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 adaptive coding with respect to classes of memoryless sources over a countable alphabet defined by an envelope function with finite and non-decreasing hazard rate (log-concave envelope distributions). 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.**

Index Terms—countable alphabets, redundancy, adaptive compression, minimax.

I. Introduction

A. From universal coding to adaptive coding

This paper is concerned with problems of *adaptive* that is *twice universal* or *hierarchical 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 $\Lambda^n = \{\mathbb{P}^n : \mathbb{P} \in \Lambda\}$.

Throughout the paper, we will rely on the correspondence between non-ambiguous codes and probability distributions and refer to codes through coding probabilities (see Cover and Thomas, 1991, for a gentle introduction to the notion of coding probability). 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,

$$D(\mathbb{P}^n, Q^n) = \sum_{\mathbf{x}\in\mathcal{X}^n} \mathbb{P}^n\{\mathbf{x}\} \log \frac{\mathbb{P}^n(\mathbf{x})}{Q^n(\mathbf{x})} = \mathbb{E}_{\mathbb{P}^n}\left[\log \frac{\mathbb{P}^n(X_{1:n})}{Q^n(X_{1:n})}\right].$$

Up to a constant, the expected redundancy of Q^n with respect to $\mathbb{P}$ is the expected difference between the length of codewords defined by encoding messages as if they were produced by Q^n and the ideal codeword length when encoding messages produced by $\mathbb{P}^n$. In the language of mathematical statistics, it is also the *cumulative entropy risk* suffered by estimating $\mathbb{P}^n$ using Q^n.

Notice that the definition of redundancy uses base 2 logarithms. Throughout this text, $\log x$ 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 known class of sources. The *maximal redundancy* of coding probability Q^n with respect to source class 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,

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

As the design of almost minimax coding probabilities is usually not a trivial task, looking for an apparently more ambitious goal, *adaptivity*, may seem preposterous. Indeed, whereas universality issues are ubiquitous in lossless coding theory (Csiszár and Körner, 1981; Cover and Thomas, 1991), and adaptivity has been a central concept in Statistics during the last two decades (See Bickel et al., 1998; Barron et al., 1999; Donoho and Johnstone, 1994; Donoho et al., 1996; Abramovich et al., 2006; Tsybakov, 2004, and references therein), the very word adaptivity barely made its way in the lexicon of Information Theory. Nevertheless, adaptivity issues have been addressed in coding theory, sometimes using different expressions to name things. Adaptive coding is sometimes called twice universal coding (Ryabko, 1984, 1990; Ryabko and Topsøe, 2002; Ryabko et al., 2008) or hierarchical universal coding (Merhav and Feder, 1998). We pursue this endeavour.

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 $m \in \mathcal{M}$,

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

as n tends to infinity. In words, a sequence of coding probabilities is adaptive with respect to a collection of source classes if it asymptotically achieves minimax redundancy over all classes. Note that this is a kind of first order requirement, the $o_m(1)$ term may tend to 0 at a rate that depends on the source class Λ_m.

This is not the only way of defining adaptive compression, more stringent definitions are possible (See Catoni, 2004, Section 1.5). This last reference describes oracle inequalities

[§]The material in this paper was presented in part at the 23rd Intern. Meeting on Probabilistic, Combinatorial, and Asymptotic Methods for the Analysis of Algorithms (AofA'12), Montréal, Québec, Canada, June 2012

[‡]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

for the context-tree weighting method (Willems, 1998), a successful attempt to achieve adaptivity with respect to an infinite collection of source classes on a finite alphabet indexed by their memory structure (Catoni, 2004).

The present paper describes another successful attempt to achieve adaptivity with respect to an infinite collection of source classes on a countable alphabet.

B. Adaptive coding with respect to a collection of envelope classes

Pioneering results by Kieffer (1978), Gyorfi, Pali, and van der Meulen (1993; 1994) show that, as soon as the alphabet is infinite, finite minimax redundancy, that is the possibility of achieving universality, is not a trivial property 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 over $\mathcal{X}$ 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). This even contrasts with recent delicate asymptotic results for coding over large finite alphabets with unknown size (Szpankowski and Weinberger, 2012; Yang and Barron, 2013).

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

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

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

An envelope function defines an *envelope distribution*. The minimax redundancy of the source classes we are interested in is, up to the first order, asymptotically determined by the tail behavior of the envelope distribution.

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. The next theorem provides an upper-bound on the minimax redundancy of envelope classes and suggests general design principles for universal coding over envelope classes and for adaptive coding over a collection of envelope classes.

Theorem 1. *(Boucheron, Garivier, and Gassiat, 2009) If Λ is an envelope class of memoryless sources, with the tail envelope function $\overline{F}$ then:*

$$R^+(\Lambda^n) \leq \inf_{u:u\leq n} \left[n\overline{F}(u) \log e + \frac{u-1}{2} \log n \right] + 2 \, .$$

If the envelope F is known, if the message length n is known, the following strategy is natural: determine u such that $\overline{F}(u) \approx \frac{1}{n}$; choose a good universal coding probability for memoryless sources over alphabet $\{0, \ldots, u\}$; escape symbols larger than u using 0 which does not belong to the source alphabet; encode the escaped sequence using the good universal coding probability; encode all symbols larger than u using a coding probability tailored to the envelope distribution. If the upper bound is tight, this strategy should achieve the minimax redundancy rate. If the message length is not known in advance, using a doubling technique should allow to derive an online extension. As naive as this approach may look, it has already proved fruitful. Bontemps (2011) showed that the minimax redundancy of classes defined by exponentially vanishing envelopes is half the upper bound obtained by choosing u so as $\overline{F}(u) \approx 1/n$.

If we face a collection of possible envelope classes and the envelope is not known in advance, we face two difficulties: there is no obvious way to guess a reasonable threshold; once a threshold is chosen, there is no obvious way to choose a coding probability for escaped symbols.

There are reasons to be optimistic. Almost adaptive coding techniques for the collection of source classes defined by power law envelopes were introduced in (Boucheron, Garivier, and Gassiat, 2009). Moreover, Bontemps designed and analyzed the AC-code (Auto-Censuring code) (described in Section II) and proved that this simple computationally efficient online code is adaptive over the union of classes of sources with exponentially decreasing envelopes (see Definition 3). As the AC-code does not benefit from any 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 (Csiszár and Shields, 1996) 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).

C. Roadmap

This paper shows that the AC-code is adaptive over the collection 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 (Theorem 2). The relevant envelopes are characterized by the fact that they have non-decreasing hazard rate (see Section III). This distributional property implies that the corresponding envelope distributions fit nicely in the framework of extreme value theory (See Section V-B), smoothed version of the envelope distribution belong to the so-called Gumbel domain of attraction, and this implies strong concentration properties for maxima of i.i.d. samples distributed according to envelope distributions. As the AC-code uses mixture coding over the *observed alphabet* in a sequential way, the intuition provided by Theorem 1 suggests that the AC-code should perform well when the largest symbol in a message of length n is close to the quantile of order $1 - 1/n$ of the envelope distribution. This concentration property is a consequence of the non-decreasing hazard rate assumption (Boucheron and Thomas, 2012). Moreover we check in the Appendix (see Section D) that if the sampling distribution has the non-decreasing hazard rate property, on average, the size of the largest symbol and the number of distinct symbols in the sample differ by a constant.

The non-decreasing hazard rate assumption has far reaching implications concerning the slow variation property of the quantile function of the envelope distribution (Section V-B) that prove instrumental in the derivation of matching lower bounds for the minimax redundancy of the corresponding envelope classes. In Section V, we revisit the powerful results concerning extensions of minimax redundancy by Haussler and Opper (1997). Advanced results from regular variation theory shed new light on the small classes where the lower bounds from (Haussler and Opper, 1997) are known to be tight

In words, borrowing ideas from extreme value theory (Falk et al., 2011; de Haan and Ferreira, 2006; Beirlant et al., 2004; Resnick, 1987), we prove that if the envelope distribution function has finite and non decreasing hazard rate (defined in Section III): i) an explicit formula connects the minimax redundancy and the envelope distribution; ii) the AC-code asymptotically 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 describes the AC-code. Section III 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 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.

II. 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 $i\colon 1 \le i \le n$, let $m_i = \max_{1\le j\le i} x_j$. 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 derived from the sequence of differences between records and terminated by a 1, $\widetilde{\mathbf{m}} = (\widetilde{m}_i - \widetilde{m}_{i-1} + 1)_{1\le i\le n_n^0} 1$. 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 $\widetilde{x}_i = x_i \mathbb{I}_{x_i \le m_{i-1}}$. The binary string C_M is obtained by arithmetic encoding of $\widetilde{x}_{1:n}0$.

Remark 1. *Let* $x_{1:n} \in \mathbb{N}_+^n$ *be*

$$5\,15\;8\,1\,30\;7\,1\,2\,1\,8\,4\,7\,15\;1\,5\,17\,13\;4\,12\;12\,,$$

(records are italicized) then $m_{1:n}$ *is*

$$5\,15\,15\,15\,30\,30\,30\,30\,30\,30\,30\,30\,30\,30\,30\,30\,30\,30\,30\,30$$

and $\tilde{x}_{1:n}0$ *is parsed into* 4 *substrings terminated by* 0

$$\underbrace{0}_{i}\;\underbrace{0}_{ii}\;\underbrace{8\,1\,0}_{iii}\;\underbrace{7\,1\,2\,1\,8\,4\,7\,15\,1\,5\,17\,13\,4\,12\,12\,0}_{iv}$$

while $\tilde{m}$ *is* $6\,11\;16\;1$.

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 (See Cesa-Bianchi and Lugosi, 2006, for a gentle introduction to Krichevsky-Trofimov mixtures).

Bontemps (2011) describes a nice simple way of interleaving the Elias codewords and the mixture code in order to perform *online* encoding and decoding. Substrings of $\tilde{x}_{1:n}0$ terminated by 0 are fed online to an arithmetic coder over the relevant alphabet using properly adapted Krichevsky-Trofimov mixtures, after each 0, the corresponding symbol from $\tilde{m}$ is encoded using the self-delimited Elias code and transmitted. The alphabet used to encode the next substring of $\tilde{x}_{1:n}0$ is enlarged and the procedure is iterated. The last symbol of $\tilde{m}$ is a 1, it signals the end of the message.

III. Hazard rate and envelope distribution

The envelope distributions we consider in this paper are characterized by the behavior of their hazard function $n \mapsto -\ln \overline{F}(n)$. The probabilistic analysis of the performance of the AC-code borrows tools and ideas from extreme value theory. As the theory of extremes for light-tailed discrete random variables is plagued by interesting but distracting paradoxes, following Anderson (1970), it proves convenient to define a continuous distribution function starting from the envelope distribution function F. This continuous distribution function will be called the *smoothed envelope distribution*, it coincides with the envelope distribution on $\mathbb{N}$. Its hazard function is defined by linear interpolation: the *hazard rate*, that is the derivative of the hazard function is well-defined on $\mathbb{R}_+ \setminus \mathbb{N}$ and it is piecewise constant. The envelope distributions we consider here are such that this hazard rate is non-decreasing and finite. 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.

We will also repeatedly deal with the quantile function of the envelope distribution and even more often with the quantile function of the smoothed envelope distribution. As the latter is continuous and strictly increasing over its support, the quantile function of the smooth envelope distribution is just the inverse function of the smooth envelope distribution. The quantile function of the piecewise constant envelope distribution is the left continuous generalized inverse:

$$F^{-1}(p) = \inf\{k \colon k \in \mathbb{N}, F(k) \geq p\}\,.$$

If the hazard rate is finite, then $\lim_{p\to 1} F^{-1}(p) = \infty$. Note that the smoothed envelope distribution has support $[l_f - 1, \infty)$. Recall that if X is distributed according to the smoothed envelope distribution $\lfloor X \rfloor + 1$ and $\lceil X \rceil$ are distributed according to the envelope distribution.

Remark 2. *Assume that F is a shifted geometric distribution: for some $l \in \mathbb{N}_+$, some $q \in (0,1)$, for all $k \geq l$, $\overline{F}(k) = (1-q)^{k-l}$, so that the hazard function is $(k-l)\ln 1/(1-q)$ for $k \geq l$. The corresponding smooth distribution is the shifted exponential distribution with tail function $t \mapsto (1-q)^{t-l}$ for $t > l$.*

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*

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

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

IV. Main result

In the text, we will repeatedly use the next shorthand for the quantile of order $1 - 1/t, t > 1$ of the smoothed envelope distribution. The function $U \colon [1,\infty) \to \mathbb{R}$ is defined by

$$U(t) = F_s^{-1}(1 - 1/t) \tag{1}$$

where F_s is the smoothed envelope distribution.

The main result may be phrased as follows.

Theorem 2. *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, and $U \colon [1,\infty) \to \mathbb{R}$ is defined by* (1)*, then*

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

while

$$R^+(\Lambda_f^n) \geq (1 + o_f(1))(\log e) \int_1^n \frac{U(x)}{2x} \mathrm{d}x$$

as n tends to infinity.

Remark 3. *Note that the* AC*-code is almost trivially adaptive over classes of memoryless sources with alphabet $\{1,\dots,k\}, k \in \mathbb{N}_+$: almost-surely eventually, the largest symbol in the sequence coincides with the right-end of the source distribution, and the minimaxity (up to the first order) of Krichevsky-Trofimov mixtures settles the matter.*

The following corollary provides the bridge with Bontemps's work on classes defined by exponentially decreasing envelopes (See Definition 3).

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*

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

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:

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

The elementary proof relies on the fact $t \mapsto U(\mathrm{e}^t)$ is the inverse of the hazard function of the smoothed envelope distribution $-\ln \overline{F}_s$, it 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 6 from Section V gives a lower bound for 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 $\mathbb{P}^n \in \Lambda_f^n$ 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 10 asserts that (I) is negligible with respect to $R^+(\Lambda_f^n)$ and Proposition 11 asserts that the expected value of (II) is equivalent to $R^+(\Lambda_f^n)$.

V. Minimax redundancies

A. The Haussler-Opper lower bound

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 (the minimax cumulative entropy risk in the language of Haussler and Opper) to the metric entropy of the class of marginal distributions with respect to Hellinger distance. Recall that the Hellinger distance (denoted by d_H) between two probability distributions P_1 and P_2 on $\mathbb{N}$, is defined as the ℓ_2 distance between the square roots of the corresponding probability mass functions p_1 and p_2:

$$d_H^2(P_1, P_2) = \sum_{k\in\mathbb{N}} \left(p_1(k)^{1/2} - p_2(k)^{1/2}\right)^2 .$$

The next lower bound on minimax redundancy can be extracted from Theorem 5 in (Haussler and Opper, 1997). It relies on the fact that the Bayes redundancy is never larger than the minimax redundancy.

Theorem 3. *Using notation and conventions from Section I, for any prior probability distribution π on Λ_1,*

$$R^+(\Lambda^n) \geq \mathbb{E}_{\pi_1}\left[-\log \mathbb{E}_{\pi_2} e^{-n\frac{d_H^2(P_1,P_2)}{2}}\right] .$$

where $\pi_1 = \pi_2 = \pi$ and $P_1 \sim \pi_1, P_2 \sim \pi_2$ are picked independently.

For the sake of self-reference, a rephrasing of the proof from (Haussler and Opper, 1997) is given in the Appendix.

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.

The connection between the minimax redundancy of Λ^n and the metric entropy of Λ^1 under Hellinger metric is a direct consequence of Theorem 3.

Theorem 4. *Let $\epsilon_n > 0$ be such that*

$$\frac{n\epsilon_n^2}{8} \geq \mathcal{H}_{\epsilon_n}(\Lambda)$$

then

$$R^+(\Lambda^n) \geq \log(e)\mathcal{H}_{\epsilon_n}(\Lambda) - 1 .$$

Proof of Theorem 4: Choose a prior π that is uniformly distributed over an $\epsilon/2$-separated set of maximum cardinality. Such a set has cardinality at least $\mathcal{D}_\epsilon = \mathcal{D}_\epsilon(\Lambda^1)$.

For each P_1 in the $\epsilon/2$-separated,

$$-\log \mathbb{E}_{\pi_2} \mathrm{e}^{-n\frac{d_H^2(P_1,P_2)}{2}} \geq -\log\left(\frac{1}{\mathcal{D}_\epsilon} + \frac{\mathcal{D}_\epsilon - 1}{\mathcal{D}_\epsilon}\mathrm{e}^{-n\frac{\epsilon^2}{8}}\right) .$$

Averaging over P_1 leads to

$$\begin{aligned}R^+(\Lambda^n) &\geq -\log\left(\frac{1}{\mathcal{D}_\epsilon} + \mathrm{e}^{-n\frac{\epsilon^2}{8}}\right) \\ &\geq \log e \sup_\epsilon \min\left(\mathcal{H}_\epsilon(\Lambda), \frac{n\epsilon^2}{8}\right) - 1 .\end{aligned}$$

■

Up to this point, no assumption has been made regarding the behavior of $\mathcal{H}_\epsilon(\Lambda)$ as ϵ tends to 0. Recall that a measurable function $f\colon (0,\infty) \to [0,\infty)$ is said to be *slowly varying* at infinity if for all $\kappa > 0$, $\lim_{x\to+\infty} \frac{f(\kappa x)}{f(x)} = 1$ (See Bingham et al., 1989, for a thorough treatment of regular variation).

Assuming that $x \mapsto \mathcal{H}_{1/x}(\Lambda)$ is slowly varying at infinity allows us to solve equation $\mathcal{H}_\epsilon(\Lambda) = n\epsilon^2/8$ as n tends to infinity.

Indeed, using $g(x)$ as a shorthand for $\mathcal{H}_{1/x}(\Lambda)$, we look for x satisfying $n/8 = x^2 g(x)$ or equivalently

$$1 = \frac{x}{(n/8)^{1/2}}\sqrt{g\left((n/8)^{1/2}\frac{x}{(n/8)^{1/2}}\right)} . \tag{3}$$

Assume that g is slowly varying. A Theorem by De Bruijn (See Bingham et al., 1989, Theorem 1.5.13) asserts that there exists slowly varying functions $g^\#$ such that $g^\#(x)g(xg^\#(x)) \sim 1$ as x tends to infinity. Moreover all such functions are asymptotically equivalent, they are called De Bruijn conjugates of g. If g is slowly varying, so is $\sqrt{g}$, and we may consider its De Bruijn conjugate $(\sqrt{g})^\#$.

Any sequence $(x_n)_n$ of solutions of Equation (3) is such that $x_n/\sqrt{n/8}$ is asymptotically equivalent to $\sqrt{g}^\#((n/8)^{1/2})$. Hence $\epsilon_n = 1/x_n \sim (n/8)^{-1/2}(\sqrt{g}^\#((n/8)^{1/2}))^{-1}$. We may deduce that

$$\begin{aligned}R^+(\Lambda^n) &\geq (1+o(1))\log \mathrm{e}\,\frac{n\epsilon_n^2}{8} \\ &= \log \mathrm{e}\left(\sqrt{g}^\#((n/8)^{1/2})\right)^{-2} .\end{aligned} \tag{4}$$

The computation of De Bruijn conjugates is usually a delicate topic (see again Bingham et al., 1989), but strengthening the slow variation assumption simplifies the matter. We have not been able to find a name for the next notion in the literature, although it appears in early work by Bojanic and Seneta (1971). We nickname it *very slow variation* in the paper.

Definition 4. *A continuous, non decreasing function $g\colon (0,\infty) \to [0,\infty)$ is said to be* very slowly varying *at infinity if for all $\eta \geq 0$ and $\kappa > 0$,*

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

Note that not all slowly varying functions satisfy these

conditions. For example, $x \mapsto \exp((\ln x)^\beta)$ with $\beta > 1/2$ does not (see Bingham et al., 1989).

Remark 4. *If g is very slowly varying, then for all $\alpha, \beta > 0$, the function defined by $x \mapsto (g(x^\beta))^\alpha$ is also very slowly varying.*

Bojanic and Seneta have proved that if g is a very slowly varying function, the De Bruijn conjugates of g are asymptotically equivalent to $1/g$ (See Bingham et al., 1989, Corollary 2.3.4). Hence, if g is very slowly varying, the De Bruijn conjugates of $x \mapsto \sqrt{g}(x)$ are asymptotically equivalent to $1/\sqrt{g}(x)$.

Taking advantage of the very slow variation property allows us to make lower bound (4) transparent.

Theorem 5. *(Haussler and Opper, 1997, Theorem 5) Assume that the function over $[1, \infty)$ defined by $x \mapsto \mathcal{H}_{1/x}(\Lambda^1)$ is very slowly varying, then*

$$R^+(\Lambda^n) \geq (\log e)\mathcal{H}_{n^{-1/2}}(\Lambda)\,(1 + o(1)) \quad \text{as } n \text{ tends to } +\infty.$$

In order to lower bound the minimax redundancy of source classes defined by envelope distribution with non-decreasing hazard rate, in the next section we establish that the function $t \mapsto \int_1^{t^2} \frac{U(x)}{x}\mathrm{d}x$ has the very slow variation property, then in Section V-C we check that $\int_1^{t^2} \frac{U(x)}{x}\mathrm{d}x \sim \mathcal{H}_{1/t}(\Lambda)$.

B. Slow variation and consequences

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 absolutely continuous distribution with finite and non-decreasing hazard rate. Let $U\colon [1, \infty) \to \mathbb{R}$ be defined by $U(t) = F^{-1}(1 - 1/t)$. Then*

(i) *the inverse of the hazard function, that is the function on $]0, \infty)$ defined by $t \mapsto U(\mathrm{e}^t) = F^{-1}(1 - \mathrm{e}^{-t})$ is concave.*
(ii) *The function U is slowly varying at infinity.*

The proof relies on some classical results from regular variation theory. For the sake of self-reference, we reproduce those results here, proofs can be found in (Bingham et al., 1989) or in (de Haan and Ferreira, 2006).

Theorem 1.2.6 in de Haan and Ferreira (2006) characterizes the so-called domains of attraction of Extreme Value Theory thanks to an integral representation of the hazard function $-\ln \overline{F}$. We reproduce here what concerns the Gumbel domain of attraction as the envelope distributions we deal with belong to this domain.

Theorem 6. *A distribution function F belongs to the Gumbel max-domain of attraction, if and only if there exists $t_0 < F^{-1}(1)$ and $c\colon [t_0, F^{-1}(1)) \to \mathbb{R}_+$, and a continuous function ψ such that $\lim_{t \to F^{-1}(1)} c(t) = c_* \in \mathbb{R}_+$ such that for $t \in [t_0, F^{-1}(1))$,*

$$-\ln \overline{F}(t) = -\ln c(t) + \int_{t_0}^t \frac{1}{\psi(s)}\mathrm{d}s$$

where $\lim_{t \to F^{-1}(1)} \psi'(t) = 0$ and $\lim_{t \to F^{-1}(1)} \psi(t) = 0$ when $F^{-1}(1) < \infty$.

If F belongs to the Gumbel max-domain of attraction, then $t \mapsto F^{-1}(1 - 1/t)$ is slowly varying at infinity.

Remark 5. *Under the so-called Von Mises conditions ψ may be chosen as the reciprocal of the hazard rate.*

Proof of Proposition 2: (i) As $t \mapsto F^{-1}(1 - \mathrm{e}^{-t})$ is the inverse of the hazard function $-\ln \overline{F}$, its derivative is equal to the reciprocal of the hazard rate evaluated at $F^{-1}(1 - \mathrm{e}^{-t})$. As the hazard rate is assumed to be non-decreasing, the derivative of $F^{-1}(1 - \mathrm{e}^{-t})$ is non-increasing with respect to t.
(ii) As we consider an absolutely continuous distribution, we may and do assume that using the notation of Theorem 6, $c(t)$ is constant and that ψ is the reciprocal of the hazard rate and that it is differentiable. The function ψ is a positive non increasing function, thus its derivative converges to 0 at infinity. Hence, by Theorem 6, the smoothed envelope distribution belongs to the Gumbel max-domain of attraction. This entails that $t \mapsto F^{-1}(1 - 1/t)$ is slowly varying at infinity. ∎

The next Theorem from Bojanic and Seneta (1971) can be found in (Bingham et al., 1989, Theorem 2.3.3). It asserts that if g is slowly varying and if for some $x > 0$, $g(tx)/g(t)$ converges sufficiently fast towards 1, then g is also very slowly varying.

Theorem 7. *If g varies slowly at infinity and for some $x \in \mathbb{R}_+$*

$$\lim_{t \to \infty} \left(\frac{g(tx)}{g(t)} - 1 \right) \ln g(t) = 0$$

then g is very slowly varying,

$$\lim_{t \to \infty} \left(\frac{g(t(g(t))^\kappa)}{g(t)} - 1 \right) = 1$$

locally uniformly in $\kappa \in \mathbb{R}$.

The next proposition establishes that if F is an envelope distribution with finite non-decreasing hazard rate, $U(t) = F^{-1}(1 - 1/t)$, then the function $t \mapsto \int_1^t \frac{U(x)}{x}\mathrm{d}x$ satisfies the condition of Theorem 7. As a byproduct of the proof, we show that $U(t) \ln U(t)$ is asymptotically negligible with respect to $\int_1^t \frac{U(x)}{x}\mathrm{d}x$ as t tends to infinity. Both properties are instrumental in the derivation of the main results of this paper. First the sequence $(\int_1^n \frac{U(x)}{x}\mathrm{d}x)_n$ is asymptotically equivalent to $\mathcal{H}_{1/n^{1/2}}(\Lambda_f^1)$ and thus to the lower bound on the minimax redundancy rate for Λ_f; second, by Proposition 11, it is also equivalent to the average redundancy of the mixture encoding produced by the AC-code (the negligibility of $U(t) \ln U(t)$ is used in the proof); third, the cost of the Elias encoding of censored symbols does not grow faster than $U(n)$.

Proposition 3. *Using notation from Proposition 2.*

i)

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

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

A function $g\colon \mathbb{R}_+ \to \mathbb{R}_+$ has the *extended regular variation property*, if there exists $\gamma \in \mathbb{R}$ and a measurable function

$a\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that for all $y \geq 1$

$$\lim_{t\to\infty} \frac{g(ty)-g(t)}{a(t)} = \int_1^y x^{\gamma-1}\mathrm{d}x\,.$$

If $\gamma = 0$, g has the extended slow variation property. The function a is called the auxiliary function. We refer to (Bingham et al., 1989) or (de Haan and Ferreira, 2006) for a thorough treatment of this notion. Basic results from regular variation theory assert that if g has the extended slow variation property, all possible auxiliary functions are slowly varying and that $\lim_{t\to\infty} a(t)/g(t) = 0$.

Proof of Proposition 3: To prove i), note that by concavity of $t \mapsto U(\mathrm{e}^t)$,

$$\begin{aligned}\int_1^t \frac{U(x)}{x}\,\mathrm{d}x &= \int_0^{\ln t} U(\mathrm{e}^s)\mathrm{d}s \\ &\geq \frac{\ln(t)}{2}U(t)\,.\end{aligned}$$

Plugging this upper bound leads to

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

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

ii) We first prove that $t \mapsto \int_1^t \frac{U(x)}{x}\mathrm{d}x$ has the *extended slow variation property with auxiliary function* $U(t)$, that is for all $y \geq 1$,

$$\lim_{t\to\infty} \frac{\int_1^{ty} \frac{U(x)}{x}\mathrm{d}x - \int_1^t \frac{U(x)}{x}\,\mathrm{d}x}{U(t)} = \log y\,.$$

Indeed

$$\begin{aligned}\int_1^{ty} \frac{U(x)}{x}\,\mathrm{d}x - \int_1^t \frac{U(x)}{x}\,\mathrm{d}x &= \int_1^y \frac{U(tx)}{x}\,\mathrm{d}x \\ &= U(t)\int_1^y \frac{U(tx)}{U(t)}\frac{1}{x}\mathrm{d}x\end{aligned}$$

Now the desired result follows from the Uniform Convergence Theorem from regular variation theory: if U is slowly varying, then $\frac{U(tx)}{U(t)}$ converges uniformly to 1 for $x \in [1, y]$ as t tends to infinity.

In order to invoke Theorem 7 to establish ii) it suffices to notice that

$$\begin{aligned}&\lim_{t\to\infty}\left(\frac{\int_1^{ty}\frac{U(x)}{x}\,\mathrm{d}x}{\int_1^t \frac{U(x)}{x}\,\mathrm{d}x} - 1\right)\ln\int_1^t \frac{U(x)}{x}\,\mathrm{d}x \\ &= \lim_{t\to\infty} \frac{\int_1^{ty}\frac{U(x)}{x}\,\mathrm{d}x - \int_1^t \frac{U(x)}{x}\,\mathrm{d}x}{U(t)}\frac{U(t)\ln\int_1^t \frac{U(x)}{x}\,\mathrm{d}x}{\int_1^t \frac{U(x)}{x}\,\mathrm{d}x} \\ &= \ln y \lim_{t\to\infty} \frac{U(t)\ln\int_1^t \frac{U(x)}{x}\,\mathrm{d}x}{\int_1^t \frac{U(x)}{x}\,\mathrm{d}x} \\ &\leq \ln y \lim_{t\to\infty} \frac{U(t)\ln\left(U(t)\ln(t)\right)}{\int_1^t \frac{U(x)}{x}\,\mathrm{d}x} \\ &\leq \ln y \left\{\lim_{t\to\infty} \frac{U(t)\ln U(t)}{\int_1^t \frac{U(x)}{x}\,\mathrm{d}x} + \lim_{t\to\infty} \frac{U(t)\ln\ln(t)}{\int_1^t \frac{U(x)}{x}\,\mathrm{d}x}\right\} \\ &\leq \ln y \lim_{t\to\infty} \frac{\ln\ln(t)}{2\ln t} \\ &= 0\end{aligned}$$

where the first inequality comes from the fact that U is non-decreasing, the second inequality from i), the third inequality from the first intermediate result in the proof of i).

■

C. Minimax rate

The ϵ-entropy of envelope classes defined by envelope distributions with finite and non-decreasing hazard rate is related to the behavior of the quantile function of the smoothed envelope distribution.

Proposition 4. *Let f be an envelope function. Assume that the smoothed envelope distribution F has finite and non-decreasing hazard rate, let $U(t) = F^{-1}(1-1/t)$ be the quantile of order $1-1/t$ of the smoothed envelope distribution, then*

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

as ϵ tends to 0.

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

The next lower bound for $R^+(\Lambda_f^n)$ follows from a direct application of Theorem 5 and Proposition 3:

Proposition 5. *Using notation from Proposition 4,*

$$R^+(\Lambda_f^n) \geq (1+o_f(1))(\log e)\int_1^n \frac{U(x)}{2x}\mathrm{d}x$$

as n tends to $+\infty$.

A concrete corollary follows easily.

Proposition 6. *The minimax redundancy of the sub-exponential envelope class with parameters (α, β, γ) satisfies*

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

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

$$\beta\left(\ln(\gamma t)\right)^{1/\alpha} - 1 \leq U(t) \leq \beta\left(\ln(\kappa\gamma t)\right)^{1/\alpha} - 1 \qquad (5)$$

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 $\overline{F}(k) \leq 1/t \Rightarrow k+1 \geq \beta(\ln(\gamma t))^{1/\alpha}$.

■

As observed in the introduction, Theorem 1 provides an easy upper-bound on the minimax redundancy of envelope classes, $R^+(\Lambda_f^n) \leq 2 + \log e + \frac{U(n)}{2}\log n$. For envelope classes with non-decreasing hazard rate, this upper-bound is (asymptotically) within a factor of 2 of the minimax redundancy. Indeed,

$$\int_1^n \frac{U(x)}{2x}\mathrm{d}x = \frac{1}{2}\int_0^{\ln n} U\left(\mathrm{e}^y\right)\mathrm{d}y$$

$$\geq \frac{1}{2}\frac{U(n)\ln n}{2} = \frac{U(n)\ln n}{4}\,,$$

where the inequality comes from concavity and positivity of $y \mapsto U(\mathrm{e}^y)$. Hence, by Proposition 6, for such envelope classes

$$R^+(\Lambda_f^n) \geq (1+o_f(1))\frac{U(n)\log n}{4}\,.$$

The merit of the AC-code is to asymptotically achieve the minimax lower bound while processing the message online and without knowing the precise form of the envelope. This is established in the next section.

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 $\widetilde{\mathfrak{m}}$ 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 10).

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 7. *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 $U(t)=F^{-1}(1-1/t)$ and $Y_{1,1}\leq\ldots\leq Y_{n,n}$ be the order statistics. Then,*

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

Remark 6. *If the hazard rate is strictly increasing, $Y_{n,n}-U(n)$ satisfies a law of large numbers (See Anderson, 1970). We refer to (Boucheron and Thomas, 2012) for more results about concentration inequalities for order statistics.*

Remark 7. *The upper bound on $\mathbb{E}[Y_{n,n}]$ may be tight. For example it allows to establish that the expected value of the maximum of n independent standard Gaussian random variables is less than $\sqrt{2H_n - \ln H_n - \ln\pi}$ (Boucheron and Thomas, 2012).*

The proof of proposition 7 relies on a quantile coupling argument and on a sequence of computational steps inspired by extreme value theory (de Haan and Ferreira, 2006) and concentration of measure theory (Ledoux, 2001). The proof also takes advantage of the Rényi representation of order statistics (See de Haan and Ferreira, 2006, Chapter 2). The next theorem rephrases this classical result.

Theorem 8. (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 8. *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; Steele, 1986; Massart, 2007; Boucheron et al., 2013) .

Theorem 9. (sub-additivity of variance and entropy.) *Let $X_1,\ldots,X_n$ be independent random variables and let Z 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*

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

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

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

Proof of Proposition 8: 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 7: 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.

By the Efron-Stein inequality (See Proposition 8),

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

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

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

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 8,

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

∎

When handling subexponential envelopes classes, Proposition 7 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 $Y_{1,n} \leq \ldots \leq Y_{n,n}$ denote again the order statistics of a sample $Y_1, \ldots, Y_n$ from the envelope distribution, then for any non-decreasing function g, $\mathbb{E}[g(M_n)] \leq \mathbb{E}[g(Y_{n,n})]$. Using (5) one gets the following.

Proposition 9. *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).

Remark 8. *From Proposition 9, it follows that the number of distinct symbols in $X_{1:n}$ grows at most logarithmically with n. A simple argument stated at the end of the Appendix shows that, if the hazard rate is non-decreasing, the number of distinct symbols may grow as fast as $U(n) = F^{-1}(1-1/n)$ where F is the envelope distribution. This suggests a brute force approach to the analysis of the redundancy of the* AC*-code based on the next two informal observations: the redundancy of Krichevsky-Trofimov mixtures for an alphabet with size $M_n = O_P(\log n)$, should not be larger than $\mathbb{E}\left[\frac{M_n}{2}\right] \log n$; the cost of the Elias encoding of records is negligible with respect to the redundancy of Krichevsky-Trofimov mixture. This leads to a simple estimate of redundancy : $(1 + o_f(1))\frac{U(n)}{2} \log n$ which is always larger than $\int_1^n \frac{U(x)}{2x} \mathrm{d}x$, but may be within a constant of the optimal bound. Indeed, the cautious analysis of the* AC*-code pioneered in (Bontemps, 2011) and pursued here allows us to recover the exact redundancy rate of the* AC*-code and to establish asymptotic minimaxity.*

B. Elias encoding

The average length of the Elias encoding for sources from a class with a smoothed envelope distribution F with non-decreasing hazard rate is $O(U(n))$ (where $U(t) = F^{-1}(1 - 1/t)$). 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 10. *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*

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

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, letting $M_n = \max(X_1, \ldots, X_n)$, 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 10: The length of the Elias codewords used to encode the sequence of record differences $\widetilde{\mathbf{m}}$ is readily upper-bounded:

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

for some universal constant ρ. The bound on the length of the Elias encoding follows from Proposition 9.

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 $\sum_{i=1}^{n_n^0}(\widetilde{m}_i-\widetilde{m}_{i-1})\leq M_n$, we also have

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

The average length of C_E satisfies:

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

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 $n_n^0\log(M_n)\leq R_n\log(\max(Y_1,\ldots,Y_n))$. 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

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

■

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 11. *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*

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

as n tends to infinity.

The proof of Proposition 11 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 12 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 12. (POINTWISE BOUND) *Let i_0 be the random integer defined by: $i_0=1\vee\lfloor M_n/4\rfloor$, then,*

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

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

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

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 $\{0,\dots,M_n\}$ 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},

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

The difference between the codelengths can be further upper bounded.

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

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

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

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

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

■

Proof of Proposition 11: 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 7.

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

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

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

Now using Propositions 2 and 3 and the fact that U tends to infinity at infinity one gets that

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

as n tends to infinity and the result follows.

■

APPENDIX

A. Haussler-Opper lower bound

Proof of Theorem 3: The proof of the Haussler-Opper lower bound consists of ingenuously using Fubini's theorem and Jensen's inequality.

Throughout this proof, we think of Λ^1 as a measurable parameter space denoted by Θ endowed with a probability distribution π, $\theta, \theta^*, \tilde{\theta}, \widehat{\theta}$ denote random elements of Θ picked

according to the prior π. Each P_θ define an element of Λ^n. The model is assumed to be dominated and for each $\theta \in \Theta$, $\mathrm{d}P_\theta^n$ denotes the density of P_θ^n with respect to the dominating probability. In this paragraph $\int_{\mathcal{X}^n} \dots$ should be understood as integration with to the dominating probability.

Recall that the Bayes redundancy under prior π satisfies

$$\begin{aligned}&\mathbb{E}_{\pi_1}\left[D\left(P_{\theta^*}^n, \mathbb{E}_{\pi_2} P_{\tilde{\theta}}^n\right)\right]\\ &= -\int_\Theta \mathrm{d}\pi(\theta^*) \int_{\mathcal{X}^n} \mathrm{d}P_{\theta^*}^n(x_{1:n}) \log \int_\Theta \mathrm{d}\pi(\tilde{\theta}) \frac{\mathrm{d}P_{\tilde{\theta}}^n(x_{1:n})}{\mathrm{d}P_{\theta^*}^n(x_{1:n})}\end{aligned}$$

where θ^* (resp. $\tilde{\theta}$) is distributed according to $\pi_1 = \pi$ (resp. $\pi_2 = \pi$). The next equation

$$\int_\Theta \mathrm{d}\pi(\theta^*) \int_{\mathcal{X}^n} \mathrm{d}P_{\theta^*}^n(x_{1:n}) \left(\frac{\int_\Theta \mathrm{d}\pi(\tilde{\theta}) \frac{\mathrm{d}P_{\tilde{\theta}}^n(x_{1:n})}{\mathrm{d}P_{\theta^*}^n(x_{1:n})}}{\int_\Theta \mathrm{d}\pi(\widehat{\theta}) \sqrt{\frac{\mathrm{d}P_{\widehat{\theta}}^n(x_{1:n})}{\mathrm{d}P_{\theta^*}^n(x_{1:n})}}} \right) = 1$$

is established by repeated invokation of Fubini's Theorem as follows:

$$\begin{aligned}&\int_\Theta \mathrm{d}\pi(\theta^*) \int_{\mathcal{X}^n} \mathrm{d}P_{\theta^*}^n(x_{1:n}) \left(\frac{\int_\Theta \mathrm{d}\pi(\tilde{\theta}) \frac{\mathrm{d}P_{\tilde{\theta}}^n(x_{1:n})}{\mathrm{d}P_{\theta^*}^n(x_{1:n})}}{\int_\Theta \mathrm{d}\pi(\widehat{\theta}) \sqrt{\frac{\mathrm{d}P_{\widehat{\theta}}^n(x_{1:n})}{\mathrm{d}P_{\theta^*}^n(x_{1:n})}}} \right)\\ &= \int_\Theta \mathrm{d}\pi(\theta^*) \int_{\mathcal{X}^n} \sqrt{\mathrm{d}P_{\theta^*}^n(x_{1:n})} \left(\frac{\int_\Theta \mathrm{d}\pi(\tilde{\theta}) \mathrm{d}P_{\tilde{\theta}}^n(x_{1:n})}{\int_\Theta \mathrm{d}\pi(\widehat{\theta}) \sqrt{\mathrm{d}P_{\widehat{\theta}}^n(x_{1:n})}} \right)\\ &= \int_{\mathcal{X}^n} \int_\Theta \mathrm{d}\pi(\theta^*) \sqrt{\mathrm{d}P_{\theta^*}^n(x_{1:n})} \left(\frac{\int_\Theta \mathrm{d}\pi(\tilde{\theta}) \mathrm{d}P_{\tilde{\theta}}^n(x_{1:n})}{\int_\Theta \mathrm{d}\pi(\widehat{\theta}) \sqrt{\mathrm{d}P_{\widehat{\theta}}^n(x_{1:n})}} \right)\\ &= \int_{\mathcal{X}^n} \left(\int_\Theta \mathrm{d}\pi(\tilde{\theta}) \mathrm{d}P_{\tilde{\theta}}^n(x_{1:n}) \right)\\ &= \int_\Theta \mathrm{d}\pi(\tilde{\theta}) \int_{\mathcal{X}^n} \mathrm{d}P_{\tilde{\theta}}^n(x_{1:n})\\ &= 1\,.\end{aligned}$$

Starting from the previous equation, using twice the Jensen inequality and the convexity of $-\log$ leads to

$$\begin{aligned}&-\int_\Theta \mathrm{d}\pi(\theta^*) \int_{\mathcal{X}^n} \mathrm{d}P_{\theta^*}^n(x_{1:n}) \log \int_\Theta \mathrm{d}\pi(\tilde{\theta}) \frac{\mathrm{d}P_{\tilde{\theta}}^n(x_{1:n})}{\mathrm{d}P_{\theta^*}^n(x_{1:n})}\\ &\geq -\int_\Theta \mathrm{d}\pi(\theta^*) \int_{\mathcal{X}^n} \mathrm{d}P_{\theta^*}^n(x_{1:n}) \log \int_\Theta \mathrm{d}\pi(\widehat{\theta}) \sqrt{\frac{\mathrm{d}P_{\widehat{\theta}}^n(x_{1:n})}{\mathrm{d}P_{\theta^*}^n(x_{1:n})}}\\ &\geq -\int_\Theta \mathrm{d}\pi(\theta^*) \log \int_{\mathcal{X}^n} \mathrm{d}P_{\theta^*}^n(x_{1:n}) \int_\Theta \mathrm{d}\pi(\widehat{\theta}) \sqrt{\frac{\mathrm{d}P_{\widehat{\theta}}^n(X_{1:n})}{\mathrm{d}P_{\theta^*}^n(X_{1:n})}}\,.\end{aligned}$$

In the sequel, $\alpha_H(P_{\widehat{\theta}}, P_{\theta^*})$ is a shorthand for the Hellinger affinity between $P_{\widehat{\theta}}$ and P_{θ^*}, recall that

$$\begin{aligned}\alpha_H(P_{\widehat{\theta}}, P_{\theta^*})^n &= \alpha_H(P_{\widehat{\theta}}^n, P_{\theta^*}^n)\\ &= \int_{\mathcal{X}^n} \sqrt{\mathrm{d}P_{\widehat{\theta}}^n(x_{1:n}) \mathrm{d}P_{\theta^*}^n(x_{1:n})}\end{aligned}$$

and that

$$\begin{aligned}\alpha_H(P_{\widehat{\theta}}, P_{\theta^*}) &= 1 - d_H^2(P_{\widehat{\theta}}, P_{\theta^*})\\ &\leq \exp\left(-d_H^2(P_{\widehat{\theta}}, P_{\theta^*})\right)\,.\end{aligned}$$

By Fubini's Theorem again,

$$\begin{aligned}&-\int_\Theta \mathrm{d}\pi(\theta^*) \log \int_{\mathcal{X}^n} \mathrm{d}P_{\theta^*}^n(x_{1:n}) \int_\Theta \mathrm{d}\pi(\widehat{\theta}) \sqrt{\frac{\mathrm{d}P_{\widehat{\theta}}^n(X_{1:n})}{\mathrm{d}P_{\theta^*}^n(X_{1:n})}}\\ &\geq -\int_\Theta \mathrm{d}\pi(\theta^*) \log \int_\Theta \mathrm{d}\pi(\widehat{\theta}) \int_{\mathcal{X}^n} \sqrt{\mathrm{d}P_{\widehat{\theta}}^n(X_{1:n}) \mathrm{d}P_{\theta^*}^n(X_{1:n})}\\ &= -\int_\Theta \mathrm{d}\pi(\theta^*) \log \int_\Theta \mathrm{d}\pi(\widehat{\theta}) \alpha_H(P_{\widehat{\theta}}, P_{\theta^*})^n\\ &\geq -\int_\Theta \mathrm{d}\pi(\theta^*) \log \int_\Theta \mathrm{d}\pi(\widehat{\theta}) \exp\left(-n \frac{d_H^2(P_{\widehat{\theta}}, P_{\theta^*})}{2}\right)\,.\end{aligned}$$

The right hand side can written as

$$\mathbb{E}_{\pi_1}\left[-\log \mathbb{E}_{\pi_2} \exp\left(-n \frac{d_H^2(P_1, P_2)}{2}\right)\right]$$

where P_1 (resp. P_2) is distributed according to $\pi_1 = \pi$ (resp. $\pi_2 = \pi$). The proof is terminated by recalling that, whatever the prior, the minimax redundancy is not smaller than the Bayes redundancy. ■

B. Proof of Proposition 4

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 $\ell_2^{\mathbb{N}^+}$:

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

Let $N_\epsilon = U(\frac{16}{\epsilon^2})$ (N_ϵ is the $1-\epsilon^2/16$ quantile of the smoothed 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 ($\mathrm{Vol}(B_m) = \Gamma(1/2)^m/\Gamma(m/2+1)$ 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

$$\begin{aligned}\mathcal{H}_\epsilon &\leq \sum_{k=1}^{l_f} \ln\left(\sqrt{f(k)}+\epsilon/4\right) + \sum_{k=l}^{N_\epsilon} \ln(\sqrt{f(k)}) \\ &\quad - \ln \mathrm{Vol}(B_{N_\epsilon}) + \frac{N_\epsilon - l_f}{\sqrt{1-e^{-b}}} + N_\epsilon \ln\frac{8}{\epsilon} \\ &\leq \sum_{k=l}^{N_\epsilon} \frac{1}{2}\ln\left(\frac{64\overline{F}(k-1)}{\epsilon^2}\right) - \ln \mathrm{Vol}(B_{N_\epsilon}) \\ &\quad + \frac{N_\epsilon - l_f}{\sqrt{1-e^{-b}}} + l_f \ln\frac{8}{\epsilon} + \sum_{k=1}^{l_f} \ln\left(\sqrt{f(k)}+\epsilon/4\right).\end{aligned} \tag{6}$$

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 $m = N_\epsilon - l_f$

$$\begin{aligned}\mathcal{H}_\epsilon &\geq \sum_{k=l}^{N_\epsilon} \ln\sqrt{f(k)} - \ln \mathrm{Vol}(B_{N_\epsilon - l_f}) + (N_\epsilon - l_f)\ln\frac{1}{\epsilon} \\ &\geq \sum_{k=l}^{N_\epsilon} \frac{1}{2}\ln\left(\frac{\overline{F}(k-1)(1-e^{-b})}{\epsilon^2}\right) - \ln \mathrm{Vol}(B_{N_\epsilon - l_f})\end{aligned} \tag{7}$$

Now,

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

as ϵ tends to 0. Since $N_\epsilon \to \infty$, we have also $\ln \mathrm{Vol}(B_{N_\epsilon - l_f}) = [N_\epsilon \ln N_\epsilon]\,(1+o(1))$, as ϵ tends to 0. Now, the term $\sum_{k=l}^{N_\epsilon} \frac{1}{2}\ln\left(\frac{\overline{F}(k-1)}{\epsilon^2}\right)$ n (6) and (7) is treated by (2). The desired result follows from the fact that U and hence $U\ln(U)$ are slowly varying (Proposition 2) and from Proposition 3.

C. Proof of equation (2)

Performing the change of variable $y = U(x)$ ($x = 1/\overline{F}(y)$, $\frac{\mathrm{d}x}{\mathrm{d}y} = \frac{F'(y)}{(\overline{F}(y))^2}$),

$$\begin{aligned}\int_1^t \frac{U(x)}{2x}\mathrm{d}x &= \int_{l_f-1}^{U(t)} \frac{yF'(y)}{2\overline{F}(y)}\mathrm{d}y \\ &= \left[-\frac{y}{2}\ln(\overline{F}(y))\right]_{l_f-1}^{U(t)} + \int_{l_f-1}^{U(t)} \frac{\ln(\overline{F}(y))}{2}\mathrm{d}y \\ &= \frac{U(t)}{2}\ln(t) + \int_0^{U(t)} \frac{\ln(\overline{F}(y))}{2}\mathrm{d}y \\ &= \int_0^{U(t)} \frac{\ln(t\overline{F}(x))}{2}\mathrm{d}x,\end{aligned}$$

where the second equation follows by integration by parts.

D. The number of distinct symbols in a geometric sample

The number of distinct symbols in a geometric sample has been thoroughly investigated using analytic techniques (See Archibald et al., 2006, and related papers). The next elementary argument from (Ben-Hamou, 2013) shows that the number of distinct symbols in a geometric sample is on average of the same order of magnitude as the maximum. This is in sharp constrast with what is known for samples from power-law distributions, See (Gnedin et al., 2007) for a general survey, (Ohannessian and Dahleh, 2012) and (Lugosi and Nobel, 1999) for statistical applications.

Assume that $X_1, \ldots, X_n$ are independently, identically distributed according to a geometric distribution with parameter $q \in (0,1)$, that is $\mathbb{P}\{X_1 > k\} = (1-q)^k$ for $k \in \mathbb{N}$. Let K_n denote the number of distinct symbols in $X_1, \ldots, X_n$ and $M_n = \max(X_1, \ldots, X_n)$.

$$\mathbb{E}M_n \geq \mathbb{E}K_n \geq \mathbb{E}M_n - \frac{1-q}{q^2}\,.$$

Proof: Let $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n}$ denote the non-decreasing rearrangement of $X_1, \ldots, X_n$. We agree on $X_{0,n} = 0$. The difference $M_n - K_n$ is upper-bounded by the sum of the gaps in the non-decreasing rearrangement:

$$M_n - K_n \leq \sum_{k=1}^{n} (X_{k,n} - X_{k-1,n} - 1)_+\,.$$

The expected value of $(X_{k,n} - X_{k-1,n} - 1)_+$ can be readily upper-bounded using Rényi's representation (Theorem 8). The order statistics $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n}$ are distributed like $\lfloor Y_{k,n}/\ln(1/(1-q))\rfloor$ where $Y_{1,n} \leq Y_{2,n} \leq \ldots \leq Y_{n,n}$ are the order statistics of a standard exponential sample. For $j \in \mathbb{N}_+$, the event $\lfloor Y_{k,n}/\ln(1/(1-q))\rfloor > j + \lfloor Y_{k-1,n}/\ln(1/(1-q))\rfloor$ is included in the event $Y_{k,n} > j\ln(1/(1-q)) + Y_{k-1,n}$. As $(n-k+1)(Y_{k,n} - Y_{k-1,n})$ is exponentially distributed, the last event has probability $(1-q)^{j(n-k+1)}$.

$$\begin{aligned}&\mathbb{E}[M_n - K_n] \\ &\quad \leq \sum_{k=1}^{n} \mathbb{E}[(X_{k,n} - X_{k-1,n} - 1)_+] \\ &\quad \leq \sum_{k=1}^{n}\sum_{j\in\mathbb{N}} \mathbb{P}\left\{(X_{k,n} - X_{k-1,n} - 1)_+ > j\right\} \\ &\quad \leq \sum_{k=1}^{n}\sum_{j\in\mathbb{N}_+} (1-q)^{j(n-k+1)} \\ &\quad \leq \sum_{k=1}^{n} \frac{(1-q)^k}{1-(1-q)^k} \\ &\quad \leq \sum_{k=1}^{n} \frac{(1-q)^k}{q}\,.\end{aligned}$$

■

As M_n is concentrated around $\ln n/\ln(1/(1-q))$, this simple argument reveals that for geometrically distributed samples, the number of distinct symbols is close to the largest value encountered in the sample.

This observation can be extended to the setting where the sampling distribution has finite non-decreasing hazard rate.

Proposition 13. *Assume that $X_1, \ldots, X_n$ are independently, identically distributed according to a distribution with finite non-decreasing hazard rate over $\mathbb{N} \setminus \{0\}$. Let K_n denote the number of distinct symbols in $X_1, \ldots, X_n$ and $M_n = \max(X_1, \ldots, X_n)$. Then there exists a constant κ that may depend on the sampling distribution but not on sample size n such that*

$$\mathbb{E}M_n \geq \mathbb{E}K_n \geq \mathbb{E}M_n - \kappa\,.$$

Acknowledgment

The authors wish to thank M. Thomas, M. Ohannessian and the referees for many insightful suggestions.

References

F. Abramovich, Y. Benjamini, D. Donoho, and I. Johnstone, "Adapting to unknown sparsity by controlling the false discovery rate," *Annals of Statistics*, vol. 34, no. 2, pp. 584–653, 2006.

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.

M. Archibald, A. Knopfmacher, and H. Prodinger, "The number of distinct values in a geometrically distributed sample," *European Journal of Combinatorics*, vol. 27, no. 7, pp. 1059–1081, 2006.

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.

A. Barron, L. Birgé, and P. Massart, "Risks bounds for model selection via penalization," *Probab. Theory Relat. Fields*, vol. 113, pp. 301–415, 1999.

J. Beirlant, Y. Goegebeur, J. Teugels, and J. Segers, *Statistics of extremes*, ser. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons Ltd., 2004.

A. Ben-Hamou, "On the missing mass in independent samples and random walks." Master's thesis, Université Paris-Diderot Paris 7, Paris, France, September 2013.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner, *Efficient and adaptive estimation for semiparametric models*. New York: Springer-Verlag, 1998, reprint of the 1993 original.

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

R. Bojanic and E. Seneta, "Slowly varying functions and asymptotic relations," *Journal of Mathematical Analysis and Applications*, vol. 34, no. 2, pp. 302–315, 1971.

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

S. Boucheron and M. Thomas, "Concentration inequalities for order statistics," *Electronic Communications in Probability*, vol. 17, 2012.

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

S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

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.

I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Channels*. Academic Press, 1981.

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

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

D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard, "Density estimation by wavelet thresholding," *Annals of Statistics*, vol. 24, no. 2, pp. 508–539, 1996.

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.

M. Falk, J. Husler, and R.-D. Reiss, *Law of small numbers: extremes and rare events*. Birhäuser, 2011.

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.

A. Gnedin, B. Hansen, and J. Pitman, "Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws," *Probab. Surv.*, vol. 4, pp. 146–171, 2007. [Online]. Available: http://dx.doi.org/10.1214/07-PS092

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.

G. Lugosi and A. Nobel, "Adaptive model selection using empirical complexities," *Ann. Statist.*, vol. 27, no. 6, pp. 1830–1864, 1999.

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.

N. Merhav and M. Feder, "Universal prediction," *IEEE Transactions on Information Theory*, vol. 44, pp. 2124–2147, 1998.

M. Ohannessian and M. Dahleh, "Rare probability estimation under regularly varying heavy tails." *Journal of Machine Learning Research-Proceedings Track*, vol. 23, pp. 21–1, 2012.

S. Resnick, *Extreme values, regular variation, and point processes*. New York: Springer-Verlag, 1987, vol. 4.

B. Ryabko, "A fast adaptive coding algorithm," *Problemy Peredachi Informatsii*, vol. 26, no. 4, pp. 24–37, 1990.

——, "Twice-universal coding," *Problemy Peredachi Informatsii*, vol. 20, no. 3, pp. 24–28, 1984.

B. Ryabko and F. Topsøe, "On asymptotically optimal methods of prediction and adaptive coding for Markov sources," *J. Complexity*, vol. 18, no. 1, pp. 224–241, 2002.

B. Ryabko, J. Astola, and A. Gammerman, "Adaptive coding and prediction of sources with large and infinite alphabets," *IEEE Trans. Inform. Theory*, vol. 54, no. 8, pp. 3808–3813, 2008.

J. Steele, "An Efron-Stein inequality for nonsymmetric statistics," *The Annals of Statistics*, vol. 14, no. 2, pp. 753–758, 1986.

W. Szpankowski and M. Weinberger, "Minimax pointwise redundancy for memoryless models over large alphabets," *IEEE Trans. Inform. Theory*, vol. 58, no. 7, pp. 4094–4104, 2012.

A. B. Tsybakov, *Introduction à l'estimation non-paramétrique*. Springer, 2004.

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.

X. Yang and A. Barron, "Large alphabet coding and prediction through poissonization and tilting," in *The Sixth Workshop on Information Theoretic Methods in Science and Engineering*, Tokyo, August 2013.

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