Adaptive compression against a countable alphabet

Dominique Bontemps[1], Stéphane Boucheron[2†] and Elisabeth Gassiat[3‡]

[1] *Institut de Mathématiques de Toulouse, Toulouse, France*
[2] *LPMA ,CNRS Université Paris-Diderot*
[3] *LMO, CNRS, Université Paris-Sud*

This paper sheds light on universal coding with respect to classes of memoryless sources over a countable alphabet defined by an envelope function with finite and non-decreasing hazard rate. We prove that the auto-censuring (AC) code introduced by Bontemps (2011) is adaptive with respect to the collection of such classes. The analysis builds on the tight characterization of universal redundancy rate in terms of metric entropy by Haussler and Opper (1997) and on a careful analysis of the performance of the AC-coding algorithm. The latter relies on non-asymptotic bounds for maxima of samples from discrete distributions with finite and non-decreasing hazard rate.

Keywords: adaptive compression; countable alphabets; lossless data compression; minimax; redundancy; universal coding

1 Introduction

The aim of lossless data compression is to encode a sequence of symbols –the message– into a hopefully short sequence of bits –the codeword–. An encoding algorithm has not only to be one-to-one: the message should be recoverable from the codeword, it has also to be non-ambiguous, any sequence of messages should be recoverable from the corresponding sequence of codewords, this is warranted if no codeword is a strict prefix of any other codeword. Such a code is called a prefix code. Data compression is as old as digital communications and computers (See Cover and Thomas 1991, for references), and efficient lossless data compression algorithms can be found on any laptop either as stand-alone applications like `gzip,bzip2` or as embedded algorithms like Huffmann coders or arithmetic coders in `jpeg,jpeg2000` processing algorithms.

There are several ways of thinking about the cost of a data compression algorithm. The number of computational steps needed to encode/decode a message of n symbols may be the most obvious one for somebody familiar with algorithm analysis (Flajolet and Sedgewick 2009), quantifying the computational cost of the Burroughs-Wheeler transform (an essential ingredient in `bzip2`) is an example of such a

[†] supported by European Network PASCAL 2
[‡] supported by European Network PASCAL 2

question. In this text, we are interested in another aspect of the performance of data compression algorithms: redundancy. In words, if $\ell(X_{1:n})$ denotes the length of the codeword for message $X_{1:n}$ under some coding scheme, the redundancy of that coding scheme under some probability distribution $\mathbb{P}$ on the set of messages of length n over the finite or countable alphabet $\mathcal{X}$ is the difference between $\mathbb{E}[\ell(X_{1:n})]$ and the minimal expected length of codewords amongst all non-ambiguous codes.

Shannon first source coding theorem (Cover and Thomas 1991) asserts that this minimum is at least as large as the Shannon (binary) entropy of the probability distribution $\mathbb{P}$: $H(\mathbb{P}^n) = \mathbb{E}_{\mathbb{P}}[-\log \mathbb{P}(X_{1:n})]$ – throughout this paper, $\log x$ denotes the base 2 logarithm of x while $\ln x$ denotes its natural logarithm –. If $\mathbb{P}$ is known, it is possible to design an encoding algorithm that achieves average codelength at most $H(\mathbb{P}^n) + 1$. A construction by Shannon, Fano and Elias called arithmetic coding shows that it is always possible to choose codeword length not larger than $-\log \mathbb{P}\{X_{1:n}\} + 1$. This is why the quantity $-\log \mathbb{P}\{X_{1:n}\}$ is called the ideal codeword length of message $X_{1:n}$ under probability $\mathbb{P}$. The Kraft-McMillan inequality asserts that for any non-ambiguous encoding algorithm, the collection $(2^{-\ell(x_{1:n})})_{x_{1:n}}$ defines a sub-probability over messages of length n (Cover and Thomas 1991). Therefore, in the sequel, we will identify encoding algorithms and probability distributions.

If an encoding algorithm tailored to probability distribution $\mathbb{Q}$ (that is with coding length $\lfloor -\log \mathbb{Q}(x_{1:n}) \rfloor + 1$) is used against probability distribution $\mathbb{P}$, the average difference between the coding length and the Shannon entropy of $\mathbb{P}$ is called the redundancy of the encoding. Up to an additive constant 1, the redundancy of the coding probability $\mathbb{Q}$ with respect to $\mathbb{P}$ coincides with the Kullback-Leibler divergence between $\mathbb{P}$ and $\mathbb{Q}$: $D(\mathbb{P} \mid \mathbb{Q}) = \mathbb{E}_{\mathbb{P}}[\log \mathbb{P}(X_{1:n})/\mathbb{Q}(X_{1:n})]$. A universal coding algorithm attempts to minimize redundancy with respect to a collection of probability distributions.The richer the collection, the more challenging is the universal coding problem. The analysis of the redundancy of many useful coding algorithms like `gzip, bzip2, zip, ctw, etc` against very large or not so large collections of probability distribution has been carried out (See Szpankowski 2001, and references therein). Most results so far deal with coding over finite, known alphabets.

Here, we are interested in coding over a countable alphabet $\mathcal{X}$ (say the set of positive integers $\mathbb{N}_+$ or the set of integers $\mathbb{N}$). Sources over alphabet $\mathcal{X}$ are probability distributions on the set $\mathcal{X}^{\mathbb{N}}$ of infinite sequences of symbols from $\mathcal{X}$ and Λ will denote various collections of sources on alphabet $\mathcal{X}$. The symbols emitted by a source are denoted by a sequence $\mathbf{X}$ of $\mathcal{X}$-valued random variable $\mathbf{X} = (X_n)_{n\geq 1}$. If $\mathbb{P}$ is the distribution of $\mathbf{X}$, $\mathbb{P}^n$ denotes the distribution of the first n symbols $X_{1:n} = (X_1, ..., X_n)$, and we let $\Lambda^n = \{\mathbb{P}^n : \mathbb{P} \in \Lambda\}$. Finally, for any countable set $\mathcal{X}$, let $\mathfrak{M}_1(\mathcal{X})$ be the set of all probability distributions over $\mathcal{X}$.

Since $\mathbb{P}^n \in \Lambda^n$ is usually not known, universal coding looks like a statistical problem even though the problem of estimating the source is best avoided. Universal coding rather attempts to develop sequences of coding probabilities $(Q^n)_n$ so as to minimize the redundancy over a whole class of sources. The *maximal redundancy* of Q^n with respect to Λ is defined by:

$$R^+(Q^n, \Lambda^n) = \sup_{\mathbb{P}\in\Lambda} D(\mathbb{P}^n, Q^n)\,.$$

The infimum of $R^+(Q^n, \Lambda^n)$ is called the *minimax redundancy* with respect to Λ:

$$R^+(\Lambda^n) = \inf_{Q^n \in \mathfrak{M}_1(\mathcal{X}^n)} R^+(Q^n, \Lambda^n).$$

Classical results by Kieffer (1978), Györfi, Pali, and van der Meulen (1993, 1994) show that finite minimax redundancy is not a trivial property when the alphabet is infinite even for classes of memoryless sources. This observation contrasts with what we know about the finite alphabet setting where coding probabilities asymptotically achieving minimax redundancies have been described (Barron et al. 1998, Yang and Barron 1998, Clarke and Barron 1994).

Universal coding against a countable alphabet is not the same thing as universal coding against a finite alphabet of unknown size which is much larger than the message length. Note that delicate asymptotic results for coding over large finite alphabets with unknown size have started to appear (Szpankowski and Weinberger 2010).

Investigating minimax redundancy and developing efficient coding algorithms achieving that minimax redundancy over (relatively) small classes of sources is not enough. A more ambitious goal consists of developing adaptive coding algorithms, that is coding algorithms that asymptotically achieve minimax redundancy simultaneously over large collections of sources classes. Adaptivity is a pivotal concept in non-parametric statistics (See Catoni 2004, for a thorough account of the interplay between statistics and data compression). Formally, a sequence $(Q^n)_n$ of coding probabilities is said to be *asymptotically adaptive* with respect to a collection $(\Lambda_m)_{m\in\mathcal{M}}$ of source classes if for all $m \in \mathcal{M}$:

$$R^+(Q^n, \Lambda^n_m) = \sup_{\mathbb{P}\in\Lambda_m} D(\mathbb{P}^n, Q^n) \leq (1 + o(1))R^+(\Lambda^n_m)$$

as n tends to infinity. In the finite alphabet context, the context-tree weighting method (See Catoni 2004, and references therein) has for example been shown to be adaptive with respect to large collections of class of sources defined by Markovian constraints, while Lempel-Ziv coders like `gzip, zip, compress,...` are not (Louchard and Szpankowski 1997).

In this paper we consider the AC-coding algorithm introduced by Bontemps (2011).

The AC-code encodes a sequence $x_{1:n} = x_1, \ldots, x_n$ of symbols from $\mathbb{N}_+ = \mathbb{N} \setminus \{0\}$ on the basis of the following idea: large symbols are few, and can be encoded separately. In practice a symbol is great if it is greater than all symbols seen so far. Large symbols are encoded using Elias penultimate code (Elias 1975), which is a prefix code over $\mathbb{N}_+$ (the length of Eilas encoding of n is larger than $\log n + \log\log n$ and smaller than $\log n + 2\log\log n$). For small(er) symbols are handled by an arithmetic coder tailored to Krichevsky-Trofimov mixtures (See Cesa-Bianchi and Lugosi 2006, Catoni 2004, and references therein). The AC-encoding algorithm progressively enlarges the alphabet of the Krichevsky-Trofimov mixtures so as to accomodate the symbols met so far. The next paragraph describes the AC-encoding in technical terms.

For i: $1 \leq i \leq n$, let $m_i = \max_{1\leq j\leq i} x_j$. The i^{th} symbol is a *record* if $m_i \neq m_{i-1}$. Let n^0_i be the number of records up to index i. The j^{th} record is denoted by $\widetilde{m}_j$. From the definitions, $\widetilde{m}_{n^0_i} = m_i$ for all i. Let $\widetilde{m}_0 = 0$ and let $\widetilde{\mathbf{m}}$ be the sequence of differences between records terminated by a 1, $\widetilde{\mathbf{m}} = (\widetilde{m}_i - \widetilde{m}_{i-1} + 1)_{1\leq i\leq n^0_n} 1$ (the last 1 in the sequence serves as a terminating symbol). The symbols in $\widetilde{\mathbf{m}}$ are encoded using Elias penultimate code. This sequence of codewords forms C_E.
The sequence of censored symbols $\widetilde{x}_{1:n}$ is defined by $\widetilde{x}_i = x_i \mathbb{I}_{x_i \leq m_{i-1}}$. The binary string C_M is obtained by arithmetic encoding of $\widetilde{x}_{1:n}0$. The coding probability used to (arithmetically) encode $\widetilde{x}_{1:n}0$ is

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

with

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

where n_i^j is the number of occurrences of symbol j amongst the first i symbols (in $x_{1:i}$). We agree on $n_0^j = 0$ for all $j > 0$. Note that 0 is always encoded as a new symbol: if $x_{i+1} = j > m_i$, the AC-code encodes a 0 just as if it were is the first occurrence of symbol 0, meanwhile the counter n_i^j is incremented.

Bontemps (2011) describes a nice way of interleaving the Elias codewords C_E and the mixture code C_M in order to perform online encoding and decoding.

When turning to classes of memoryless sources over a countable alphabet, we consider the simplest possible ones, classes defined by an envelope function.

Definition 1 *Let f be a mapping from $\mathbb{N}_+$ to $[0,1]$, with $1 < \sum_{j>0} f(j) < \infty$. The* envelope class Λ_f *defined by the function f is the collection of stationary memoryless sources with first marginal distribution dominated by f: $\Lambda_f = \Big\{\mathbb{P} \ : \ \forall x \in \mathbb{N},\ \mathbb{P}^1\{x\} \leq f(x)$ and $\mathbb{P}$ is stationary and memoryless.$\Big\}$. The associated* envelope distribution *has lower endpoint $l_f = \max\{k\colon \sum_{j \geq k} f(j) \geq 1\}$. The envelope distribution F is defined by $F(k) = 0$ for $k < l_f$, and $F(k) = 1 - \sum_{j>k} f(j)$ for $k \geq l_f$. The tail function $\overline{F}$ is defined by $\overline{F} = 1 - F$. The associated probability mass function coincides with f for $u > l_f$ and is equal to $F(l_f) \leq f(l_f)$ at $u = l_f$.*

This envelope probability distribution plays a special role in the analysis of the minimax redundancy $R^+(\Lambda_f^n)$. Boucheron, Garivier, and Gassiat (2009) related the summability of the envelope function and the minimax redundancy of the envelope class. They proved almost matching upper and lower bounds on minimax redundancy for envelope classes as for example: $R^+(\Lambda_f^n) \leq \inf_{u \leq n} \big[n \log(1 + \overline{F}(u)) + \frac{u-1}{2} \log n\big] + 2$. The minimax redundancy of classes defined by exponentially vanishing envelopes was fully characterized by Bontemps (2011) using arguments borrowed from Haussler and Opper (1997), it avers that the minimax redundancy is half the upper bound obtained by choosing u so as $\overline{F}(u) \approx 1/n$ in the above-stated inequality. This suggested the possibility of describing the minimax redundancy as a simple functional of the envelope distribution without referring to the precise form of the envelope function. Bontemps proved that the AC-code is adaptive over the union of classes of sources with exponentially decreasing envelopes. As the AC-code does not benefit from side information concerning the envelope, it is natural to ask whether it is adaptive to a larger class of sources. That kind of question has been addressed in data compression over finite alphabets by Garivier (2006), who proved that Context-Tree-Weighting (Willems 1998, Catoni 2004) is adaptive over Renewal sources while it had been designed to compress sources with bounded memory. In a broader context, investigating the situations where an appealing procedure is minimax motivates the maxiset approach pioneered in (Cohen et al. 2001, Kerkyacharian and Picard 2002).

Haussler and Opper (1997) characterize the minimax redundancy of a collection of sources using the metric entropy of the class of marginal distributions, when the class is not too large. Intuition suggests that an envelope class is not too large when the envelope decreases fast enough. On the other hand, a bird's eye-view at the AC-code shows that it uses mixture coding over the observed alphabet in a sequential way. Intuition suggests that adaptivity depends on the fact that the observed alphabet does not grow too fast.

We prove that if the envelope distribution function has finite and non decreasing hazard rate (defined in Section 2): an explicit formula connects the minimax redundancy and the envelope distribution; the AC-code achieves the minimax redundancy, that is the AC-code is adaptive with respect to the collection of envelope classes with finite and non decreasing hazard rate.

The paper is organized as follows. Section 2 provides notation and definitions concerning hazard rates. The main result concerning the adaptivity of the AC-code over classes with envelopes with finite and non-decreasing hazard rate is stated in Section 3. The minimax redundancy of source classes defined by envelopes with finite and non-decreasing hazard rate is characterized in Section 4. Section 5 is dedicated to the characterization of the redundancy of the AC-code over source classes defined by envelopes with finite and non-decreasing hazard rate. Proofs are given in the Appendix.

2 Definitions and notation

Following Anderson (1970), it proves convenient to define a continuous distribution function F_c starting from the envelope distribution function F. The distribution function is characterized by its hazard function $h_c\colon [l_f-1,\infty)\to\mathbb{R}_+$, defined by $h_c(n)=-\ln\overline{F}(n)$ for $n\in\mathbb{N}$, and $h_c(t)=h_c(\lfloor t\rfloor)+(t-\lfloor t\rfloor)(h_c(\lfloor t\rfloor+1)-h_c(\lfloor t\rfloor))$ for $t\geq l_f-1$. The tail function of F_c is $\overline{F}_c(t)=\exp(-h_c(t))$ for $t\geq l_f-1$. For all integers n, $\overline{F}_c(n)=\overline{F}(n)$. The hazard rate h'_c is piecewise constant, it equals

$$\begin{aligned} h_c(\lfloor t\rfloor+1)-h_c(\lfloor t\rfloor) &= \ln(\overline{F}(\lfloor t\rfloor)/\overline{F}(\lfloor t\rfloor+1)) \\ &= \ln\left(1+f(\lfloor t\rfloor+1)/\overline{F}(\lfloor t\rfloor+1)\right) \end{aligned}$$

if $t\geq l_f$, and $-\ln\overline{F}(l_f)$ if $t\in(l_f-1,l_f)$. Notice that the hazard rate is finite on (l_f-1,∞) if and only if f has infinite support. Henceforth, given an envelope function f, $F,F_c,\overline{F},\overline{F}_c$ will be defined accordingly. We will also consistently define $U,U_c\colon(1,\infty)\to\mathbb{R}$ by

$$U(t)=\inf\{x\colon F(x)\geq 1-1/t\}=\inf\{x\colon 1/\overline{F}(x)\geq t\}$$

and $U_c(t)=\inf\{x\colon 1/\overline{F}_c(x)\geq t\}$. The last two functions prove illuminating in extreme value theory. If the hazard rate is finite, then $U(n)\to\infty$ and $U_c(n)\to\infty$ as n tends to infinity. Note that if F is the envelope distribution defined by f, then $F_c(t)=0$ for $t\leq l_f-1$. Recall that if X is distributed according to F_c then $\lfloor X\rfloor+1$ is distributed according to F or equivalently that $U(t)=\lfloor U_c(t)\rfloor+1$ for $t>1$.

In this paper, abusing notation, an envelope function f is said to have finite and non-decreasing hazard rate if the associated continuous distribution function F_c has finite and non-decreasing hazard rate. In this case, the essential infimum of the hazard rate is $b=-\ln\overline{F}(l_f)>0$. The envelopes introduced in the next definition provide examples of such envelopes. Poisson distributions offer other examples.

Definition 2 *The* sub-exponential envelope class *with parameters* $\alpha\geq 1$ *(shape),* $\beta>0$ *(scale) and* $\gamma>1$ *is the set* $\Lambda(\alpha,\beta,\gamma)$ *of probability mass functions* $(p(k))_{k\geq 1}$ *on the positive integers such that*

$$\forall k\geq 1,\ p(k)\leq f(k),\ \text{ where } f(k)=\gamma e^{-\left(\frac{k}{\beta}\right)^{\alpha}}.$$

Exponentially vanishing envelopes (Boucheron et al. 2009) are obtained by fixing $\alpha=1$.

To define source classes small enough so that the metric entropy of the class of marginal distributions characterizes the minimax redundancy, Haussler and Opper (1997) introduce functions with properties we set below. We shall see that source classes defined by envelopes with finite and non-decreasing hazard rate are small enough in that respect. Recall that a measurable function $h\colon (0,\infty) \to [0,\infty)$ is said to be *slowly varying* at infinity if for all $\kappa > 0$, $\lim_{x\to+\infty} \frac{h(\kappa x)}{h(x)} = 1$ (See Bingham et al. 1989).

Definition 3 *A continuous, non decreasing function $h\colon (0,\infty) \to [0,\infty)$ is said to be* very slowly varying *at infinity if for all $\eta \geq 0$ and $\kappa > 0$,*

$$\lim_{x\to+\infty} \frac{h(\kappa x (h(x))^\eta)}{h(x)} = 1 \quad \textit{and} \quad \lim_{x\to+\infty} \frac{h(\kappa x (\ln x)^\eta)}{h(x)} = 1.$$

The next proposition, proved in the appendice (Section A), allows to check that source classes defined by envelopes with finite and non-decreasing hazard rate are indeed small.

Proposition 1 *Let f be an envelope function with finite and non-decreasing hazard rate. Then*

(i) *U_c is slowly varying at infinity;*

(ii) *$U_c \circ \exp$ is a concave function, and its derivative is equal to $\overline{F}_c(U_c(\exp(t)))/f_c(U_c(\exp(t)))$;*

(iii) *The function $\widetilde{h}\colon [1,\infty) \to \mathbb{R}$, $\widetilde{h}(t) = \int_1^{t^2} \frac{U_c(x)}{2x}\mathrm{d}x$ is very slowly varying;*

(iv)

$$\lim_{t\to+\infty} \frac{U_c(t)\ln U_c(t)}{\int_1^t \frac{U_c(x)}{x}\mathrm{d}x} = 0.$$

3 Main result

Theorem 1 *The AC-code is adaptive with respect to source classes defined by envelopes with finite and non-decreasing hazard rate.*

Let Q^n be the coding probability associated with the AC-code, then if f is an envelope with non-decreasing hazard rate,

$$R^+(Q^n;\Lambda_f^n) \leq (1+o(1))(\log e) \int_1^n \frac{U_c(x)}{2x}\mathrm{d}x$$

while

$$R^+(\Lambda_f^n) = (1+o(1))(\log e) \int_1^n \frac{U_c(x)}{2x}\mathrm{d}x$$

as n tends to infinity.

The following corollary provides the bridge with Bontemps's work.

Corollary 1 *The AC-code is adaptive with respect to sub-exponential envelope classes: $\cup_{\alpha\geq 1,\beta>0,\gamma>1}\Lambda(\alpha,\beta,\gamma)$. Let Q^n be the coding probability associated with the AC-code, then*

$$R^+(Q^n;\Lambda^n(\alpha,\beta,\gamma)) \leq (1+o(1))R^+(\Lambda^n(\alpha,\beta,\gamma))$$

as n tends to infinity.

Bontemps (2011) showed that the AC-code is adaptive over exponentially decreasing envelopes, that is over $\cup_{\beta>0,\gamma>1}\Lambda(1,\beta,\gamma)$. Corollary 1 shows that the AC-code is adaptive to both the scale and the shape parameter.

The next equation helps in understanding the relation between the redundancy of the AC-code and the metric entropy:

$$\int_1^t \frac{U_c(x)}{2x}\mathrm{d}x = \int_0^{U_c(t)} \frac{\ln(t\overline{F}_c(x))}{2}\mathrm{d}x\,. \tag{1}$$

The elementary proof is given at the end of the appendix. The left-hand-side of the equation appears (almost) naturally in the derivation of the redundancy of the AC-code. The right-hand-side or rather an equivalent of it, appears during the computation of the minimax redundancy of the envelope classes considered in this paper.

The proof of Theorem 1 is organized in two parts : Proposition 3 from Section 4 describes the minimax redundancy of source classes defined by envelopes with finite and non-decreasing hazard rate.

The redundancy of the AC-coding probability Q^n with respect to $\mathbb{P}^n \in \Lambda^n(f)$ is analyzed in Section 5. The pointwise redundancy is decomposed in the following way:

$$-\log Q^n(X_{1:n}) + \log \mathbb{P}^n(X_{1:n}) = \underbrace{\ell(C_E)}_{\mathrm{I}} + \underbrace{\ell(C_M) + \log \mathbb{P}^n(X_{1:n})}_{\mathrm{II}}\,.$$

Proposition 8 asserts that (I) is negligible with respect to $R^+(\Lambda_f^n)$, uniformly for $\mathbb{P} \in \Lambda_f$, and Proposition 9 asserts that the expected value of (II) is bounded, uniformly for $\mathbb{P} \in \Lambda_f$, by a term which is equivalent to $R^+(\Lambda_f^n)$.

4 Minimax redundancies

The minimax redundancy of source classes defined by envelopes f with finite and non-decreasing hazard rate is characterized using Theorem 5 from (Haussler and Opper 1997). This theorem relates the minimax redundancy to the metric entropy of the class of marginal distributions with respect to Hellinger distance. Recall that the Hellinger distance between two probability distributions P_1 and P_2 on $\mathbb{N}$, defined by the corresponding probability mass functions p_1 and p_2, is $\left(\sum_{k\in\mathbb{N}}\left(\sqrt{p_1(k)} - \sqrt{p_2(k)}\right)^2\right)^{1/2}$. If probability distributions over $\mathbb{N}$ are parametrized by the square root of their probability mass function, the Hellinger metric is just the ℓ_2 distance between parameters. For a source class Λ, Let $\mathcal{H}_\epsilon(\Lambda)$ be the ϵ-entropy of Λ^1 with respect to the Hellinger metric. That is, $\mathcal{H}_\epsilon(\Lambda) = \ln \mathcal{D}_\epsilon(\Lambda)$ where $\mathcal{D}_\epsilon(\Lambda)$ is the cardinality of the smallest finite partition of Λ^1 into sets of diameter at most ϵ when such a finite partition exists.

Theorem 2 *(Haussler and Opper 1997, Theorem 5) Let Λ be a class of stationary memoryless sources. Assume there exists a very slowly varying function h such that:*

$$\mathcal{H}_\epsilon(\Lambda) = h\left(\tfrac{1}{\epsilon}\right)(1+o(1)) \quad \textit{as } \epsilon \textit{ tends to } 0.$$

Then

$$R^+(\Lambda^n) = (\log e)h(\sqrt{n})\,(1+o(1)) \quad \textit{as } n \textit{ tends to } +\infty.$$

This theorem tightly characterizes the asymptotic redundancy of small source classes. Notice that the definition of redundancy uses base 2 logarithms while ϵ-entropy is usually defined using natural logarithms.

Proposition 2 *(Entropy of envelope classes with finite and non-decreasing hazard rate.) Let f be an envelope function with finite and non-decreasing hazard rate, then*

$$\mathcal{H}_\epsilon(\Lambda_f) = (1+o(1)) \int_0^{1/\epsilon^2} \frac{U_c(x)}{2x}\mathrm{d}x \quad \text{as } \epsilon \text{ tends to } 0 \ .$$

The proof follows the approach of (Bontemps 2011). It is stated in the appendix.

The characterization of $R^+(\Lambda_f^n)$ follows from a direct application of Theorem 2 and Proposition 1 (iii):

Proposition 3 *(Minimax redundancy of envelope classes with finite and non-decreasing hazard rate.) Let f be an envelope function with finite and non-decreasing hazard rate, then*

$$R^+(\Lambda_f^n) = (1+o(1))(\log e) \int_1^{n} \frac{U_c(x)}{2x}\mathrm{d}x \quad \text{as } n \text{ tends to } +\infty \ .$$

A concrete corollary follows easily.

Proposition 4 *The minimax redundancy of the sub-exponential envelope class with parameters (α, β, γ) satisfies*

$$R^+(\Lambda^n(\alpha,\beta,\gamma)) = \frac{\alpha}{2(\alpha+1)}\beta \left(\ln(2)\right)^{1/\alpha} \left(\log n\right)^{1+1/\alpha} (1+o(1)) \quad \text{as } n \text{ tends to } +\infty \ .$$

Proof: Indeed, if f is a sub-exponential envelope function with parameters (α, β, γ) one has, for $t > 1$,

$$\beta \left(\ln(\gamma t)\right)^{1/\alpha} - 1 \le U_c(t) \le \beta \left(\ln(\kappa\gamma t)\right)^{1/\alpha} - 1$$

where $\kappa = 1/(1-\exp(-\alpha/\beta^\alpha))$.

The lower bound follows from $\overline{F}(k) \ge f(k+1) = \gamma \exp(-((k+1)/\beta)^\alpha)$ which entails $\overline{F}(k) \le 1/t \Rightarrow k+1 \ge \beta(\ln(\gamma t))^{1/\alpha}$.

The upper bound follows from

$$\overline{F}(k) \le \sum_{j\ge 0} \gamma \exp\left(-\left(\tfrac{k+1}{\beta}\right)^\alpha - j\alpha \tfrac{(k+1)^{\alpha-1}}{\beta^\alpha}\right) \le \frac{f(k+1)}{1-\exp\left(-\alpha(k+1)^{\alpha-1}/\beta^\alpha\right)} \le \frac{f(k+1)}{\kappa}$$

for $\alpha \ge 1$. □

5 Redundancy of the AC-encoding algorithm

The length of the AC-encoding of $x_{1:n}$, is the sum of the length of the Elias encoding C_E of the sequence of differences between records $\widetilde{\mathbf{m}}$ and of the length of the mixture encoding C_M of the censored sequence $\widetilde{x}_{1:n}0$. In order to establish Theorem 1, we first establish an upper bound on the average length of C_E (Proposition 8).

5.1 Maximal inequalities

Bounds on the codeword length of Elias encoding and on the redundancy of the mixture code essentially rely on bounds on the expectation of the largest symbol $\max(X_1, \ldots, X_n)$ collected in the next propositions. In the sequel, H_n denotes the n^{th} harmonic number $\ln(n) \leq H_n = \sum_{i=1}^n \frac{1}{i} \leq \ln(n) + 1$.

Proposition 5 *Let $Y_1, \ldots, Y_n$ be independently identically distributed according to an absolutely continuous distribution function F with density $f = F'$ with support included in $[1, \infty)$ and non-decreasing hazard rate $f/\overline{F}$. Let b be the infimum of the hazard rate. Let $U(t) = \inf\{x \colon F(x) \geq 1 - 1/t\}$ and $Y_{1,n} \leq \ldots \leq Y_{n,n}$ be the order statistics. Then,*

$$\mathbb{E}[Y_{n,n}] \leq U(\exp(H_n))$$
$$\mathbb{E}[Y_{n,n} \ln(Y_{n,n})] \leq (\mathbb{E}Y_{n,n}) \ln(\mathbb{E}Y_{n,n})) + 2/b^2 \, .$$

The proof relies on a quantile coupling argument and on a sequence of computational steps inspired by extreme value theory (de Haan and Fereira 2006) and concentration of measure theory (Ledoux 2001). The proof also takes advantage of the Rényi representation of order statistics (See de Haan and Fereira 2006, Chapter 2). The next theorem rephrases this classical result.

Theorem 3 (RÉNYI'S REPRESENTATION) *Let $(X_{1,n}, \ldots, X_{n,n})$ denote the order statistics of an independent sample picked according to a distribution function F. Then $(X_{1,n}, \ldots, X_{n,n})$ is distributed as $(U(\exp(E_{1,n})), \ldots, U(\exp(E_{n,n})))$ where $U \colon (1, \infty) \to \mathbb{R}$ is defined by $U(t) = \inf\{x \colon F(x) \geq 1 - 1/t\}$ and $(E_{1,n}, \ldots, E_{n,n})$ are the order statistics of an n-sample of the exponential distribution with scale parameter 1. Agreeing on $E_{0,n} = 0$, $(E_{i,n} - E_{i-1,n})_{1 \leq i \leq n}$ are independent and exponentially distributed with scale parameter $1/(n + 1 - i)$.*

We will also use the following general relations on moments of maxima of independent random variables, proved below.

Proposition 6 *Let $(Y_{1,n}, \ldots, Y_{n,n})$ denote the order statistics of an independent sample picked according to a common probability distribution with support included in $(0, \infty)$, then*

$$\mathbb{E}\left[Y_{n,n} \ln Y_{n,n}\right] \leq \mathbb{E}Y_{n,n} \ln(\mathbb{E}Y_{n,n}) + \mathbb{E}\left[\frac{(Y_{n,n} - Y_{n-1,n})^2}{Y_{n-1,n}}\right] .$$

The proof of the next theorem, can be found in (Ledoux 2001) .

Theorem 4 (SUB-ADDITIVITY OF ENTROPY.) *Let $X_1, \ldots, X_n$ be independent random variables and let $Z = f(X)$ be a non-negative function of $X = (X_1, \ldots, X_n)$. For each $1 \leq i \leq n$, let Z_i be non-negative function of $(X_1, \ldots, X_{i-1}, X_{i+1}, X_n)$. Then*

$$\mathbb{E}\left[Z \ln(Z)\right] - \mathbb{E}Z \ln(\mathbb{E}Z) \leq \sum_{i=1}^n \mathbb{E}\left[Z \ln\left(\frac{Z}{Z_i}\right) - (Z - Z_i)\right] .$$

Proof of Proposition 6: Note that $Z = Y_{n,n}$ is a function of the n independent random variables $Y_1, \ldots, Y_n$. Choose the Z_i as the maximum of $Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n$, that is $Y_{n,n}$ if $Y_i < Y_{n,n}$ and

$Y_{n-1,n}$ otherwise. We have $Z_i = Z$ except possibly when $X_i = Z$, and $Z_i = Y_{n-1,n}$ otherwise. Using the sub-additivity of entropy,

$$\begin{aligned}\mathbb{E}\left[Y_{n,n}\ln Y_{n,n}\right] - \mathbb{E}Y_{n,n}\ln(\mathbb{E}Y_{n,n}) &\le \mathbb{E}\left[Y_{n,n}\ln\frac{Y_{n,n}}{Y_{n-1,n}} - (Y_{n,n}-Y_{n-1,n})\right] \\ &\le \mathbb{E}\left[Y_{n,n}\ln\left(1+\frac{Y_{n,n}-Y_{n-1,n}}{Y_{n-1,n}}\right) - (Y_{n,n}-Y_{n-1,n})\right] \\ &\le \mathbb{E}\left[\frac{(Y_{n,n}-Y_{n-1,n})^2}{Y_{n-1,n}}\right] \text{ as } \ln(1+u)\le u \text{ for } u>-1.\end{aligned}$$

□

Proof of Proposition 5: Thanks to Rényi's representation of order statistics, $\mathbb{E}[Y_{n,n}] = \mathbb{E}[U(\exp(E_{n,n}))]$. The first statement follows from the concavity of $t \mapsto U(\exp(t))$ (Proposition 1, ii).

By Proposition 6,

$$\mathbb{E}[Y_{n,n}\ln(Y_{n,n})] \le (\mathbb{E}Y_{n,n})\ln(\mathbb{E}Y_{n,n})) + \mathbb{E}\left[\frac{(Y_{n,n}-Y_{n-1,n})^2}{Y_{n-1,n}}\right].$$

Thanks to Rényi's representation, $Y_{n,n} - Y_{n-1,n}$ is distributed like $U(\exp(E_{n,n})) - U(\exp(E_{n-1,n}))$. Thanks to the concavity of $U \circ \exp$, this difference is upper-bounded by

$$U(\exp(E_{n,n})) - U(\exp(E_{n-1,n})) \le \frac{\overline{F}(U(\exp(E_{n-1,n})))}{f(U(\exp(E_{n-1,n})))}(E_{n,n}-E_{n-1,n})$$

and the two factors on the right-hand-side are independent. Meanwhile $\mathbb{E}[(E_{n,n} - E_{n-1,n})^2] = 2$, $\frac{\overline{F}(U(\exp(E_{n-1,n})))}{f(U(\exp(E_{n-1,n})))} \le \frac{1}{b}$.

□

When handling envelopes classes with finite and non decreasing hazard rate, Proposition 5 provides a handy way to upper bound the various statistics that are used to characterize the redundancy of the AC-code. If the source belongs to Λ_f, let $Y_1, \dots, Y_n$ be identically independently distributed according to the probability distribution with tail function $\overline{F}_c$. The quantile coupling argument ensures that there exists a probability space with random variables $(X_1', \dots, X_n')$ distributed like $(X_1, \dots, X_n)$ and random variables $(Y_1', \dots, Y_n')$ distributed like $(Y_1, \dots, Y_n)$ and $X_i' \le Y_i' + 1$ for all $i \le n$ almost surely.

Let $Y_{(1)} \le \dots \le Y_{(n)}$ denote the order statistics of $Y_1, \dots, Y_n$, and let M_n denote the maximum of $X_1, \dots, X_n$. Then for any non-decreasing function g, $\mathbb{E}[g(M_n)] \le \mathbb{E}[g(Y_{(n)} + 1)]$. Using Proposition 5 one gets the following.

Proposition 7 *Let $X_1, \dots, X_n$ be independently identically distributed according to $P \in \Lambda_f^1$, let $M_n = \max(X_1, \dots, X_n)$, then,*

$$\begin{aligned}\mathbb{E}M_n &\le U_c(en) + 1 \\ \mathbb{E}[M_n \log M_n] &\le [U_c(en)+1]\ln[U_c(en)+1] + 2/b^2.\end{aligned}$$

5.2 Elias encoding

The average length of the Elias encoding for sources from a class defined by an envelope with non-decreasing hazard rate is $O(U_c(n))$. It does not grow as fast as the minimax redundancy and contributes in a negligible way to the total redundancy. Indeed, $U_c(en) = o\big(\int_1^n \frac{U_c(x)}{2x}\mathrm{d}x\big)$ thanks to Proposition 1.

Proposition 8 *Let f be an envelope function with associated non-decreasing hazard rate. Then, for all $\mathbb{P} \in \Lambda_f$, the expected length of the Elias encoding of the sequence of record increments amongst the first n symbols is upper-bounded by*

$$\mathbb{E}\left[\ell(C_E)\right] \leq (2\log(e) + \rho)(U_c(\exp(H_n)) + 1)$$

where ρ is a universal constant (which may be chosen as $\rho = 2$).

Proof of Proposition 8: The length of the Elias codewords used to encode the sequence of record differences $\widetilde{\mathbf{m}}$ is readily upper-bounded:

$$\ell(C_E) \leq \sum_{i=1}^{n_n^0} (2\log(1 + \widetilde{m}_i - \widetilde{m}_{i-1}) + \rho) \leq \sum_{i=1}^{n_n^0} 2\log(e)\,(\widetilde{m}_i - \widetilde{m}_{i-1}) + \rho n_n^0 \leq (2\log(e) + \rho)M_n$$

for some universal constant ρ. The bound on the length of the Elias encoding follows from Proposition 7. □

5.3 Adaptive mixture coding

The next proposition compares the length of the mixture encoding C_M with the ideal codeword length of $X_{1:n}$.

Proposition 9 *Let $f\colon \mathbb{N}_+ \to [0,1]$ be an envelope with finite and non-decreasing hazard rate. The expected difference between the mixture encoding of the censored sequence $\widetilde{X}_{1:n}$ and the ideal codeword length of $X_{1:n}$ is upper-bounded as*

$$\mathbb{E}\left[\ell(C_M) + \log \mathbb{P}(X_{1:n})\right] \leq \log(e)\int_1^n \frac{U_c(x)}{2x}\mathrm{d}x\,(1 + o(1))$$

as n tends to infinity.

The proof of Proposition 9 is organized in two steps. The first step consists in establishing a pointwise upper bound on the difference between the ideal codeword length and codeword length of the AC-code (Proposition 10 below). This upper-bound consists of three summands. The expected value of the three summands is then upper-bounded under the assumption that the source belongs to an envelope class with non-decreasing hazard rate.

Proposition 10 (POINTWISE BOUND) *Let i_0 be the random integer defined by: $i_0 = 1 \vee \lfloor M_n/4 \rfloor$, then,*

$$-\ln Q^n(\widetilde{X}_{1:n}) + \ln \mathbb{P}^n(X_{1:n}) \leq \underbrace{\frac{M_n(\ln(M_n) + 10)}{2} + \frac{\ln n}{2}}_{\text{(A.I)}} + \underbrace{\sum_{i=i_0}^{n-1}\left(\frac{M_i}{2i+1}\right)}_{\text{(A.II)}}$$

Proof: The pointwise redundancy can be decomposed into

$$-\ln Q^n(\widetilde{X}_{1:n}) + \ln \mathbb{P}^n(X_{1:n}) = \underbrace{-\ln \mathrm{KT}_{M_n+1}(\widetilde{X}_{1:n}) + \ln \mathbb{P}^n(X_{1:n})}_{(\mathrm{A})} \underbrace{-\ln Q^n(\widetilde{X}_{1:n}) + \ln \mathrm{KT}_{M_n+1}(\widetilde{X}_{1:n})}_{(\mathrm{B})}$$

where KT_{M_n+1} is the Krichevsky-Trofimov mixture coding probability over an alphabet of cardinality M_n+1. Summand (A) may be upper bounded thanks to the next bound, the proof of which can be found in (Boucheron, Garivier, and Gassiat 2009),

$$(\mathrm{A}) = -\ln(\mathrm{KT}_{M_n+1}(\widetilde{X}_{1:n})) + \ln(\mathbb{P}^n(X_{1:n})) \leq \frac{M_n+1}{2}\ln(n) + 2\ln(2)\,.$$

The second summand (B) is negative, this is the codelength the AC-code pockets by progressively enlarging the alphabet rather than using $\{0,\dots,M_n\}$ as the alphabet. Bontemps (2011, in the proof of Proposition 4) points out a simple and useful connexion between the coding lengths under Q^n and KT_{M_n+1}:

$$(\mathrm{B}) = -\ln Q^n(\widetilde{X}_{1:n}) + \ln \mathrm{KT}_{M_n+1}(\widetilde{X}_{1:n}) = -\sum_{i=1}^{n-1} \ln\left(\frac{2i+1+M_n}{2i+1+M_i}\right)\,.$$

The difference between the codelengths can be further upper bounded.

$$\begin{aligned}
-\sum_{i=1}^{n-1} \ln\left(\frac{2i+1+M_n}{2i+1+M_i}\right) &= -\sum_{i=1}^{n-1} \ln\left(1+\frac{M_n-M_i}{2i+1+M_i}\right) \\
&\leq -\sum_{i=i_0}^{n-1}\left(\frac{M_n-M_i}{2i+1+M_i}\right) + \frac{1}{2}\sum_{i=i_0}^{n-1}\left(\frac{M_n-M_i}{2i+1+M_i}\right)^2 \\
&\quad \text{as } \ln(1+x) \geq x - x^2/2 \text{ for } x \geq 0 \\
&= \underbrace{\sum_{i=i_0}^{n-1}\left(\frac{-M_n}{2i+1+M_i}\right)}_{(\mathrm{B.I})} + \underbrace{\sum_{i=i_0}^{n-1}\left(\frac{M_i}{2i+1+M_i}\right)}_{(\mathrm{B.II})} + \underbrace{\frac{1}{2}\sum_{i=i_0}^{n-1}\left(\frac{M_n-M_i}{2i+1+M_i}\right)^2}_{(\mathrm{B.III})}\,.
\end{aligned}$$

The upper bound on (A) can be used to build an upper bound on (A)+(B.I).

$$\begin{aligned}
(\mathrm{A}) + (\mathrm{B.I}) &\leq M_n\left(\frac{\ln(n)}{2} - \sum_{i=i_0}^{n-1}\frac{1}{2i+1+M_i}\right) + \frac{\ln n}{2} \\
&= M_n\left(\sum_{i=i_0}^{n-1}\left(\frac{1}{2i} - \frac{1}{2i+1+M_i}\right) + \frac{\ln(n)}{2} - \sum_{i=i_0}^{n-1}\frac{1}{2i}\right) + \frac{\ln n}{2} \\
&\leq M_n\left(\sum_{i=i_0}^{n-1}\frac{M_i+1}{(2i+1+M_i)(2i)} + \frac{H_{i_0}}{2} + \frac{1}{2n}\right) + \frac{\ln n}{2} \\
&\leq M_n\sum_{i=i_0}^{n-1}\frac{M_i+1}{(2i+1)(2i)} + \frac{M_n(\ln(M_n)+2)}{2} + \frac{\ln n}{2}\,.
\end{aligned}$$

Adding (B.III) to the first summand in the last expression,

$$
\begin{aligned}
M_n \sum_{i=i_0}^{n-1} \frac{M_i+1}{(2i+1)(2i)} + (\text{B.III}) &\le M_n \sum_{i=i_0}^{n-1} \frac{M_i}{(2i+1)^2(2i)} + M_n \sum_{i=i_0}^{n-1} \frac{1}{(2i+1)(2i)} + \frac{1}{2}\sum_{i=i_0}^{n-1} \frac{M_n^2+M_i^2}{(2i+1)^2} \\
&\le M_n^2 \sum_{i\ge i_0} \left(\frac{1}{2i(2i+1)^2} + \frac{1}{(2i+1)^2} \right) + \frac{M_n}{2i_0} \\
&\le M_n \left(\frac{M_n}{2i_0} + \frac{1}{2i_0} \right) \\
&\le 4M_n\,.
\end{aligned}
$$

□

Proof of Proposition 9: First recall that $\ell(C_M) = \lfloor -\log Q^{n+1}(\widetilde{X}_{1:n}0)\rfloor + 1 \le 1 + \log(2n + M_n + 1) - \log Q^n(\widetilde{X}_{1:n})$. Therefore, using Proposition 7 and Proposition 10, the average redundancy of the mixture code is upper bounded by

$$
3 + U_c(en) + \log(e) \left(\ln n + \mathbb{E}\Big[\underbrace{\frac{M_n(\ln(M_n)+10)}{2} + \frac{\ln n}{2}}_{(\text{A.I})} \Big] + \mathbb{E}\Big[\underbrace{\sum_{i=i_0}^{n-1} \left(\frac{M_i}{2i+1} \right)}_{(\text{A.II})} \Big] \right).
$$

We may now use the maximal inequalities from Proposition 7.

$$
\begin{aligned}
\sum_{i=1}^{n-1} \frac{\mathbb{E}M_i}{2i+1} &\le \sum_{i=1}^{n-1} \frac{U_c(\exp(H_i))+1}{2i+1} \\
&\le \sum_{i=1}^{n-1} \frac{U_c(ei)+1}{2i+1} \\
&\le \int_1^n \frac{U_c(ex)}{2x}\mathrm{d}x + \frac{U_c(e)}{3} + \frac{\ln(n)}{2}\,.
\end{aligned}
$$

Meanwhile, letting b be the infimum of the hazard rate of the envelope,

$$
\mathbb{E}\Big[\frac{M_n(\ln(M_n)+10)}{2} + \frac{\ln n}{2} \Big] \le \frac{(U_c(en)+1)(\ln(U_c(en)+1)+10)}{2} + \frac{2}{b^2} + \frac{\ln n}{2}.
$$

Now using Proposition 1 (i) and (iv) and the fact that U_c tends to infinity at infinity one gets that

$$
\ln n + U_c(n)\ln U_c(n) = o\left(\int_1^n \frac{U_c(ex)}{2x}\mathrm{d}x \right)
$$

as n tends to infinity and the result follows.

□

References

C. Anderson. Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *J. Appl. Probability*, 7:99–113, 1970.

A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44(6):2743–2760, 1998.

N. Bingham, C. Goldie, and J. Teugels. *Regular variation*, volume 27. Cambridge University Press, 1989.

D. Bontemps. Universal coding on infinite alphabets: exponentially decreasing envelopes. *IEEE Trans. Inform. Theory*, 57(3):1466–1478, 2011.

S. Boucheron, A. Garivier, and E. Gassiat. Coding over infinite alphabets. *IEEE Trans. Inform. Theory*, 55:358–373, 2009.

O. Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, 2004. Ecole d'Ete de Probabilites de Saint-Flour XXXI.

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

B. Clarke and A. Barron. Jeffrey's prior is asymptotically least favorable under entropy risk. *J. Stat. Planning and Inference*, 41:37–60, 1994.

A. Cohen, R. DeVore, G. Kerkyacharian, and D. Picard. Maximal spaces with given rate of convergence for thresholding algorithms. *Appl. Comput. Harmon. Anal.*, 11(2):167–191, 2001.

T. Cover and J. Thomas. *Elements of information theory*. John Wiley & sons, 1991.

L. de Haan and A. Fereira. *Extreme value theory*. Springer-Verlag, 2006.

P. Elias. Universal codeword sets and representations of the integers. *IEEE Trans. Information Theory*, IT-21:194–203, 1975.

P. Flajolet and R. Sedgewick. *Analytic combinatorics*. Cambridge University Press, Cambridge, 2009.

A. Garivier. Redundancy of the context-tree weighting method on renewal and Markov renewal processes. *IEEE Trans. Inform. Theory*, 52(12):5579–5586, 2006.

L. Györfi, I. Pali, and E. van der Meulen. On universal noiseless source coding for infinite source alphabets. *Eur. Trans. Telecommun. & Relat. Technol.*, 4(2):125–132, 1993.

L. Györfi, I. Páli, and E. C. van der Meulen. There is no universal source code for an infinite source alphabet. *IEEE Trans. Inform. Theory*, 40(1):267–271, 1994. ISSN 0018-9448.

D. Haussler and M. Opper. Mutual information, metric entropy and cumulative relative entropy risk. *Annals of Statistics*, 25(6):2451–2492, 1997. ISSN 0090-5364.

G. Kerkyacharian and D. Picard. Minimax or maxisets? *Bernoulli*, 8(2):219–253, 2002.

J. C. Kieffer. A unified approach to weak universal source coding. *IEEE Trans. Inform. Theory*, 24(6): 674–682, 1978.

M. Ledoux. *The concentration of measure phenomenon*. AMS, 2001.

G. Louchard and W. Szpankowski. On the average redundancy rate of the Lempel-Ziv code. *IEEE Trans. on Information Theory*, 43(1):2–8, 1997.

W. Szpankowski. *Average case analysis of algorithms on sequences.* J. Wiley, 2001.

W. Szpankowski and M. Weinberger. Minimax redundancy for large alphabets. In *Proceeding 2010 Int. Symp. Inf. Theory* ISIT., pages 1488–1492, Austin, 2010.

F. M. Willems. The context-tree weighting method: extensions. *IEEE Trans. Inform. Theory*, 44(2): 792–798, 1998.

Y. Yang and A. Barron. An asymptotic property of model selection criteria. *IEEE Trans. Inform. Theory*, 44:95–116, 1998.

A Proof of Proposition 1

Proof: (i) The inverse of the hazard rate h'_c is a positive non increasing function, thus its derivative converges to 0 at infinity, and (i) follows from Theorem 1.2.6 in de Haan and Fereira (2006).
(ii) Recall that $h'_c(t) = f_c(t)/\overline{F}_c(t)$. Since the hazard rate is non-decreasing, the derivative of $U_c \circ \exp$ is non-increasing.
To prove (iii), notice first that since the hazard rate is finite, U_c tends to infinity at infinity. U_c is non decreasing, so that for large enough t,

$$3 \ln t \leq \widetilde{h}(t) \leq U_c(t^2) \ln t. \tag{2}$$

Thus, it is enough to prove that for all $\eta \geq 0$ and $\kappa > 0$,

$$\lim_{x \to +\infty} \frac{\widetilde{h}(\kappa x (\widetilde{h}(x))^\eta)}{\widetilde{h}(x)} = 1.$$

Let $g \colon \mathbb{R}_+ \to \mathbb{R}_+$ be defined by $g(t) = \ln(\widetilde{h}(\exp(t))) = \ln\left(\int_0^t U_c(\exp(2x))\mathrm{d}x\right)$. It is enough to check that

$$\lim_{t \to \infty} g(t + \eta g(t) + z) - g(t) = 0$$

for $z \in \mathbb{R}, \eta > 0$. But,

$$g(t + \eta g(t) + z) - g(t) = \ln\left(1 + \frac{\int_t^{t+\eta g(t)+z} U_c(\exp(2x))\mathrm{d}x}{\int_0^t U_c(\exp(2x))\mathrm{d}x}\right).$$

For large enough t, $\eta g(t)+z>0$, and by concavity of $U_c \circ \exp$,

$$\int_t^{t+\eta g(t)+z} U_c(\exp(2x))\mathrm{d}x \le (z+\eta g(t))U_c\left(\exp\left(2t+\eta g(t)+z\right)\right)$$
$$\le (z+\eta g(t))U_c\left(\exp(2t)\right) + \frac{\overline{F}_c\left(U_c\left(\exp(2t)\right)\right)}{f_c\left(U_c\left(\exp(2t)\right)\right)}(z+\eta g(t))^2 .$$

Letting b be the infimum of the hazard rate,

$$g(t+\eta g(t)+z)-g(t) \le (z+\eta g(t))g'(t) + \frac{1}{b}\frac{(z+\eta g(t))^2}{\exp(g(t))} .$$

The second summand tends to 0 as t tends to infinity. Since $g(t)$ tends to infinity at infinity, there remains to prove that $g(t)g'(t)$ tends to 0 at infinity, that is to establish that $U_c(u^2)\ln\widetilde{h}(u)/e^{\widetilde{h}(u)}$ tends to 0 at infinity. But as $U_c(x)/x$ is regularly varying with index -1, $t\mapsto\int_0^t U_c(x)/x\mathrm{d}x$ is slowly varying (regularly varying with index 0) (See de Haan and Fereira 2006, Proposition B.1.9, Point 4) and so is $t\mapsto\widetilde{h}(t)=\int_0^{t^2} U_c(x)/x\mathrm{d}x$, this follows from Karamata integral representation Theorem (de Haan and Fereira 2006, Theorem B.1.6). So that using (2),

$$\frac{U_c(u^2)\ln\widetilde{h}(u)}{e^{\widetilde{h}(u)}} \le \frac{U_c(u^2)}{u^2}\frac{\ln\widetilde{h}(u)}{\ln(u)}\frac{\ln(u)}{u}$$

now, by (de Haan and Fereira 2006, Proposition B.1.9, Point 1), the first two factors tend to 0 as u tends to infinity, and (iii) follows.
To prove (iv), note that

$$\int_1^t \frac{U_c(x)}{x}\mathrm{d}x = \int_0^{\ln t} U_c(\exp(s))\mathrm{d}s \ge \frac{\ln(t)}{2}U_c(t)\,, \quad \text{by concavity of } U_c\circ\exp.$$

Plugging this upper bound leads to:

$$\frac{U_c(t)\ln(U_c(t))}{\int_1^t\frac{U_c(x)}{x}\mathrm{d}x} \le 2\frac{U_c(t)\ln(U_c(t))}{U_c(t)\ln(t)} = 2\frac{\ln(U_c(t))}{\ln(t)}$$

which tend to 0 as t tends to infinity (Again by de Haan and Fereira 2006, Proposition B.1.9, Point 1). □

B Proof of Proposition 2

In order to alleviate notation $\mathcal{H}_\epsilon$ is used as a shorthand for $\mathcal{H}_\epsilon(\Lambda_f)$. Upper and lower bounds for $\mathcal{H}_\epsilon$ follow by adapting the "flat concentration argument" in Bontemps (2011). The cardinality $\mathcal{D}_\epsilon$ of the smallest partition of Λ_f^1 into subsets of diameter less than ϵ is not larger than the smallest cardinality of a covering by Hellinger balls of radius smaller than $\epsilon/2$. Recall that Λ_f^1 endowed with the Hellinger distance may be considered as a subset of $\ell_2^{\mathbb{N}_+}$:

$$C=\left\{(x_i)_{i>0}\colon \sum_{i>0}x_i^2=1\right\}\bigcap\left\{(x_i)_{i>0}\colon \forall i>0, 0\le x_i\le\sqrt{f(i)}\right\}.$$

Let $N_\epsilon = U(\frac{16}{\epsilon^2})$ (N_ϵ is the $1-\epsilon^2/16$ quantile of the envelop distribution). Let D be the projection of C on the subspace generated by the first N_ϵ vectors of the canonical basis. Any element of C is at distance at most $\epsilon/4$ of D. Any $\epsilon/4$-cover for D is an $\epsilon/2$-cover for C. Now D is included in the intersection of the unit ball of a N_ϵ-dimensional Euclidian space and of an hyper-rectangle $\prod_{i=1}^{N_\epsilon}[0, \sqrt{f(k)}]$. An $\epsilon/4$-cover for D can be extracted from any maximal $\epsilon/4$-packing of points from D. From such a maximal packing, a collection of pairwise disjoint balls of radius $\epsilon/8$ can be extracted that fits into $\epsilon/8$-blowup of D. Let B_m be the m-dimensional Euclidean unit ball ($\mathrm{Vol}(B_m) = \Gamma(1/2)^m/\Gamma(m+1/2)$ with $\Gamma(1/2) = \sqrt{\pi}$). By volume comparison, $\mathcal{D}_\epsilon \times (\epsilon/8)^{N(\epsilon)}\mathrm{Vol}(B_{N_\epsilon}) \leq \prod_{i=1}^{N_\epsilon}\left(\sqrt{f(k)}+\epsilon/4\right)$, or

$$\mathcal{H}_\epsilon \leq \sum_{k=1}^{N_\epsilon} \ln\left(\sqrt{f(k)}+\epsilon/4\right) - \ln \mathrm{Vol}(B_{N_\epsilon}) + N_\epsilon \ln\frac{8}{\epsilon}\,.$$

Let $l = U(1)$ ($l = l_f + 1$). For $k \geq l$, $f(k) = \overline{F}(k-1)(1-\overline{F}(k)/\overline{F}(k-1))$. As the hazard rate of the envelope distribution is assumed to be non-decreasing, denoting the essential infimum of the hazard rate by b, $\overline{F}(k-1)(1-e^{-b}) \leq f(k) \leq \overline{F}(k-1)$. Hence, for $l \leq k \leq N_\epsilon$, $\sqrt{f(k)} \geq \epsilon/4\sqrt{1-e^{-b}}$. Thus

$$\begin{aligned}
\mathcal{H}_\epsilon &\leq \sum_{k=1}^{l_f} \ln\left(\sqrt{f(k)}+\epsilon/4\right) + \sum_{k=l}^{N_\epsilon} \ln(\sqrt{f(k)}) - \ln \mathrm{Vol}(B_{N_\epsilon}) + \frac{N_\epsilon - l_f}{\sqrt{1-e^{-b}}} + N_\epsilon \ln\frac{8}{\epsilon} \\
&\leq \sum_{k=l}^{N_\epsilon} \frac{1}{2}\ln\left(\frac{64\overline{F}(k-1)}{\epsilon^2}\right) - \ln \mathrm{Vol}(B_{N_\epsilon}) + \frac{N_\epsilon - l_f}{\sqrt{1-e^{-b}}} + l_f \ln\frac{8}{\epsilon} + \sum_{k=1}^{l_f} \ln\left(\sqrt{f(k)}+\epsilon/4\right). \quad (3)
\end{aligned}$$

Following Bontemps (2011), a lower bound is derived by another volume comparison argument. From any partition into sets of diameter smaller than ϵ, one can extract a covering by balls of radius ϵ. Then for any positive integer m, $\mathcal{D}_\epsilon \geq \frac{\prod_{k=l}^{l_f+m}\sqrt{f(k)}}{\epsilon^m \mathrm{Vol}(B_m)}$. Hence, choosing $m = N_\epsilon - l_f$

$$\begin{aligned}
\mathcal{H}_\epsilon &\geq \sum_{k=l}^{N_\epsilon} \ln\sqrt{f(k)} - \ln \mathrm{Vol}(B_{N_\epsilon - l_f}) + (N_\epsilon - l_f)\ln\frac{1}{\epsilon} \\
&\geq \sum_{k=l}^{N_\epsilon} \frac{1}{2}\ln\left(\frac{\overline{F}(k-1)(1-e^{-b})}{\epsilon^2}\right) - \ln \mathrm{Vol}(B_{N_\epsilon - l_f})\,. \quad (4)
\end{aligned}$$

Now,

$$\ln \mathrm{Vol}(B_{N_\epsilon}) = [N_\epsilon \ln N_\epsilon]\,(1+o(1)) = \left[U_c\left(\frac{16}{\epsilon^2}\right)\ln U_c\left(\frac{16}{\epsilon^2}\right)\right](1+o(1))$$

as ϵ tends to 0. Since $N_\epsilon \to \infty$, we have also $\ln \mathrm{Vol}(B_{N_\epsilon - l_f}) = [N_\epsilon \ln N_\epsilon]\,(1+o(1))$, as ϵ tends to 0. Now, the term $\sum_{k=l}^{N_\epsilon}\frac{1}{2}\ln\left(\frac{\overline{F}(k-1)}{\epsilon^2}\right)$ in (3) and (4) is treated by (1). The desired result follows from the fact that U_c and hence $U_c\ln(U_c)$ are slowly varying (Proposition 1 (i)) and from Proposition 1 (iv).

C Proof of equation (1)

Making the change of variable $y = U_c(x)$ ($x = 1/\overline{F}_c(y)$, $\frac{\mathrm{d}x}{\mathrm{d}y} = \frac{f_c(y)}{(\overline{F}_c(y))^2}$),

$$\int_1^t \frac{U_c(x)}{2x}\mathrm{d}x = \int_{l_f-1}^{U_c(t)} \frac{y f_c(y)}{2\overline{F}_c(y)}\mathrm{d}y = \frac{U_c(t)}{2}\ln(t) + \int_0^{U_c(t)} \frac{\ln(\overline{F}_c(y))}{2}\mathrm{d}y = \int_0^{U_c(t)} \frac{\ln(t\overline{F}_c(x))}{2}\mathrm{d}x,$$

where the second equation follows by integration by parts.