln l_n + O(1))\sum_{k=3}^{n-1}\frac{1}{k+1}.
\end{aligned}$$

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

$$\begin{aligned}
(A_4) &\leq \frac{1}{\alpha}\int_3^n \frac{\ln x}{x}\,\mathrm{d}x + O(\ln\ln n)\int_3^n \frac{\mathrm{d}x}{x} \\
&\leq \frac{1}{2\alpha}\ln^2 n + o(\ln^2 n).
\end{aligned}$$

∎

C. Contribution of C2

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

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

Proof: Like in the previous subsection, we give first the sketch of the proof, and we delay the proofs of (13) and (14).

$$\begin{aligned}
&\sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} \mathbb{E}_{P^n}[l(\mathtt{C2})] \\
&\qquad \leq 1 + \sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} \sum_{i=1}^{n} \mathbb{E}_{P^n}\left[\mathbb{1}_{X_i > M_{i-1}} l_E(X_i + 1)\right].
\end{aligned}$$

We deal with this sum thanks to the following. Let g be a positive and non-decreasing function on $[1, \infty)$. Let $(K_n)_{n \geq 1}$ be a non-decreasing sequence of positive integers. Then, for all $n \geq 1$,

$$\begin{aligned}
&\sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} \sum_{i=1}^{n} \mathbb{E}_{P^n}\left[\mathbb{1}_{X_i > M_{i-1}} g(X_i)\right] \\
&\leq (1 + o(1))\frac{g(K_n)\ln n}{\alpha} + C\,n\int_{K_n}^{\infty} g(x+1)e^{-\alpha x}\,\mathrm{d}x. \quad (13)
\end{aligned}$$

To apply (13), 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

$$\begin{aligned}
&\sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} \mathbb{E}_{P^n}[l(\mathtt{C2})] \\
&\leq (1 + o(1))\frac{l_E(K_n + 1)}{\alpha}\ln n + C\,n\int_{K_n}^{\infty} l_E(x+2)e^{-\alpha x}\,\mathrm{d}x
\end{aligned}$$

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

$$l_E(K_n + 1) \underset{n \to \infty}{\sim} \log K_n \sim \log\log n = o(\log n),$$

and therefore

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

With this choice of K_n we also get

$$n\int_{K_n}^{\infty} l_E(x+2)e^{-\alpha x}\,\mathrm{d}x = o(\log n) \quad (14)$$

which achieves the proof of Proposition 5. ∎

Proof of (13): 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

$$(B_1) = \sum_{i=1}^{n} \mathbb{E}_{P^n}\left[\mathbb{1}_{X_i > M_{i-1}} g(X_i)\right].$$

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

$$\begin{aligned}
(B_1) &= \sum_{i=1}^{n}\sum_{k \geq 0} P^n(M_{i-1} = k)\sum_{m \geq k+1} P(m)g(m) \\
&= \sum_{m \geq 1} P(m)g(m)\sum_{i=1}^{n}\sum_{k=0}^{m-1}\mathbb{P}(M_{i-1} = k) \\
&= P(1)g(1) + \sum_{m \geq 2} P(m)g(m)\sum_{i=1}^{n}(1 - \bar{p}(m-1))^{i-1} \\
&= \sum_{m \geq 1} P(m)g(m)\frac{1 - (1 - \bar{p}(m-1))^n}{\bar{p}(m-1)}.
\end{aligned}$$

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

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

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

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

■

Proof of (14):

$$\begin{aligned}&n\int_{K_n}^{\infty} l_E(x+2)e^{-\alpha x}\,\mathrm{d}x \\ &\quad\leq n\int_{K_n}^{\infty}\left(\log(x+2)+2\log\log(x+3)+1\right)e^{-\alpha x}\,\mathrm{d}x \\ &\quad\leq ne^{-\alpha K_n}\log(K_n+3) \\ &\qquad \int_{K_n+3}^{\infty}\frac{\log x+2\log\log x+1}{\log(K_n+3)}e^{-\alpha(x-K_n-3)}\,\mathrm{d}x \\ &\quad\leq e^{\alpha}\log(K_n+3)\left(\sup_{x\geq K_n+3}\frac{\log x+2\log\log x+1}{\log x}\right) \\ &\qquad \int_{K_n+3}^{\infty}\frac{\log x}{\log(K_n+3)}e^{-\alpha(x-K_n-3)}\,\mathrm{d}x \\ &\quad= O(\log K_n)\int_0^{\infty}\left(1+\frac{\log\left(1+\frac{x}{K_n+3}\right)}{\log(K_n+3)}\right)e^{-\alpha x}\,\mathrm{d}x \\ &\quad= o(\log n).\end{aligned}$$

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. ■

D. Proof of Theorem 4

The message sent by the `ACcode` algorithm is compound of two strings `C1` and `C2`. `C1` corresponds to the part of the message encoded by the arithmetic code, with coding probability Q^{n+1}. The arithmetic code encodes a message $\widetilde{x}_{1:n}0$ with $\lceil -\log Q^{n+1}(\widetilde{x}_{1:n}0)\rceil + 1$ bits. We have

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

thanks to Lemma 4. Therefore the redundancy of `ACcode` can be upper bounded, for all $n\geq 2$, by

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

We conclude thanks to Propositions 4 and 5.

Acknowledgment

I would like to thank E. Gassiat for her helpful advice, for the many ideas in this paper she suggested me, and for her constant availability to my questions.

Thanks also to A. Garivier and to S. Boucheron for the useful discussions we had. I don't forget the ideas I got from my classmates, especially R. Imekraz.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. 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.
[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. Plann. 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.
[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.
[31] A. Garivier, “A lower bound for the maximin redundancy in pattern coding,” *Entropy*, vol. 11, no. 4, pp. 634–642, 2009.
[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. ke He and E. hui 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*, vol. 55, no. 1, pp. 358–373, 2009.
[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.
[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-Sud 11 Univ., sep 2007. [Online]. Available: http://www.math.u-psud.fr/~bontemps/Documents/rapportM2.pdf
[40] 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.

Dominique Bontemps is currently (september 2010) PhD student at Laboratoire de Mathématiques d’Orsay, Univ Paris-Sud, F-91405 Orsay. His research fields are lossless compression, Bayesian statistics, and model selection.