About adaptive coding on countable alphabets

Dominique Bontemps, Stéphane Boucheron, Elisabeth Gassiat

About Adaptive Coding on Countable Alphabets

Dominique Bontemps[‡] Stéphane Boucheron[*] Elisabeth Gassiat[†]

Abstract

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: countable alphabets; redundancy; adaptive compression; minimax;

I. Introduction

A. Universal coding over countable alphabets

This paper is concerned with problems of universal 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 $\mathcal{X}$. In this paper, the symbol Λ will be used to 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 $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$. 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 $\mathcal{X}$.

The *expected redundancy* of any (coding) distribution $Q^n \in \mathfrak{M}_1(\mathcal{X}^n)$ with respect to $\mathbb{P}$ is equal to the Kullback-Leibler divergence (or relative entropy) between $\mathbb{P}^n$ and Q^n: $D(\mathbb{P}^n, Q^n) = \sum_{\mathbf{x}\in\mathcal{X}^n} \mathbb{P}^n\{\mathbf{x}\} \log \frac{\mathbb{P}^n(\mathbf{x})}{Q^n(\mathbf{x})} = \mathbb{E}_{\mathbb{P}^n}\left[\log \frac{\mathbb{P}^n(X_{1:n})}{Q^n(X_{1:n})}\right]$. Notice that the definition of redundancy uses base 2 logarithms. Throughout this note, $\log x$ denotes the base 2 logarithm of x while $\ln x$ denotes its natural logarithm.

Universal coding attempts to develop sequences of coding probabilities $(Q^n)_n$ so as to minimize expected 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:

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

A corollary of early results by Kieffer (1978), Gyorfi, Pali, and van der Meulen (1993; 1994) shows that finite minimax redundancy is not a trivial property as soon as the alphabet is infinite even for classes of memoryless sources.

Proposition 1: If a class Λ of stationary sources over a countable alphabet $\mathcal{X}$ has finite *minimax redundancy* then there exists a probability distribution $Q \in \mathfrak{M}_1(\mathcal{X})$ such that for every $\mathbb{P} \in \Lambda$ with $\lim_n H(\mathbb{P}^n)/n < \infty$ where $H(\mathbb{P}^n) = \sum_{\mathbf{x}\in\mathcal{X}} -\mathbb{P}^n(\mathbf{x}) \log \mathbb{P}^n(\mathbf{x})$ (finite Shannon entropy rate), Q satisfies $D(\mathbb{P}^1, Q) < \infty$.

This observation contrasts with what we know about the finite alphabet setting where coding probabilities asymptotically achieving minimax redundancies have been described (Xie and Barron, 2000; Barron et al., 1998; Yang and Barron, 1998; Xie and Barron, 1997; Clarke and Barron, 1994). Note that delicate asymptotic results for coding over large finite alphabets with unknown size have started to appear (Szpankowski and Weinberger, 2010).

This prompted Boucheron, Garivier, and Gassiat (2009) to study the redundancy of specific memoryless source classes, namely classes defined by an envelope function.

B. Envelope classes

Definition 1: Let f be a mapping from $\mathbb{N}_+$ to $[0, 1]$, with $1 \le \sum_{j>0} f(j) < \infty$. The *envelope class* Λ_f defined by the function f is the collection of stationary memoryless sources with first marginal distribution dominated by f:

$$\Lambda_f = \Big\{\mathbb{P} \;:\; \forall x \in \mathbb{N},\ \mathbb{P}^1\{x\} \le f(x)\ ,\ \text{and } \mathbb{P} \text{ is stationary and memoryless.}\Big\}\,.$$

[‡]supported by Institut de Mathématiques de Toulouse, Université de Toulouse
[*]supported by Network of Excellence PASCAL II, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris-Diderot
[†]supported by Network of Excellence PASCAL II, Laboratoire de Mathématiques d'Orsay, Université Paris-Sud

Definition 2: Let f be an envelope function. 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: $R^+(\Lambda_f^n) \leq \inf_{u\leq n} \left[n\log(1+\overline{F}(u)) + \frac{u-1}{2}\log n \right] + 2$. 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 raises the question: Is it possible to describe the minimax redundancy as a simple functional of the envelope distribution?

C. Adaptive coding

A sequence $(Q^n)_n$ of coding probabilities (see Cover and Thomas, 1991, for a gentle introduction to the notion of coding probability) is said to be *asymptotically adaptive* with respect to a collection $(\Lambda_m)_{m\in\mathcal{M}}$ of source classes if for all $m \in \mathcal{M}$:

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

as n tends to infinity. Effective and (almost) adaptive coding techniques for the collection of source classes defined by algebraically vanishing envelopes were introduced in (Boucheron, Garivier, and Gassiat, 2009). Moreover, Bontemps designed and analyzed the AC-code (Auto-Censuring code) and proved that this 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 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). This raises a second question: Is it possible to design coding probabilities that are adaptive with respect to larger families of envelope classes, and how do these coding probabilities depend on the collection of envelope distribution functions?

D. Answers and organization of the paper

This paper aims at clarifying the difficulty of universal coding with respect to envelope classes. We provide positive and precise answers to the aforementioned questions for a family of envelope classes that lie between the exponential envelope classes investigated in (Boucheron et al., 2009; Bontemps, 2011) and the classes of sources with finite alphabets.

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.

Borrowing ideas from extreme value theory, we prove that if the envelope distribution function has finite and non decreasing hazard rate (defined in Section II): i) an explicit formula connects the minimax redundancy and the envelope distribution; ii) 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 II provides notation and definitions concerning hazard rates. Section III describes the AC-code. 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 IV. The minimax redundancy of source classes defined by envelopes with finite and non-decreasing hazard rate is characterized in Section V. Section VI 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.

II. Hazard rate and continuous envelope distribution function

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 $t \geq 0$. The tail function of F_c is $\overline{F}_c(t) = \exp(-h_c(t))$ for $t > 0$. For all integers n, $\overline{F}_c(n) = \overline{F}(n)$. The hazard rate h_c' is piecewise constant, it equals

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

The essential infimum of the hazard rate is $b = -\ln \overline{F}(l_f) > 0$. Notice that the hazard rate is finite on $[l_f - 1, \infty)$ 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 $t > 1$.

The envelopes introduced in the next definition provide examples where the associated continuous distribution function has non-decreasing hazard rate. Poisson distributions offer other examples.

Definition 3: 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

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

Exponentially vanishing envelopes (Boucheron et al., 2009; Bontemps, 2011) are obtained by fixing $\alpha = 1$.

III. The AC-code

The AC-code encodes a sequence $x_{1:n} = x_1, \ldots, x_n$ of symbols from $\mathbb{N}_+ = \mathbb{N} \setminus \{0\}$ in the following way. For $i \colon 1 \leq i \leq n$, 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_i^0 be the number of records up to index i. The j^{th} record is denoted by $\widetilde{m}_j$. From the definitions, $\widetilde{m}_{n_i^0} = m_i$ for all i. Let $\widetilde{m}_0 = 0$ and let $\widetilde{\mathbf{m}}$ be the sequence of differences between records terminated by a 1, $\widetilde{\mathbf{m}} = (\widetilde{m}_i - \widetilde{m}_{i-1} + 1)_{1 \leq i \leq n_n^0} 1$ (the last 1 in the sequence serves as a terminating symbol). The symbols in $\widetilde{\mathbf{m}}$ are encoded using Elias penultimate code (Elias, 1975). 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

$$Q_{i+1}\left(\widetilde{X}_{i+1} = j \mid X_{1:i} = x_{1:i}\right) = \frac{n_i^j + \frac{1}{2}}{i + \frac{m_i + 1}{2}}$$

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 $j > 0$. If $i < n$, the event $\{\widetilde{X}_{i+1} = 0\} = \{X_{i+1} = M_{i+1} > M_i\}$ has conditional probability $Q_{i+1}\big(\widetilde{X}_{i+1} = 0 \mid X_{1:i} = x_{1:i}\big) = \frac{1/2}{i + (m_i+1)/2}$. Note that 0 is always encoded as a new symbol: if $x_{i+1} = j > m_i$, the AC-code encodes a 0, but n_i^j rather than n_i^0 is incremented. In words, the mixture code consists of progressively enlarging the alphabet and feeding an arithmetic coder with Krichevsky-Trofimov mixtures over the smallest alphabet seen so far (Cesa-Bianchi and Lugosi, 2006). Bontemps (2011) describes a nice way of interleaving the Elias codewords and the mixture code in order to perform online encoding and decoding.

IV. Main result

The main result may be phrased as follows.

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 $\cup_{\beta>0,\gamma>1}\Lambda(1,\beta,\gamma)$. Theorem 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 5 from Section V 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 VI. The pointwise redundancy is upper bounded in the following way:

$$\begin{aligned}&-\log Q^n(X_{1:n}) + \log \mathbb{P}^n(X_{1:n})\\ &\leq \underbrace{\ell(C_E)}_{\text{I}} + \underbrace{\ell(C_M) + \log \mathbb{P}^n(\widetilde{X}_{1:n})}_{\text{II}}\,.\end{aligned}$$

Proposition 9 asserts that (I) is negligible with respect to $R^+(\Lambda_f^n)$ and Proposition 10 asserts that the expected value of (II) is equivalent to $R^+(\Lambda_f^n)$.

V. 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 miximax 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.

We also need to introduce further notation.

Definition 4: 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 $\kappa > 0$,

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

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).

Theorem 2: (Haussler and Opper, 1997, Theorem 5) Assume there exists a very slowly varying function h such that:

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

Then

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

This theorem tightly characterizes the asymptotic redundancy of small source classes. We shall see that source classes defined by envelopes with finite and non-decreasing hazard rate are small. Notice that the definition of redundancy uses base 2 logarithms while ϵ-entropy is usually defined using natural logarithms.

Let us now state some analytic properties that will prove useful when checking that source classes defined by envelopes with finite and non-decreasing hazard rate are indeed small.

Proposition 2: 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,
(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)

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

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) The derivative of $U_c \circ \exp$ is equal to $\overline{F}_c(U_c(\exp(t)))/f_c(U_c(\exp(t)))$ which is non-increasing as the hazard rate is non-decreasing.
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,

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

Thus, it is enough to prove that for all $\eta \ge 0$ 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,

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

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

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

Letting b be the infimum of the hazard rate,

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

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

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

Plugging this upper bound leads to:

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

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

Proposition 3: (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

$$\mathcal{H}_\epsilon(\Lambda_f) = (1+o(1))\int_0^{1/\epsilon^2} \frac{U_c(x)}{2x}\mathrm{d}x$$

as ϵ 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 2 (iii):

Proposition 4: (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

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

as n tends to $+\infty$.

A concrete corollary follows easily.

Proposition 5: The minimax redundancy of the sub-exponential envelope class with parameters (α,β,γ) satisfies

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

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

$$\beta\left(\ln(\gamma t)\right)^{1/\alpha} - 1 \leq U_c(t) \leq \beta\left(\ln(\kappa\gamma t)\right)^{1/\alpha} - 1 \tag{3}$$

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

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

The upper bound follows from

$$\begin{aligned} \overline{F}(k) &\leq \sum_{j\geq 0}\gamma\exp\left(-\left(\tfrac{k+1}{\beta}\right)^\alpha - j\alpha\tfrac{(k+1)^{\alpha-1}}{\beta^\alpha}\right) \\ &\leq \frac{f(k+1)}{1-\exp\left(-\alpha(k+1)^{\alpha-1}/\beta^\alpha\right)} \leq \frac{f(k+1)}{\kappa}, \end{aligned}$$

for $\alpha \geq 1$. ■

VI. Redundancy of AC-code

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 9).

A. 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 6: Let $Y_1,\ldots,Y_n$ be independently identically distributed according to an absolutely continuous distribution function F with density $f=F'$ 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 $Y_{1,1}\leq\ldots\leq Y_{n,n}$ be the order statistics. Then,

$$\begin{aligned} \mathbb{E}[Y_{n,n}] &\leq U(\exp(H_n)) \\ \mathbb{E}[Y_{n,n}^2] &\leq \mathbb{E}[Y_{n,n}]^2 + 2/b^2 \\ \mathbb{E}[Y_{n,n}\ln(Y_{n,n})] &\leq (\mathbb{E}Y_{n,n})\ln(\mathbb{E}Y_{n,n})) + 2/b^2 \text{ if } Y_i \geq 1 \text{ a.s.} \end{aligned}$$

Note that if the hazard rate is strictly increasing, $Y_{n,n} - U(n)$ satisfies a law of large numbers (See Anderson, 1970).

The proof of proposition 6 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.

Proposition 7: Let $(Y_{1,n}, \ldots, Y_{n,n})$ denote the order statistics of an independent sample picked according to a common probability distribution, then

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

and if the random variables $(Y_i)_{i \leq n}$ are non-negative,

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

In the proof of this proposition, $\mathbb{E}^{(i)}$ denotes conditional expectation with respect to $Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n$ and for each Z_i denotes 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. Order statistics are functions of independent random variables. The next theorem, the proof of which can be found in (Ledoux, 2001) has proved to be a powerful tool when investigating the fluctuations of independent random variables (See for example Efron and Stein, 1981; Massart, 2007) .

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

$$\begin{aligned}\mathrm{Var}\left(Z\right) &\leq \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - \mathbb{E}^{(i)}Z\right)^2\right] \\ &\leq \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - Z_i\right)^2\right] ,\end{aligned}$$

and if Z and all $Z_i, 1 \leq i \leq n$, are positive,

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

Proof of Proposition 7: As $\mathbb{E}[Y_{n,n}^2] = \mathrm{Var}(Y_{n,n}) + (\mathbb{E}Y_{n,n})^2$, it is enough to bound $\mathrm{Var}(Y_{n,n})$. As $Z = Y_{n,n}$ is a function of n independent random variables $Y_1, \ldots, Y_n$, choosing the Z_i as $\max(X^{(i)})$, $Z_i = Z$ except possibly when $X_i = Z$, and then $Z_i = Y_{n-1,n}$. The sub-additivity property of the variance imply that

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

Using the sub-additivity of entropy with the convention about Z_i,

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

as $\ln(1+u) \leq u$ for $u > -1$. ■

Proof of Proposition 6: Thanks to Rényi's representation of order statistics, $\mathbb{E}[Y_{n,n}] = \mathbb{E}[U(\exp(E_{n,n}))]$, the proof of the first statement follows from the concavity of $t \mapsto U(\exp(t))$, that is from Proposition 2 ii).

By the Efron-Stein inequality (See Proposition 7),

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

Thanks again 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

$$\begin{aligned} &U(\exp(E_{n,n})) - U(\exp(E_{n-1,n})) \\ &\leq \frac{\overline{F}(U(\exp(E_{n-1,n})))}{f(U(\exp(E_{n-1,n})))}(E_{n,n} - E_{n-1,n}) . \end{aligned}$$

The two factors are independent. While $\mathbb{E}[(E_{n,n} - E_{n-1,n})^2] = 2$,

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

By Proposition 7,

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

■

When handling subexponential envelopes classes, Proposition 6 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 $\Lambda(\alpha, \beta, \gamma)$, let $Y_1, \ldots, Y_n$ be identically independently distributed according to the probability distribution with tail function $\overline{F}(u) = 1 \wedge \sum_{k>u} f(k)$ where $f(u) = \gamma \exp(-(u/\beta)^\alpha)$. The quantile coupling argument ensures that there exists a probability space with random variables $(X'_1, \ldots, X'_n)$ distributed like $(X_1, \ldots, X_n)$ and random variables $(Y'_1, \ldots, Y'_n)$ distributed like $(Y_1, \ldots, Y_n)$ and $X'_i \leq Y'_i$ for all $i \leq n$ almost surely.

Let $Y_{(1)} \leq \ldots \leq Y_{(n)}$ denote the order statistics of $Y_1, \ldots, Y_n$, then for any non-decreasing function g, $\mathbb{E}[g(M_n)] \leq \mathbb{E}[g(Y_{(n)})]$. Using (3) one gets the following.

Proposition 8: Let $X_1, \ldots, X_n$ be independently identically distributed according to $P \in \Lambda^1(\alpha, \beta, \gamma)$, let $M_n = \max(X_1, \ldots, X_n)$, then,

$$\begin{aligned} \mathbb{E}M_n &\leq \beta \left(\ln\left(\kappa\gamma e n\right)\right)^{1/\alpha} . \\ \mathbb{E}[M_n \log M_n] &\leq \beta \left(\ln\left(\kappa\gamma e n\right)\right)^{1/\alpha} \\ &\qquad \times \left(\ln\beta + \frac{1}{\alpha}\ln(\ln(\kappa\gamma e n))\right) + 2\kappa^2 \\ \mathbb{E}[M_n^2] &\leq \beta^2 \left(\ln\left(\kappa\gamma e n\right)\right)^{2/\alpha} + 2\kappa^2. \end{aligned}$$

It provides a simple refinement of Lemma 4 from (Bontemps, 2011).

B. 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 as far as subexponential envelope classes are concerned, it contributes in a negligible way to the total redundancy.

Proposition 9: 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 $\rho = 2$).
In general if $X_1, \ldots, X_n$ are independently identically distributed, the following holds

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

For classes defined by power law envelopes, M_n grows like a power of n, the last upper bound shows that the length of the Elias encoding of records remains polylogarithmic with respect to n while the minimax redundancy grows like a power

of n (Boucheron et al., 2009). However the AC-code is not likely to achieve the minimax redundancy over classes defined by power-law envelopes.

The last statement stems from the fact that the Elias codelength is less than a concave function of the encoded value. The average Elias codelength of record differences should not be larger than the Elias codelength of the average record difference which is the maximum divided by the number of records.

Proof of Proposition 9: The length of the Elias codewords used to encode the sequence of record differences $\widetilde{\mathbf{m}}$ is readily upper-bounded:

$$\begin{aligned}\ell(C_E) &\le \sum_{i=1}^{n_n^0} (2\log\left(1+\widetilde{m}_i-\widetilde{m}_{i-1}\right)+\rho)\\ &\le \sum_{i=1}^{n_n^0} 2\log(e)\left(\widetilde{m}_i-\widetilde{m}_{i-1}\right)+\rho n_n^0\\ &\le 2\log(e)M_n+\rho n_n^0\\ &\le (2\log(e)+\rho)M_n\end{aligned}$$

for some universal constant ρ. The bound on the length of the Elias encoding follows from Proposition 8.

If we were only interested in subexponential envelope classes, this would be enough. The next lines may be used to establish that the length of the Elias encoding remains negligible with respect to minimax redundancy for larger envelope source classes.

Using the arithmetic-geometric mean inequality and $\sum_{i=1}^{n_n^0}\left(\widetilde{m}_i-\widetilde{m}_{i-1}\right)\le M_n$, we also have

$$\begin{aligned}\ell(C_E) &\le 2\sum_{i=1}^{n_n^0}\log\left(1+\widetilde{m}_i-\widetilde{m}_{i-1}\right)+n_n^0\rho\\ &\le 2n_n^0\log(1+M_n/n_n^0)+n_n^0\rho\,.\end{aligned}$$

The average length of C_E satisfies:

$$\begin{aligned}&\mathbb{E}\left[\ell(C_E)\right]\\ &\le \mathbb{E}\left[n_n^0(2\log(1+M_n/n_n^0)+\rho)\right]\\ &\le 2\left(\mathbb{E}\left[n_n^0\log(2M_n)\right]-\mathbb{E}n_n^0(\log\mathbb{E}n_n^0-\rho)\right)\\ &\quad\text{by concavity of } x\mapsto -x\log x\\ &\le 2\left(\Big(\sum_{i=1}^{n}\frac{1}{i}\Big)\mathbb{E}\log(2(M_n+1))-\mathbb{E}n_n^0(\log\mathbb{E}n_n^0-\rho)\right)\\ &\le 2\ln(en)(\log(2\mathbb{E}[M_n+1])+\rho)\,.\end{aligned}$$

The penultimate inequality comes from the following observation. Any integer valued random variable can be represented as the integer part of a real valued random variable with absolutely continuous distribution. For example, $X_1,\ldots,X_n$ may be represented as $\lfloor Y_1\rfloor,\ldots,\lfloor Y_n\rfloor$ where $Y_1,\ldots,Y_n$ are i.i.d. and supported by $\cup_{n\in\mathbb{N}_+}[n,n+1/2]$. Any record in $X_1,\ldots,X_n$ comes from a record in $Y_1,\ldots,Y_n$ (but the converse may not be true). Letting R_n denote the number of records in $Y_1,\ldots,Y_n$, we have $n_n^0\log(M_n)\le R_n\log(\max(Y_1,\ldots,Y_n))$. Moreover R_n and $\max(Y_1,\ldots,Y_n)$ are independent, and R_n is a sum of independent Bernoulli random variables with parameters $1,1/2,\ldots,1/n$. This entails

$$\begin{aligned}&\mathbb{E}[R_n\log(\max(Y_1,\ldots,Y_n))]\\ &=\mathbb{E}R_n\mathbb{E}[\log(\max(Y_1,\ldots,Y_n))]\\ &\le\sum_{i=1}^{n}\frac{1}{i}\log(2\mathbb{E}\max(Y_1,\ldots,Y_n))\\ &\le H_n\log(2(\mathbb{E}M_n+1))\,.\end{aligned}$$

■

C. Adaptive mixture coding

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

Proposition 10: 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 its ideal codeword length is upper-bounded by

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

as n tends to infinity.

The proof of Proposition 10 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 11 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 11: (POINTWISE BOUND) Let i_0 be the random integer defined by: $i_0 = 1 \vee \lfloor M_n/4 \rfloor$, then,

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

Proof: Let C_M be the mixture encoding of $\widetilde{X}_{1:n}$, then $\ell(C_M) = -\log Q^n(\widetilde{X}_{1:n})$. The pointwise redundancy can be decomposed into

$$\begin{aligned} &- \ln Q^n(\widetilde{X}_{1:n}) + \ln \mathbb{P}^n(\widetilde{X}_{1:n}) \\ &= \underbrace{- \ln \text{KT}_{M_n+1}(\widetilde{X}_{1:n}) + \ln \mathbb{P}^n(\widetilde{X}_{1:n})}_{\text{(A)}} \\ &\quad \underbrace{- \ln Q^n(\widetilde{X}_{1:n}) + \ln \text{KT}_{M_n+1}(\widetilde{X}_{1:n})}_{\text{(B)}} \end{aligned}$$

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),

$$\begin{aligned} \text{(A)} &= - \ln(\text{KT}_{M_n+1}(\widetilde{X}_{1:n})) + \ln(\mathbb{P}^n(\widetilde{X}_{1:n})) \\ &\leq \frac{M_n+1}{2} \ln(n) + 2\ln(2)\,. \end{aligned}$$

The second summand (B) is negative, this is the codelength the AC-code pockets by progressively enlarging the alphabet rather than using $\{0, \ldots, M_n\}$ 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}:

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

The difference between the codelengths can be further upper bounded.

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

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

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

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

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

■

Proof of Proposition 10: The average redundancy of the mixture code is thus upper bounded by

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

We may now use the maximal inequalities from Proposition 6.

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

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

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

Now using Proposition 2 (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*, vol. 7, pp. 99–113, 1970.

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

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

D. Bontemps, "Universal coding on infinite alphabets: exponentially decreasing envelopes." *IEEE Trans. Inform. Theory*, vol. 57, no. 3, pp. 1466–1478, 2011.

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

O. Catoni, *Statistical learning theory and stochastic optimization*, ser. Lecture Notes in Mathematics. Springer-Verlag, 2004, vol. 1851, 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*, vol. 41, pp. 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.*, vol. 11, no. 2, pp. 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.
B. Efron and C. Stein, "The jackknife estimate of variance," *Annals of Statistics*, vol. 9, no. 3, pp. 586–596, 1981.
P. Elias, "Universal codeword sets and representations of the integers," *IEEE Trans. Information Theory*, vol. IT-21, pp. 194–203, 1975.
A. Garivier, "Redundancy of the context-tree weighting method on renewal and Markov renewal processes," *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5579–5586, 2006.
L. Gyorfi, I. Pali, and E. van der Meulen, "On universal noiseless source coding for infinite source alphabets." *Eur. Trans. Telecommun. & Relat. Technol.*, vol. 4, no. 2, pp. 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*, vol. 40, no. 1, pp. 267–271, 1994.
D. Haussler and M. Opper, "Mutual information, metric entropy and cumulative relative entropy risk," *Annals of Statistics*, vol. 25, no. 6, pp. 2451–2492, 1997.
G. Kerkyacharian and D. Picard, "Minimax or maxisets?" *Bernoulli*, vol. 8, no. 2, pp. 219–253, 2002.
J. C. Kieffer, "A unified approach to weak universal source coding," *IEEE Trans. Inform. Theory*, vol. 24, no. 6, pp. 674–682, 1978.
M. Ledoux, *The concentration of measure phenomenon*. AMS, 2001.
P. Massart, *Concentration inequalities and model selection. Ecole d'Eté de Probabilité de Saint-Flour* XXXIV, ser. Lecture Notes in Mathematics. Springer-Verlag, 2007, vol. 1896.
W. Szpankowski and M. Weinberger, "Minimax redundancy for large alphabets." in *Proceeding 2010 Int. Symp. Inf. Theory* ISIT., Austin, 2010, pp. 1488–1492.
F. M. Willems, "The context-tree weighting method: extensions," *IEEE Trans. Inform. Theory*, vol. 44, no. 2, pp. 792–798, 1998.
Q. Xie and A. R. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Trans. Inform. Theory*, vol. 43, pp. 646–656, 1997.
——, "Asymptotic minimax regret for data compression, gambling and prediction," *IEEE Trans. Inform. Theory*, vol. 46, pp. 431–445, 2000.
Y. Yang and A. Barron, "An asymptotic property of model selection criteria," *IEEE Trans. Inform. Theory*, vol. 44, pp. 95–116, 1998.

APPENDIX

A. Proof of Proposition 3

In order to alleviate notation $\mathcal{H}_\epsilon$ is used as a shorthand for $\mathcal{H}_\epsilon(\Lambda_f^1)$. 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,

$$\mathcal{D}_\epsilon \times (\epsilon/8)^{N(\epsilon)} \mathrm{Vol}(B_{N_\epsilon}) \le \prod_{i=1}^{N_\epsilon} \left(\sqrt{f(k)} + \epsilon/4\right),$$

or

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

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

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

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,

$$\mathcal{D}_\epsilon \geq \frac{\prod_{k=l}^{l_f+m} \sqrt{f(k)}}{\epsilon^m \text{Vol}(B_m)}.$$

Hence, choosing $m = N_\epsilon - l_f$

$$\begin{aligned}
\mathcal{H}_\epsilon &\geq \sum_{k=l}^{N_\epsilon} \ln \sqrt{f(k)} - \ln \text{Vol}(B_{N_\epsilon - l_f}) + (N_\epsilon - l_f) \ln \frac{1}{\epsilon} \\
&\geq \sum_{k=l}^{N_\epsilon} \frac{1}{2} \ln\left(\frac{\overline{F}(k-1)(1-e^{-b})}{\epsilon^2}\right) - \ln \text{Vol}(B_{N_\epsilon - l_f})
\end{aligned} \tag{5}$$

Now,

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

as ϵ tends to 0. Since $N_\epsilon \to \infty$, we have also $\ln \text{Vol}(B_{N_\epsilon - l_f}) = [N_\epsilon \ln N_\epsilon]\,(1 + o(1))$, as ϵ tends to 0.
Now, the term $\sum_{k=l}^{N_\epsilon} \frac{1}{2} \ln\left(\frac{\overline{F}(k-1)}{\epsilon^2}\right)$ n (4) and (5) 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 2 (i)) and from Proposition 2 (iv).

B. Proof of equation (1)

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

$$\begin{aligned}
\int_1^t \frac{U_c(x)}{2x} \mathrm{d}x &= \int_{l_f - 1}^{U_c(t)} \frac{y f_c(y)}{2\overline{F}_c(y)} \mathrm{d}y \\
&= \left[-\frac{y}{2} \ln(\overline{F}_c(y))\right]_{l_f-1}^{U_c(t)} + \int_{l_f-1}^{U_c(t)} \frac{\ln(\overline{F}_c(y))}{2} \mathrm{d}y \\
&= \frac{U_c(t)}{2} \ln(t) + \int_0^{U_c(t)} \frac{\ln(\overline{F}_c(y))}{2} \mathrm{d}y \\
&= \int_0^{U_c(t)} \frac{\ln(t\overline{F}_c(x))}{2} \mathrm{d}x,
\end{aligned}$$

where the second equation follows by integration by parts.