Universal Weak Variable-Length Source Coding on Countably Infinite Alphabets

Jorge F. Silva*, *Member, IEEE* and Pablo Piantanida, *Senior Member, IEEE*

Abstract

Motivated from the fact that universal source coding on countably infinite alphabets (∞-alphabets) is not feasible, this work introduces the notion of "almost lossless source coding". Analog to the *weak variable-length source coding* problem studied by Han (IEEE TIT, 2000, 46, 1217-1226), almost lossless source coding aims at relaxing the lossless block-wise assumption to allow an average per-letter distortion that vanishes asymptotically as the block-length tends to infinity. In this setup, we show on one hand that Shannon entropy characterizes the minimum achievable rate (similarly to the case of finite alphabet sources) while on the other that almost lossless universal source coding becomes feasible for the family of finite-entropy stationary memoryless sources with ∞-alphabets. Furthermore, we study a stronger notion of almost lossless universality that demands uniform convergence of the average per-letter distortion to zero, where we establish a necessary and sufficient condition for the so-called family of "envelope distributions" to achieve it. Remarkably, this condition is the same necessary and sufficient condition needed for the existence of a strongly minimax (lossless) universal source code for the family of envelope distributions. Finally, we show that an almost lossless coding scheme offers faster rate of convergence for the (minimax) redundancy compared to the well-known information radius developed for the lossless case at the expense of tolerating a non-zero distortion that vanishes to zero as the block-length grows. This shows that even when lossless universality is feasible, an almost lossless scheme can offer different regimes on the rates of convergence of the (worst case) redundancy versus the (worst case) distortion.

Index Terms

Universal source coding, countably infinite alphabets (∞-alphabets), weak source coding, envelope distributions, information radius (i-radius), metric entropy analysis.

I. Introduction

The problem of Universal Source Coding (USC) has a long history on information theory [3]–[7]. This topic started with the seminal work of Davisson [6] that formalizes the variable-length lossless coding and introduces relevant information quantities (mutual information and channel capacity [4]). In lossless variable-length source coding, it is well-known that if we know the statistics of a source (memoryless or stationary and ergodic) the Shannon entropy (or Shannon entropy rate) provides the minimum achievable rate [4]. However, when the statistics of the source is not known but it belongs to family of distributions Λ, then the problem reduces to characterize the worst-case expected overhead (or *worse-case redundancy*) that a pair of encoder and decoder experiences due to the lack of knowledge about true distribution governing the source samples to be encoded [3], [8].

A seminal information-theoretic result states that the least worst-case overhead (or minimax redundancy of Λ) is fully characterized by the *information radius* of Λ [3]. The information radius (i-radius) has been richly studied by the community and there are numerous contributions [9]–[13], including applications to universal prediction of individual sequences [14]. In particular, it is well-known that the i-radius growths sub-linearly for the family of finite alphabet stationary and memoryless sources [3], which implies the existence of an universal source code that achieves Shannon entropy for every distribution in this family provided that the block length tends to infinity. What is intriguing in this positive result obtained for finite alphabet memoryless sources is that it does not longer extend to the case of stationary and memoryless sources on *countably infinite alphabets* (∞-alphabets), as was clearly shown in [5], [7], [9]. From an information complexity perspective, this infeasibility result implies that the i-radius of this family is unbounded for any finite block-length and, consequently, lossless universal source coding for ∞-alphabet stationary and memoryless sources is an intractable problem. In this regard, the proof presented by Györfi *et al.* [5, Theorem 1] is constructed over a connection between variable-length prefix-free codes and distribution estimators, and the fact that the redundancy of a given code upper bounds the expected divergence between the true distribution and the induced (through the code) estimate of the distribution. Then, the existence of an universal source code implies the existence of an universal estimator in the sense of expected information divergence [15].[1] The impossibility of achieving this learning objective for the family of finite entropy memoryless sources [5, Theorem 2] motives the main question addressed in this work that is, the study of a "weak notion" of universal variable-length source coding.

[1]Distribution estimator consistent in information divergence and reversed order information divergence were introduced by Barron *et al.* in [15]. The connection between distribution estimation consistent in expected information divergence and universal source coding can be found in [5] and [15, Sec. III.A].

In this framework, we propose to address the problem of universal source coding for ∞-alphabet stationary and memoryless sources by studying a weaker (lossy) notion of coding instead of the classical lossless definition [3], [4]. This notion borrows ideas from the seminal work by Han [16] that allows reconstruction errors but assuming known statistic. In this paper, we investigate the idea of relaxing the lossless block-wise assumption with the goal that the corresponding weak universal source coding formulation will be reduced to a learning criterion that becomes feasible for the whole family finite entropy stationary and memoryless sources on countably infinite alphabets. In particular, we move from lossless coding to an asymptotic vanishing distortion fidelity criterion based on the *Hamming distance* as a fidelity metric.

A. Contributions

Assuming that the distribution of the source is known, we first introduce the problem of "almost lossless source coding" for memoryless sources defined on countably infinite alphabets. Theorem 3 shows that *Shannon entropy* characterizes the minimum achievable rate for this problem. The proof of this theorem adopts a result from Ho *et al.* [17] that provides a closed-form expression for the *rate-distortion function* $R_\mu(d)$ on ∞-alphabets. From this characterization, we show that $\lim_{d \to 0} R_\mu(d) = \mathcal{H}(\mu)$ which is essential to prove this result[2].

Then, we address the problem of almost lossless universal source coding. The main difficulty arises in finding a lossy coding scheme that achieves asymptotically zero distortion, i.e., point-wise over the family, while guaranteeing that the worst-case average redundancy –w.r.t. the minimum achievable rate– vanishes with the block-length [3]. The proof of existence of an universal code with the desired property relies on a two-stage coding scheme that first quantizes (symbol-by-symbol) the ∞-alphabet and then applies a lossless variable-length code over the resulting quantized symbols. Our main result, stated in Theorem 4, shows that almost lossless universal source coding is feasible for the family of finite entropy stationary and memoryless sources.

We further study the possibility of obtaining rates of convergence for the worst-case distortion and the worst-case redundancy. To this end, we restrict our analysis to the family of stationary and memoryless sources with 1D-densities dominated by an envelope function f, which was previously studied in [9], [10], [20], [21]. Theorem 5 presents a necessary and sufficient condition on f to achieve an uniform convergence (over the family) of the distortion to zero and, simultaneously, a vanishing worst-case average redundancy. Remarkably, this condition (f being a summable function) is the same necessary and sufficient condition needed for the existence of a strongly minimax (lossless) universal source code [9, Theorems 3 and 4].

Finally, we provide an analysis of the potential benefit of an almost lossless two-stage coding scheme by exploring the family of envelope distributions that admits strong minimax universality in lossless source coding [8], [9]. In this context, Theorem 6 shows that we can have an almost lossless approach that offers a non-trivial reduction to the rate of convergence of the worst-case redundancy, with respect to the well-known i-radius developed for the lossless case, at the expense of tolerating a non-zero distortion that vanishes with the blocklength. This result provides evidence that even in the case where lossless universality is feasible, an almost lossless scheme can reduce the rate of the worst-case redundancy and consequently, it offers ways of achieving different regimes for the rate of convergence of the redundancy versus the distortion. The proof of this result uses advanced tools by Haussler and Opper [11] to relate the minimax redundancy of a family of distributions with its metric entropy with respect to the *Hellinger distance*. Indeed, this metric entropy approach has shown to be instrumental to derive tight bounds on the i-radius for summable envelope distributions in [10]. We extended this metric entropy approach to our almost lossless coding setting with a two-stage coding scheme to characterize the precise regime in which we can achieve gains in the rate of convergence of the redundancy.

B. Organization of the Paper

The rest of the paper is organized as follows. Section II introduces some definitions and preliminary results. Section III introduces our main weak source coding problem and shows that Shannon entropy is the minimum achievable rate provided that the statistics of the source is known. Section IV presents the problem of almost lossless universal source coding and proves its feasibility for the family of finite entropy memoryless distributions on ∞-alphabets. Section V elaborates a result for a stronger notion of almost lossless universality, and Section VI studies the gains in the rate of convergence of the minimax-redundancy that can be obtained with an almost lossless scheme for families of distributions that admit lossless USC. Finally, Section VII concludes with a summary of the work. The proofs of the main results of this paper are presented in Section VIII, while some supporting results are relegated to the Appendix section.

C. Basic Notation

The following notations and conventions are used throughout this article: $(x_n)_n$ will denote an infinite dimensional sequence in $\mathbb{R}^{\mathbb{N}}$; $(x_n)_n \ll (y_n)_n$ or, alternatively, $(x_n)_n$ being $o(y_n)$, means that $\lim_{n \to \infty} \frac{x_n}{y_n} = 0$; $(x_n)_n$ being $o(1)$ means that $\lim_{n \to \infty} x_n = 0$; $(x_n)_n \leq (y_n)_n$ means that $x_n \leq y_n$ for all $n \geq 1$; $(x_n)_n \leq (y_n)_n$ eventually in n means that there exists

---

[2]This result is well-known for finite alphabets, however the extension on countably infinite alphabets is not straightforward due to the discontinuity of the entropy [18], [19].

$N > 1$ such that $x_n \leq y_n$ for all $n \geq N$; $(x_n)_n \sim (y_n)_n$ means that $\lim_{n \to \infty} \frac{x_n}{y_n} = 1$; $\mathcal{X}$ will denote the countably infinite alphabet and $\mathcal{P}(\mathcal{X})$ the collection of probability measures in $\mathcal{X}$; for any function $f : \mathcal{X} \longrightarrow \mathbb{R}$, f is said to be summable (denoted by $f \in \ell_1(\mathcal{X})$) if $\sum_{n \in \mathcal{X}} |f(x)| < \infty$; and for $\mu, v \in \mathcal{P}(\mathcal{X})$, $\mu \ll v$ means that if $v(B) = 0$ then $\mu(B) = 0$ for any $B \subset \mathcal{X}$.

II. PRELIMINARIES

This section introduces some useful concepts, definitions and results that will be needed across the paper. Let $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$ be a stationary and memoryless process (or i.i.d. source) with values in a countably infinite alphabet (∞-alphabets) $\mathcal{X}$ equipped with a probability measure μ defined on the measurable space $(\mathcal{X}, \mathcal{B}(X))$[3]. Let $X^n = (X_1, \ldots, X_n)$ denote a finite block of length n of the process following the product measure μ^n on $(\mathcal{X}^n, \mathcal{B}(X^n))$[4]. Let us denote by $\mathcal{P}(\mathcal{X})$ the family of probability measures in $(\mathcal{X}, \mathcal{B}(X))$, where for every $\mu \in \mathcal{P}(\mathcal{X})$, we understand $f_\mu(x) := \frac{d\mu}{d\lambda}(x) = \mu(\{x\})$ to be a short-hand for its probability mass function (pmf). Let $\text{supp}(f) = \{x \in \mathcal{X} : |f(x)| > 0\}$ and let $\mathcal{P}_{\mathcal{H}}(\mathcal{X}) := \{\mu : H(\mu) < \infty\} \subset \mathcal{P}(\mathcal{X})$ denote the collection of finite Shannon entropy probabilities [22] where

$$H(\mu) = - \sum_{x \in \mathcal{X}} f_\mu(x) \log f_\mu(x), \tag{1}$$

with $\log$ function on base 2.

Given an i.i.d. source $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$ with distribution $\mu \in \mathcal{P}(\mathcal{X})$, let us consider a (variable length) lossless code f_n of length n as a prefix-free mapping from $\mathcal{X}^n$ to finite sequences of symbols in $\{0, 1\}$ [4]. It is well-known that $\mathbb{E}_{X^n} \{\mathcal{L}(f_n(X^n))\} \geq H(\mu^n)$ [4], where $\mathcal{L}(\cdot)$ indicates the functional that returns the length of binary sequences in $\{0, 1\}^\star := \cup_{k \geq 1} \{0, 1\}^k$. Then, the average length (in bits) used to encode X^n with f_n can be measured relative to $H(\mu^n)$, which motivates the introduction of the average redundancy (or redundancy) of f_n by the expression: $\mathbb{E}_{X^n} \{\mathcal{L}(f_n(X^n))\} - H(\mu^n)$. When μ is known, the *Huffman code* uses that information to offer an optimal prefix-free mapping (minimizing the average code-length) whose redundancy is upper bound by 1 [4].

A. Strong Minimax Universality, Information Radius and Envelope Families

In universal source coding (USC), we need to encode a stationary memoryless source $\mathbf{X}$ with an unknown probability μ that belongs to a class of models $\Lambda \subset \mathcal{P}(\mathcal{X})$. In this context, a natural performance for a prefix-free encoder $f_n : \mathcal{X}^n \to \{0, 1\}^\star$ is the worse-case (over the family Λ) redundancy expressed by:

$$R(f_n, \mu^n) := \sup_{\mu^n \in \Lambda^n} \left(\mathbb{E}_{X^n \sim \mu^n} \{\mathcal{L}(f_n(X^n))\} - H(\mu^n) \right),$$

where $\Lambda^n := \{\mu^n, \mu \in \Lambda\} \subset \mathcal{P}(\mathcal{X}^n)$ is a short-hand for the family of n-fold (product) distributions induced by Λ. This worse-case performance indicator motivates the adoption of the minimax design principle: $\min_{f_n} R(f_n, \mu^n)$ frequently used in USC [3], where the optimization is carried over the family of prefix-free codes. Importantly, there is a well-documented correspondence between prefix-free codes for $\mathcal{X}^n$ and probabilities in $\mathcal{P}(\mathcal{X}^n)$ [4]. In fact, the *Kraft-MacMillan inequality* defines a probability in $\mathcal{P}(\mathcal{X}^n)$ from a prefix-free code of length n [4], and conversely arithmetic coding provides a prefix-free code for $\mathcal{X}^n$ from a probability $v \in \mathcal{P}(\mathcal{X}^n)$, where the length of this code (in bits) is $\lceil - \log v(x^n) \rceil + 1$ for any $x^n \in \mathcal{X}^n$ [4], [23]. Then the minimax redundancy problem for USC reduces to the solution of the i-radius problem [3][5]:

$$R^+(\Lambda^n) := \inf_{v \in \mathcal{P}(\mathcal{X}^n)} \sup_{\mu^n \in \Lambda^n} \mathcal{D}(\mu^n | v) \tag{2}$$

and

$$\mathcal{D}(\mu^n | v) = \sum_{x^n \in \mathcal{X}^n} f_{\mu^n}(x^n) \log \frac{f_{\mu^n}(x^n)}{f_v(x^n)} \tag{3}$$

is the divergence of μ^n with respect to v [3], [4], [24]. Again using this connection between codes and distributions, a class $\Lambda \subset \mathcal{P}(\mathcal{X})$ of i.i.d. sources will be said to be *weakly universal* if there is a sequence of probabilities $(v_n)_n$ (where $v_n \in \mathcal{P}(X^n)$ for all n) such that $\sup_{\mu \in \Lambda} \lim_{n \to \infty} \frac{1}{n} \mathcal{D}(\mu^n | v_n) = 0$, and it will be *strongly universal* (or strongly minimax universal) if $\lim_{n \to \infty} \sup_{\mu \in \Lambda} \frac{1}{n} \mathcal{D}(\mu^n | v_n) = 0$. For the last stringent USC objective, the minimax redundancy sequence $(R^+(\Lambda^n))_n$ of Λ in (2) determines if the family is strongly minimax universal [3], [8]. For ∞-alphabets i.i.d. sources, it is known that $R^+(\mathcal{P}(X)^n) = \infty$ and, furthermore, weak minimax universality is not feasible [5], [7], [9]. This motivates the study of strong minimax universality over sub-collections of distributions whose 1D densities are dominated by an envelop function [9], [10], [20]:

---

[3]$\mathcal{B}(\mathcal{X})$ denotes the power set of $\mathcal{X}$.

[4]The product measure satisfies the memoryless condition for all $\mathcal{B}_1 \times \cdots \times \mathcal{B}_n \in \mathcal{X}^n$ then $\mu^n(\mathcal{B}_1 \times \cdots \times \mathcal{B}_n) = \mu(\mathcal{B}_1) \cdots \mu(\mathcal{B}_n)$.

[5]In fact, it follows that $R^+(\Lambda^n) + 2 \geq \min_{f_n} R(f_n, \mu^n) \geq R^+(\Lambda^n)$ [4].

Definition 1: Given a non-negative function $f : \mathcal{X} \longrightarrow \mathbb{R}^+$, the *envelope family* indexed by f is given by:

$$\Lambda_f := \left\{ \mu \in \mathcal{P}(\mathcal{X}) : f_\mu(x) \leq f(x), \quad \text{for} \quad x \in \mathcal{X} \right\}. \tag{4}$$

The next result by Boucheron *et al.* [9] establishes a necessary and sufficient condition to make Λ_f strongly minimax universal:

Theorem 1: [9, Ths. 3 and 4] Let us consider $f : \mathcal{X} \to \mathbb{R}^+$, with $\mathcal{X}$ an ∞-alphabet, and the family of i.i.d. envelope distributions $\left\{ \Lambda_f^n, n \geq 1 \right\}$. It follows that:

i) If f is summable, i.e., $f \in \ell_1(\mathcal{X})$, then $R^+(\Lambda_f^n) < \infty$ for all $n \geq 1$, and furthermore $(R^+(\Lambda_f^n))_n$ is $o(n)$.

ii) Otherwise, $R^+(\Lambda_f^n) = \infty$ for all $n \geq 1$.

Therefore for any envelope family in an ∞-alphabet, either it is strongly minimax universal (i.e.,$(R^+(\Lambda_f^n))_n$ is $o(n)$) or its i-radius in (2) is unbounded for any finite block-length. The last unbounded scenario means that for any n and a ny prefix-free code f_n there is a distribution μ in the family where the average redundancy of f_n (with respect to the entropy $H(\mu^n)$) is unbounded. Furthermore for a summable envelope function f, a series of relevant results stipulate the way $(R^+(\Lambda_f^n)/n)_n$ tends to zero function of specific tail attributes of f [9], [10], [20]. We select a result here that will be important for our exposition, for which some definitions are needed:

Definition 2: (Bontemps *et al.* [10]) For a non-negative envelope function $f \in \ell_1(\mathcal{X})$ with $|\text{supp}(f)| = \infty$, we can determine

$$l_f := \max \left\{ k : \sum_{j \geq k} f(j) \geq 1 \right\}$$

and the associated envelope probability $\mu_f \in \Lambda_f$ by:

$$\mu_f(\{k\}) := \begin{cases} 0, & \text{for} \quad k < l_f \\ f(k), & \text{for} \quad k > l_f \\ 1 - \sum_{j > l_f} f(j), & \text{for} \quad k = l_f. \end{cases} \tag{5}$$

Definition 3: (Bontemps *et al.* [10]) If we denote by $F_f(u) := \mu_f(\{1, \ldots, u\})$ the envelope distribution and by $\bar{F}_f(u) := 1 - F_f(u)$ the tail function of f (for all $u \geq 1$), we can define the quantile of order $\frac{1}{n}$ of μ_f as the solutions of [10]:

$$u_f^*(n) := \min \left\{ u \geq 1 : \bar{F}_f(u) < \frac{1}{n} \right\} \quad \text{for all } n \geq 1. \tag{6}$$

Theorem 2: [9, Th. 4]& [10, Th. 2, Prop. 5] Let us consider the envelope family $\left\{ \Lambda_f^n, n \geq 1 \right\}$ in Def. 1 with $f \in \ell_1(\mathcal{X})$. Then there is a sequence $(\xi_n)_n$ being $o(1)$ such that

$$(1 + \xi_n) \frac{(u_f^*(n) - 1)}{4} \log n \leq R^+(\Lambda_f^n) \leq 2 + \log e + \frac{(u_f^*(n) - 1)}{2} \log n$$

holds eventually with n.

It has been shown that when $f \in \ell_1(\mathcal{X})$ then $(u_f^*(n) \log n)_n$ is $o(n)$ [10], therefore Theorem 2 is consistent with Theorem 1. Importantly, $(u_f^*(n))_n$ captures the complexity of the envelop family by determining the worse-case redundancy (and its velocity of convergence to zero with n) that an optimal universal code can achieve in compressing (losslessly) a stationary and memoryless source in this family.

III. Almost Lossless Source Coding

We now introduce the notion of a lossy variable-length coding of n source symbols, which consists of a pair (f_n, g_n) where $f_n : \mathcal{X}^n \longrightarrow \{0,1\}^\star$ is a prefix free variable-length code (encoder) [4] and $g_n : \{0,1\}^\star \longrightarrow \mathcal{X}^n$ is the inverse mapping from bits to source symbols (decoder). Inspired by the weak coding setting introduced by Han [16], the possibility that $\{x^n : g_n(f_n(x^n)) \neq x^n\} \neq \emptyset$ is allowed. In order to quantify the loss induced by this encoding process, a per letter distortion measure characterization $\rho : \mathcal{X} \times \mathcal{X} :\longrightarrow \mathbb{R}^+$ is considered [25], [26], where for $x^n, y^n \in \mathcal{X}^n$ the distortion is given by

$$\rho_n(x^n, y^n) := \frac{1}{n} \sum_{i=1}^n \rho(x_i, y_i). \tag{7}$$

Given an information source $\mathbf{X} = \{X_i\}_{i=1}^n$, the average distortion induced by the pair (f_n, g_n) is

$$d(f_n, g_n, \mu^n) := \mathbb{E}_{X^n \sim \mu^n} \left\{ \rho_n \left(X^n, g_n(f_n(X^n)) \right) \right\}. \tag{8}$$

For the rest of the paper, we will focus on the special case where $\rho(x, y) = \mathbf{1}_{\{x \neq y\}}$. Then, $\rho_n(x^n, y^n)$ is the normalized *Hamming distance* between the sequences (x^n, y^n). On the other hand, the rate of the pair (f_n, g_n) (in bits per sample) is

$$r(f_n, \mu^n) := \frac{1}{n} \mathbb{E}_{X^n \sim \mu^n} \left\{ \mathcal{L}(f_n(X^n)) \right\}. \tag{9}$$

At this stage, we can introduce the almost-lossless source coding problem and with this, the standard notion of minimum achievable rate.

Definition 4 (Achievability): Given an information source $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$, we say that a rate $R > 0$ is achievable for almost-losslessly encoding $\mathbf{X}$, i.e., with zero asymptotic distortion, if there exists a sequence of encoder and decoder mappings $\{(f_n, g_n)\}_{n \geq 1}$ satisfying:

$$\limsup_{n \longrightarrow \infty} r(f_n, \mu^n) \leq R, \tag{10}$$

$$\lim_{n \longrightarrow \infty} d(f_n, g_n, \mu^n) = 0. \tag{11}$$

The minimum achievable rate is then defined as:

$$R_{al}(\mathbf{X}) := \min \{R : R \text{ is achievable for } \mathbf{X}\}. \tag{12}$$

Let $R_{al}(\mu)$ denotes the minimum achievable rate of a stationary and memoryless source driven by $\mu \in \mathcal{P}(\mathcal{X})$. The next theorem characterizes $R_{al}(\mu)$ provided that the source statistics is known.

Theorem 3 (Known statistics): Given a stationary and memoryless source on a ∞-alphabet driven by the probability measure $\mu \in \mathcal{P}_{\mathcal{H}}(\mathcal{X})$, it follows that $R_{al}(\mu) = H(\mu)$.

The proof is presented in Section VIII-A.

As it is expected, Shannon entropy characterizes the minimum achievable rate for the almost lossless source coding problem formulated in Definition 4. In the proof of Theorem 3, we adopt a result from Ho *et al.* [17] that provides a closed-form expression for the rate-distortion function $R_\mu(d)$ of μ on ∞-alphabet through a tight upper bound on the conditional entropy for a given minimal error probability [17, Theorem 1]. From this characterization, we show that $\lim_{d \to 0} R_\mu(d) = H(\mu)$, which is essential to show the result[6].

A. A Two-Stage Source Coding Scheme

In this section, we consider a two-stage source coding scheme that first applies a lossy (symbol-wise) reduction of the alphabet, and second a variable-length lossless source code over the restricted alphabet. Let us define the finite set $\Gamma_k := \{1, \ldots, k\}$. We say that a *two-stage lossy code* of block-length n and size k is the composition of: a lossy mapping of the alphabet, represented by a pair of functions (ϕ_n, ψ_n), where $\phi_n : \mathcal{X} \longrightarrow \Gamma_k$ and $\psi_n : \Gamma_k \longrightarrow \mathcal{X}$, and a fixed to variable-length prefix-free pair of lossless encoder-decoder $(\mathcal{C}_n, \mathcal{D}_n)$, where $\mathcal{C}_n : \Gamma_k^n \longrightarrow \{0, 1\}^\star$ and $\mathcal{D}_n : \{0, 1\}^\star \longrightarrow \Gamma_k^n$.

Given a source $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$ and an (n, k_n)-lossy source encode $(\phi_n, \psi_n, \mathcal{C}_n, \mathcal{D}_n)$[7], the lossy encoding of $\mathbf{X}$ induced by $(\phi_n, \psi_n, \mathcal{C}_n)$ is a two-stage process where first a quantization of size k_n over $\mathcal{X}^n$ is made (letter-by-letter) to generate a finite alphabet random sequence $Y^n := (\phi_n(X_1), \ldots, \phi_n(X_n))$ and then, a variable-length coding is applied to produce $\mathcal{C}_n(Y^n)$. Associated to the pair (ϕ_n, ψ_n), there is an induced partition of $\mathcal{X}$ given by:

$$\pi_n := \{\mathcal{A}_{n,i} := \phi_n^{-1}(\{i\}) \,|\, i \in \Gamma_{k_n}\} \subset \mathcal{B}(\mathcal{X}), \tag{13}$$

and a collection of prototypes[8] $\{y_{n,i} := \psi_n(i) \in \mathcal{A}_{n,i} \,|\, i \in \Gamma_{k_n}\} \subset \mathcal{X}$. The resulting distortion incurred by this code is given by

$$d(\phi_n, \psi_n, \mu^n) := \mathbb{E}_{X^n \sim \mu^n}\{\rho_n(X^n, \Psi_n(\Phi_n(X^n)))\}, \tag{14}$$

where $\hat{X}^n = \Psi_n(\Phi_n(X^n))$ is a short-hand to denote $(\psi_n(\phi_n(X_1)), \ldots, \psi_n(\phi_n(X_n)))$. On the other hand, the coding rate is:

$$r(\phi_n, \mathcal{C}_n, \mu^n) := \frac{1}{n} \mathbb{E}_{X^n \sim \mu^n}\{\mathcal{L}(\mathcal{C}_n(\Phi_n(X^n)))\}, \tag{15}$$

with $\Phi_n(X^n)$ denoting $(\phi_n(X_1), \ldots, \phi_n(X_n))$. An illustration of this two-stage process is presented in Figure 1.

At this point, it is worth mentioning some basic properties on the partitions induced by $(\phi_n, \psi_n, \mathcal{C}_n)$ on $\mathcal{X}$.

Definition 5: A sequence of partitions $\{\pi_n\}_{n \geq 1}$ of $\mathcal{X}$ is said to be asymptotically sufficient with respect to $\mu \in \mathcal{P}(\mathcal{X})$, if for all $x \in \text{supp}(\mu)$

$$\lim_{n \to \infty} \pi_n(x) = \{x\}, \ \mu\text{-almost everywhere}, \tag{16}$$

where $\pi_n(x) \subset \mathcal{X}$ denotes the cell in π that contains x and the almost-sure limit with respect to μ stated in (16) refers to the condition:

$$\mu\left(\limsup_{n \to \infty} \pi_n(x) \setminus \{x\}\right) = 0,$$

which is equivalent to $\lim_{n \to \infty} \mu\left(\bigcup_{k \geq n} \pi_k(x) \setminus \{x\}\right) = 0$.

---

[6]Theorem 3 is well-known for finite alphabets stationary memoryless sources, however its extension to countably infinite alphabets is not straightforward due to the discontinuity of the entropy. The interested reader may be refer to [18], [19], [27] for further details.

[7]For brevity, the decoding function $\mathcal{D}_n : \{0, 1\}^\star \longrightarrow \Gamma_k^n$ will be omitted and considered implicit in the rest of the exposition.

[8]Without loss of generality, we assume that $y_{n,i} \in \mathcal{A}_{n,i}$.

Encoding Process

first stage | second stage

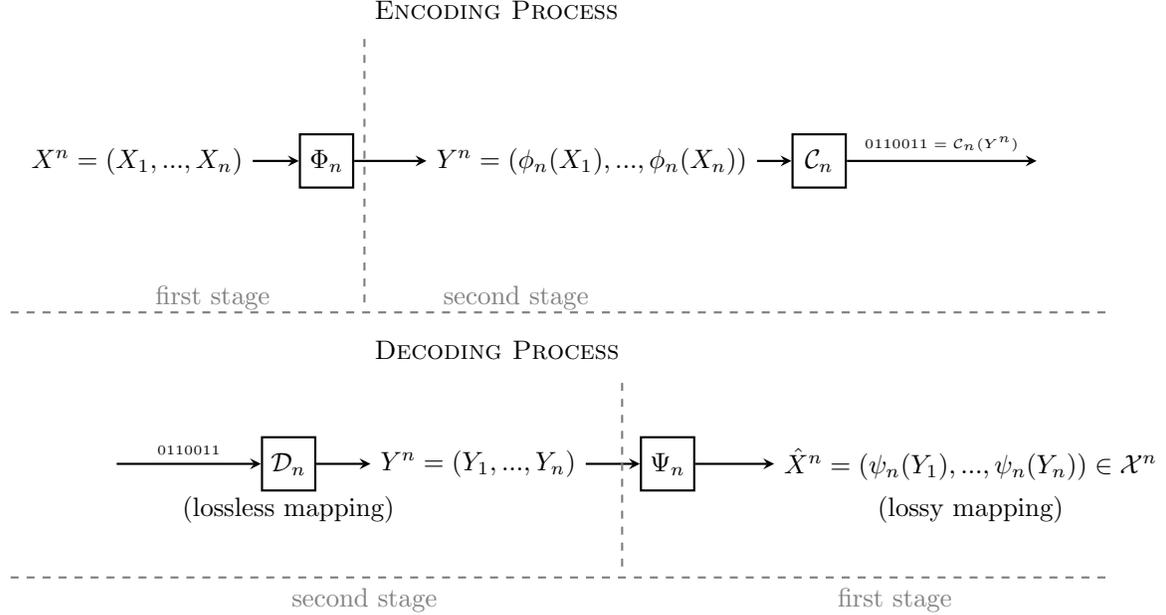Decoding Process

second stage | first stage

Fig. 1: Illustration of the two-stage lossy coding scheme $(\phi_n, \psi_n, \mathcal{C}_n, \mathcal{D}_n)$ presented in Section III-A.

Consider now almost lossless coding for which we can state the following.

Lemma 1: Let $\mathbf{X}$ be a stationary and memoryless source driven by μ. A necessary and sufficient condition for $\{(\phi_n, \psi_n, \mathcal{C}_n) : n \geq 1\}$ to have that $\lim_{n \to \infty} d(\phi_n, \psi_n, \mu^n) = 0$ is that $\{\pi_n\}_{n \geq 1}$ in (13) is asymptotically sufficient for μ.

The proof of Lemma 1 is presented in Appendix I-A.

Studying the minimum achievable rate for zero-distortion coding requires the following definition.

Definition 6: For $\mu \in \mathcal{P}_\mathcal{H}(\mathcal{X})$ and a partition π of $\mathcal{X}$, the entropy of μ restricted to the sigma-field induced by π, which is denoted by $\sigma(\pi)$, is given by

$$H_{\sigma(\pi)}(\mu) := -\sum_{\mathcal{A} \in \pi} \mu(\mathcal{A}) \log \mu(\mathcal{A}). \tag{17}$$

A basic inequality [3], [4] shows that if $\sigma(\pi) \subset \sigma(\bar{\pi})$, then $H_{\sigma(\pi)}(\mu) \leq H_{\sigma(\bar{\pi})}(\mu)$ for every μ. In particular, $H_{\sigma(\pi)}(\mu) \leq H(\mu)$, where it is simple to show that $H(\mu) = \sup_{\pi \in \Pi(\mathcal{X})} H_{\sigma(\pi)}(\mu)$ with $\Pi(\mathcal{X})$ representing the collection of finite partitions of $\mathcal{X}$. Furthermore, it is possible to state the following result.

Lemma 2: If a sequence of partitions $\{\pi_n\}_{n \geq 1}$ of $\mathcal{X}$ is asymptotically sufficient with respect to μ (Def. 5), then

$$\lim_{n \longrightarrow \infty} H_{\sigma(\pi_n)}(\mu) = H(\mu). \tag{18}$$

The proof of this result is presented in Appendix I-B.

This implies that if a two-stage scheme $\{(\phi_n, \psi_n, \mathcal{C}_n) : n \geq 1\}$ achieves zero distortion, then

$$\left\{ \pi_n = \left\{ \phi_n^{-1}(\{i\}) : i \in \Gamma_k \right\} : n \geq 1 \right\} \tag{19}$$

is asymptotically sufficient for μ (cf. Lemma 1). From the well-known result in lossless variable-length source coding [4], we have that:

$$\begin{aligned} r(\phi_n, \mathcal{C}_n, \mu^n) &\geq \frac{1}{n} H\left(\Phi_n(X^n)\right) \\ &= H_{\sigma(\pi_n)}(\mu) \end{aligned} \tag{20}$$

and consequently, Lemma 2 implies that $\liminf_{n \to \infty} r(\phi_n, \mathcal{C}_n, \mu^n) \geq H(\mu)$. Hence, letting $\bar{R}_{al}(\mu)$ to be the minimum achievable rate w.r.t. the family of two-stage lossy schemes in Definition 4, we obtain that $\bar{R}_{al}(\mu) \geq R_{al}(\mu) = H(\mu)$.

The next result shows that there is no additional overhead (in terms of bits per sample), if we restrict the problem to the family of two-stage lossy schemes.

Proposition 1: For a stationary and memoryless source $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$ driven by $\mu \in \mathcal{P}_\mathcal{H}(\mathcal{X})$,

$$\bar{R}_{al}(\mu) = R_{al}(\mu) = H(\mu).$$

The proof is presented in Appendix I-C.

IV. Universal Almost Lossless Source Coding

Consider a stationary and memoryless source $\{X_n\}_{n=1}^{\infty}$ on a ∞-alphabet with unknown distribution but belonging to a family $\Lambda \subset \mathcal{P}(\mathcal{X})$. The main question to address here is if there exists a lossy coding scheme whose rate achieves the minimum feasible rate in Theorem 3, for every possible distribution in Λ, while the distortion goes to zero as the block-length tends to infinity as defined below.

Definition 7: A family of distribution Λ is said to admit an almost lossless USC scheme, if there is a lossy source code $\{(f_n, g_n)\}_{n \geq 1}$ simultaneously satisfying:

$$\sup_{\mu \in \Lambda} \lim_{n \to \infty} d(f_n, g_n, \mu^n) = 0, \tag{21}$$

and

$$\lim_{n \to \infty} \sup_{\mu \in \Lambda} \left(r(f_n, \mu^n) - H(\mu) \right) = 0. \tag{22}$$

An almost lossless universal code provides a point-wise convergence of the distortion to zero for every $\mu \in \Lambda$ while constraining the worst-case expected redundancy to vanish as the block length tends to infinity. It is obvious from Definition 7 that if Λ admits a classical lossless universal source code [6], [7], i.e., the worst-case average redundancy vanishes with zero distortion for every finite n, then it admits an almost lossless USC. The next result shows that there is a richer family of distributions that admits an almost lossless USC scheme:

Theorem 4 (Feasibility): The family $\mathcal{P}_{\mathcal{H}}(\mathcal{X})$ admits an almost lossless USC scheme.

The proof is presented in Section VIII-B.

Remarkably, Theorem 4 shows that a weak notion of universality allows to code the complete collection of finite entropy stationary memoryless sources defined on ∞-alphabets. Since the same result for lossless source coding is not possible [5], an interpretation of Theorem 4 is that a non-zero distortion (for any finite block-length) is strictly needed to make the average redundancy of an universal coding scheme vanishing with the block-length. To obtain this result, the two-stage approach presented in Section III-A was considered.

If we restrict the family of two-stage schemes to have an exhaustive first-stage mapping, i.e., $\pi_n(x) = \{x\}$ for all $x \in \mathcal{X}$ and $n \geq 1$, then we reduce the approach to the lossless setting (i.e., zero distortion for every finite block-length). In this case, if we apply the condition to obtain Theorem 4 (stated in Lemma 4 in Section VIII-C), this reduces to verify that the i-radius of the family grows sub-linearly with the block-length (more details presented in Sections VIII-B and VIII-C), which is the condition known for a family of distributions to have a nontrivial minimax redundancy rate [3], [5]–[7], [9].

A. Entropy Estimation with an Almost Lossless Universal Code: A Side Comment

In the lossless case, the existence of a weak minimax source coding scheme $\{f_n : \mathcal{X}^n \to \{0,1\}^*, n \geq 1\}$ for a family of distribution Λ implies that $\sup_{\mu \in \Lambda} \lim_{n \to \infty} \left(r(f_n, \mu^n) - H(\mu) \right) = 0$ [3]. Consequently, the average length of the code $r(f_n, \mu^n) = \mathbb{E}\left\{ \mathcal{L}(\mathcal{C}_n(\Phi_n(X^n))) \right\}/n$ is a weak consistent estimator of the entropy distribution-free in $\mu \in \Lambda$ [28]. For the family of finite entropy stationary and memoryless sources, we have that it is not feasible to have a weak minimax USC scheme in ∞-alphabets. In fact, [5, Theorem 2] says that for every code f_n and $n \geq 1$, there exists $\mu \in \mathcal{P}_{\mathcal{H}}(\mathcal{X})$ such that $r(f_n, \mu^n) = \infty$. In other words, there is no lossless variable-length source coding scheme that offers a weakly consistent estimator of the entropy using its average block-length (per letter). In contrast, Theorem 4 shows that there is an almost lossless USC scheme $\{(\phi_n, \psi_n, \mathcal{C}_n, \mathcal{D}_n) : n \geq 1\}$ with an average length that offers a distribution-free weakly consistent estimation of the entropy in $\mathcal{P}_{\mathcal{H}}(\mathcal{X})$. In fact from the proof of Lemma 4 (Sec.VIII-C), we have that

$$\lim_{n \to \infty} \sup_{\mu \in \mathcal{P}_{\mathcal{H}}(\mathcal{X})} \left(r(\phi_n, \mathcal{C}_n, \mu^n) - H_{\sigma(\pi_n)}(\mu) \right) = 0,$$

and from the fact that $\{\pi_n : n \geq 1\}$ is asymptotically sufficient for $\mathcal{P}_{\mathcal{H}}(\mathcal{X})$ (Definition 12 in Section VIII-B), it follows that: $\lim_{n \to \infty} r(\phi_n, \mathcal{C}_n, \mu^n) = H(\mu)$, for all $\mu \in \mathcal{P}_{\mathcal{H}}(\mathcal{X})$. Then, by relaxing the lossless block-wise assumption (introducing a non-zero distortion), we control the worse-case redundancy, which is bounded by the i-radius of $\mathcal{P}_{\mathcal{H}}(\mathcal{X})$ restricted to a sub-sigma field (see (38)). This flexibility enables the capacity to balance two sources of errors: $r(\phi_n, \mathcal{C}_n, \mu^n) - H_{\sigma(\pi_n)}(\mu)$ (a kind of estimation error) and $H(\mu) - H_{\sigma(\pi_n)}(\mu)$ (an approximation error), that at the end offers an distribution-free estimate of the entropy (point-wise) using the average length of the code.

V. Uniform Convergence of the Distortion

In this section, we further focus on a stronger notion of universal weak source coding. We study whether is possible to achieve an uniform convergence of the distortion to zero (over the entire family Λ), instead of the point-wise convergence stated in Definition 7. To this end, we restrict the analysis to the rich family of envelope distributions introduced in Section II-A. We can state the following dichotomy:

Theorem 5 (Uniform convergence): Let us consider the family of envelope distributions Λ_f.

i) If $f \in \ell_1(\mathcal{X})$, then there is a two-stage coding scheme $\{(\phi_n, \psi_n, \mathcal{C}_n) : n \geq 1\}$ with $|\pi_n| < \infty$ (finite size) such that

$$\lim_{n \to \infty} \sup_{\mu \in \Lambda_f} d(\phi_n, \psi_n, \mu^n) = 0, \text{ and}$$

$$\lim_{n \to \infty} \sup_{\mu \in \Lambda_f} \left(r(\phi_n, \mathcal{C}_n, \mu^n) - H(\mu) \right) = 0.$$

ii) Otherwise, i.e., $f \notin \ell_1(\mathcal{X})$, for any two-stage code $(\phi_n, \psi_n, \mathcal{C}_n)$ of length n with $|\pi_n| < \infty$ it follows that

$$\sup_{\mu \in \Lambda_f} d(\phi_n, \psi_n, \mu^n) = 1,$$

while if $|\pi_n| = \infty$, then

$$\sup_{\mu \in \Lambda_f} \left(r(\phi_n, \mathcal{C}_n, \mu^n) - H(\mu) \right) = \infty.$$

More generally, for a lossy code (f_n, g_n) of length n, provided that

$$\sup_{\mu \in \Lambda_f} d(f_n, g_n, \mu^n) < 1,$$

then

$$\sup_{\mu \in \Lambda_f} \left(r(f_n, \mu^n) - H(\mu) \right) = \infty.$$

The proof is presented in Section VIII-D.

Theorem 5 states that if the envelope function is summable, there is a two-stage coding scheme of finite size that offers a uniform convergence of the distortion to zero (over Λ_f), while ensuring that the worst-case average redundancy (over Λ_f) vanishes with the block-length. On the negative side, for all stationary memoryless sources indexed by a non-summable envelope function, it is not possible to achieve an uniform convergence of the distortion to zero with a finite size two-stage coding rule. An infinite size rule is indeed needed, i.e., $|\pi_n| = \infty$, eventually with n, that on the down-side it has an unbounded i-radius (details presented in Lemmas 5 and 1 at Section VIII-D). Importantly, this impossibility result remains when enriching the analysis with the adoption of general lossy coding rules (details in Sec. VIII-D.3).

Finally, it worths noting that the family Λ_f with $f \in \ell_1(\mathcal{X})$ has a finite regret and redundancy in the context of lossless universal source coding [9, Ths. 3 and 4]. Furthermore, summability is the necessary and sufficient condition on f that makes this collection strongly minimax universal in lossless source coding [9]. Then, based on this strong almost lossless source coding criterion (with a uniform convergence to zero of the distortion and the redundancy) it is not possible to code (universally) a richer family of distributions when restricting the analysis to envelope families.

VI. REDUNDANCY GAINS FOR SUMMABLE ENVELOPE FAMILIES

Theorem 5 states that we can achieve a uniform convergence of the distortion to zero while the worst-case redundancy vanishes with a two-stage lossy scheme if, and only if, Λ_f has a summable envelope function. On the lossless variable-length source coding side, if $f \in \ell_1(\mathcal{X})$ we know from Theorem 1 that the i-radius of the family

$$R^+(\Lambda_f^n) = \min_{v^n \in \mathcal{P}(\mathcal{X}^n)} \sup_{\mu^n \in \Lambda_f^n} \mathcal{D}(\mu^n | v^n),$$

is $o(n)$ [9, Ths. 3 and 4], which is equivalent to state that Λ_f is strongly minimax universal [3]. Therefore, under the assumption that $f \in \ell_1(\mathcal{X})$, the lossy approach with asymptotic vanishing distortion may appear to not be useful if no gains are observed in the way the worst-case redundancy approaches zero in (22), with respect to the normalized i-radius sequence $(R^+(\Lambda_f^n)/n)_{n \geq 1}$ that governs minimax redundancy in the lossless case [3], [9].

This section explores the feasibility of obtaining gains in terms of the minimax redundancy of a two-stage lossy approach tends to zero, when compared to the minimax redundancy (of the lossless scenario) for Λ_f when $f \in \ell_1(\mathcal{X})$. We focus on the finite size *tail-based partition* scheme used to prove Theorem 4 and the achievability part of Theorem 5.

A. Preliminaries

Let us consider a positive and non-decreasing sequence of integers $(k_n)_n$ and the collection of *tail partitions* induced by:

$$\tilde{\pi}_{k_n} := \left\{ \{1\}, \cdots, \{k_n - 1\}, \Gamma_{k_n-1}^c \right\}, \ \forall n \geq 1, \tag{23}$$

where $\Gamma_k = \{1, \ldots, k\}$. Note that $\tilde{\pi}_{k_n}$ resolves all the elements of $\Gamma_{k_n-1} = \{1, \ldots, k_n - 1\}$ and consequently, there is a pair $(\tilde{\phi}_n, \tilde{\psi}_n)$ associated with $\tilde{\pi}_{k_n}$ such that $\forall \mu \in \Lambda_f$,

$$d(\tilde{\phi}_n, \tilde{\psi}_n, \mu^n) \leq \mu(\Gamma_{k_n-1}^c) \leq \sum_{x \geq k_n} f(x). \tag{24}$$

Consequently, it follows that

$$\sup_{\mu \in \Lambda_f} d(\tilde{\phi}_n, \tilde{\psi}_n, \mu^n) \leq \sum_{x \geq k_n} f(x) < \infty. \tag{25}$$

It is then easy to verify that $(1/k_n)_n$ being $o(1)$ is the necessary and sufficient condition for the tail-based scheme to have the uniform convergence (over Λ_f) of the distortion to zero.

Concerning the worst-case minimax redundancy of the two-stage scheme induced by $\{\tilde{\pi}_{k_n}\}$ in (22), we can consider a lossy mapping $(\tilde{\phi}_n, \tilde{\psi}_n)$ consistent with $\tilde{\pi}_{k_n}$ (first-stage), where it is clear that the entropy of $Y^n = (\tilde{\phi}_n(X_1), .., \tilde{\phi}_n(X_n))$, which is $H(\tilde{\Phi}_n(X^n)) = nH_{\sigma(\tilde{\pi}_{k_n})}(\mu) \leq nH(\mu)$, is a lower bound for the performance of any prefix-free code acting on Y^n. Then given the first-stage $(\tilde{\phi}_n, \tilde{\psi}_n)$, we define the worse-case redundancy of any $\mathcal{C}_n : \Gamma_{k_n}^n \to \{0,1\}^*$ as follows:

$$\bar{R}(\Lambda_f^n, \tilde{\pi}_{k_n}, \mathcal{C}_n) := \sup_{\mu \in \Lambda_f} \left(r(\tilde{\phi}_n, \mathcal{C}_n, \mu^n) - H_{\sigma(\tilde{\pi}_{k_n})}(\mu) \right).$$

Therefore for any finite n and first-stage partition $\tilde{\pi}_{k_n}$, the minimax redundancy of the second-stage is:

$$\min_{\mathcal{C}_n:\Gamma_{k_n}^n \to \{0,1\}^*} \bar{R}(\Lambda_f^n, \tilde{\pi}_{k_n}, \mathcal{C}_n). \tag{26}$$

For the second stage, we can use again the connection between prefix-free codes and distributions to map $\mathcal{C}_n$ to a probability v in $\mathcal{P}(\Gamma_{k_n}^n)$, where the redundancy $r(\tilde{\phi}_n, \mathcal{C}_n, \mu^n) - H_{\sigma(\tilde{\pi}_{k_n})}(\mu)$ can be expressed as one over n the divergence restricted to the cells of $\tilde{\pi}_{k_n}$ [24], more precisely as $\frac{1}{n}\mathcal{D}_{\sigma(\tilde{\pi}_{k_n} \times \cdots \times \tilde{\pi}_{k_n})}(\mu^n|v)$ where:

$$\mathcal{D}_{\sigma(\tilde{\pi}_{k_n} \times \cdots \times \tilde{\pi}_{k_n})}(\mu^n|v) := \sum_{\mathcal{A} \in \tilde{\pi}_{k_n} \times \cdots \times \tilde{\pi}_{k_n}} \mu^n(\mathcal{A}) \log \frac{\mu^n(\mathcal{A})}{v(\mathcal{A})},$$

with $\tilde{\pi}_{k_n} \times \cdots \times \tilde{\pi}_{k_n}$ being a short-hand for the product partition of $\mathcal{X}^n$ induced by $\tilde{\pi}_{k_n}$. Consequently, the USC problem of the second-stage in (26) given the first stage, i.e., given $\tilde{\pi}_{k_n}$ in (23), can be expressed by $\frac{1}{n}R^+(\Lambda_f^n, \sigma(\tilde{\pi}_{k_n}))$ where:

$$R^+(\Lambda_f^n, \sigma(\tilde{\pi}_{k_n})) := \min_{v \in \mathcal{P}(\mathcal{X}^n)} \sup_{\mu^n \in \Lambda_f^n} \mathcal{D}_{\sigma(\tilde{\pi}_{k_n} \times \cdots \times \tilde{\pi}_{k_n})}(\mu^n|v). \tag{27}$$

$R^+(\Lambda_f^n, \sigma(\tilde{\pi}_{k_n}))$ can be interpreted as the i-radius of Λ_f^n restricted to the events of the sub-sigma field $\sigma(\tilde{\pi}_{k_n} \times \cdots \times \tilde{\pi}_{k_n})$.[9]

B. Redundancy Gain Analysis

Returning to our question, in the context of Eq. (27) we know that $\mathcal{D}_{\sigma(\tilde{\pi}_{k_n} \times \cdots \times \tilde{\pi}_{k_n})}(\mu^n|v) \leq \mathcal{D}(\mu^n|v)$ [24], therefore for any sequence $(k_n)_n$ of positive integers, it follows that $R^+(\Lambda_f^n) \geq R^+(\Lambda_f^n, \sigma(\tilde{\pi}_{k_n}))$ for all n, Consequently, we have that

$$\liminf_{n \to \infty} \frac{R^+(\Lambda_f^n)}{R^+(\Lambda_f^n, \sigma(\tilde{\pi}_{k_n}))} \geq 1. \tag{28}$$

In particular, we want to determine regimes on $(k_n)_n$ that guarantee an asymptotic gain in minimax redundancy in the sense that

$$\lim_{n \to \infty} \frac{R^+(\Lambda_f^n, \sigma(\tilde{\pi}_{k_n}))}{R^+(\Lambda_f^n)} = 0, \tag{29}$$

subject to the condition that $(1/k_n)_n$ is $o(1)$. If $(k_n)_n$ offers an asymptotic gain on minimax redundancy in the sense of Eq. (29), then any sequence where $(\tilde{k}_n)_n$ such that $(\tilde{k}_n)_n \leq (k_n)_n$ eventually in n offers a gain in the minimax redundancy[10]. Therefore it is important to determine the largest size sequence for the tail partition such that (29) is satisfied.

Note that any sequence $(k_n)_n$ offers a non-zero worst-case distortion for a finite block-length, and if $(1/k_n)_n$ is $o(1)$ this worst case distortion goes to zero at rate function of $(k_n)_n$ and the envelope function f. From this, one could suspect a gain in the minimax redundancy, in the sense established by (29), no matter how fast $(k_n)_n$ tends to infinity with the block-length as long as $k_n < \infty$ for any n. In other words, one simple conjecture is that the complexity of a family of distributions with infinite degrees of freedom, measured in terms of the rate of convergence to zero of the minimax redundancy per sample $(R^+(\Lambda_f^n)/n)_n$, cannot be reached by projecting this family into finite but dynamic (with the block-length) alphabets. However, the following result refutes this initial guess and determines a non-trivial regime for $(k_n)_n$ with no minimax redundancy gain. Importantly, this regime is fully determined by $(u_f^*(n))_n$ (see Def.3 in Section II-A), which can be interpreted as a sequence of critical dimensions for Λ_f that was introduced by Bontemps *et al.* [10], [20] in the context of lossless USC.

Theorem 6 (Minimax redundancy gains): Let $\Lambda_f \subset \mathcal{P}(\mathcal{X})$ be an *envelope family* with $f \in \ell_1(\mathcal{X})$ and $|\text{supp}(f)| = \infty$. In addition, let $\{\tilde{\pi}_{k_n} : n \geq 1\}$ be the collection of tail-based partitions in (23) driven by a positive non-decreasing sequence $(k_n)_n$. It follows that:

---

[9]More details are presented in (44) and (45) in Sec. VIII-C.
[10]This is a simple consequence of the fact that $k \geq \tilde{k}$ implies $\sigma(\tilde{\pi}_{\tilde{k}}) \subset \sigma(\tilde{\pi}_k)$.

i) If $(k_n)_n \geq (u_f^*(n))_n$ eventually in n, then there is no gain in minimax redundancy in the sense that:

$$\lim_{n \to \infty} \frac{R^+(\Lambda_f^n, \sigma(\tilde{\pi}_{k_n}))}{R^+(\Lambda_f^n)} = 1.$$

ii) Conversely, if $(k_n)_n$ is $o(u_f^*(n))$, i.e., $\lim_{n \to \infty} k_n/u_f^*(n) = 0$, then we have a minimax redundancy gain:

$$\lim_{n \to \infty} \frac{R^+(\Lambda_f^n, \sigma(\tilde{\pi}_{k_n}))}{R^+(\Lambda_f^n)} = 0.$$

The proof is presented in Section VIII-E.

Analysis and interpretation of Theorem 6:

1) First, we note that the sequence $(u_f^*(n))_n$ in (6) defines a notion of critical dimension (or cardinality) for the family Λ_f, as it characterizes a boundary (or phase transition) on the size of the tail-based two-stage coding schemes above from which no gains in terms of the rate of minimax redundancy are obtained.

2) If we consider the regime of redundancy gain, i.e., where $(k_n)_n$ is $o(u_f^*(n))$, it is simple to note that any arbitrary partition scheme $\{\pi_n : n \geq 1\}$ such that $|\pi_n| = |\tilde{\pi}_{k_n}| = k_n$ satisfies:

$$\lim_{n \to \infty} \frac{R^+(\Lambda_f^n, \sigma(\pi_n))}{R^+(\Lambda_f^n)} = 0.$$

Then, this scenario of redundancy gain can be extended to any finite alphabet partition strategy, and consequently, we can say that the condition $(k_n)_n$ is $o(u_f^*(n))$ offers a *trivial* regime of minimax redundancy gain. However, what is not evident is the fact that the tail-based partition offers a non-trivial regime of redundancy gain, in the sense that the condition $(k_n)_n \geq (u_f^*(n))_n$ eventually with n suffices to guarantee that:

$$\lim_{n \to \infty} \frac{R^+(\Lambda_f^n, \sigma(\tilde{\pi}_{k_n}))}{R^+(\Lambda_f^n)} = 1.$$

From this angle, the tail-based partition is efficient (or sufficient) to capture the asymptotic complexity of Λ_f with a minimum alphabet size. Complementing this richness property of $\{\tilde{\pi}_{k_n}\}$, it is simple to verify that the tail partition is an optimal solution when the objective is to minimize the worst-case distortion of a two-stage lossy coding scheme restricting the finite size $k > 0$ on the quantization.

3) From a complexity view-point, $R^+(\Lambda_f^n)$ measures the complexity of the lossless coding task. Then, for a given finite partition $\pi_n \in \mathcal{B}(\mathcal{X})$, $R^+(\Lambda_f^n) - R^+(\Lambda_f^n, \sigma(\pi_n)) \geq 0$ can be interpreted as the reduction on complexity by the process of projecting Λ_f^n into a finite alphabet, i.e.,

$$\Lambda_f^n/\sigma(\pi_n) := \left\{\mu^n/\sigma(\pi_n \times \cdots \times \pi_n) : \mu \in \Lambda_f\right\}$$

where $\mu/\sigma(\pi) := \{\mu(\mathcal{A}) : \mathcal{A} \in \sigma(\pi)\} \in \mathcal{P}(\mathcal{X}, \sigma(\pi))$ is a short-hand for the probability μ restricted to the sub-sigma field induced by π (details presented in Section VIII-E.3). Then, it is interesting to know if the i.i.d. family of envelope distributions $\left\{\Lambda_f^n : n \geq 1\right\}$ with $f \in \ell_1(\mathcal{X})$ admits a finite but dynamic alphabet reduction that captures its complexity asymptotically with n. For that question, we can introduce the following:

Definition 8: We say that $\Lambda \subset \mathcal{P}(\mathcal{X})$ has a finite alphabet reduction, if there exists a partition scheme $\{\pi_n : n \geq 1\}$ with $|\pi_n| = k_n < \infty$ such that $\lim_{n \to \infty} \frac{R^+(\Lambda^n, \sigma(\pi_n))}{R^+(\Lambda^n)} = 1$, or, equivalently, that $\{\Lambda^n : n \geq 1\}$ is equivalent to $\{\Lambda^n/\sigma(\pi_n) : n \geq 1\}$ in terms of asymptotic information complexity. In this case, we say that $(k_n)_n$ is a sequence of *sufficient sizes* (or sufficient) to represent Λ.

Definition 9: We say that $(k_n^*)_n$ is the *critical (or minimal) size to represent* Λ_f, if $(k_n^*)_n$ is a sequence of sufficient size to represent Λ (Def. 8), and no sequence $(l_n)_n$ exists such that: $(l_n)_n$ is sufficient to represent Λ and $(l_n)_n \ll (k_n^*)_n$. In this context, the proof of Theorem 6 shows as a corollary that $(u_f^*(n))_n$ is the critical size to represent Λ_f. The achievability part is obtained using the tail-based partition and some metric entropy lower bound for the i-radius extended from [10], [11], [20]. On the other hand, the converse argument derives from basic i-radius results for i.i.d. sources over finite alphabets [3] and results for envelope families on countably infinite alphabets [9].

Finally, from Boucheron *et al.* [9, Ths. 3 and 4, and Cor. 2], we have that Λ_f has either a finite alphabet reduction with a sub-linear critical size sequence given by $(u_f^*(n))_n$ (if $f \in \ell_1(\mathcal{X})$), or infinite minimax redundancy for all $n \geq 1$.

4) To illustrate the result, let us consider the exponentially decreasing envelop class studied in [9], [20]:

$$\Lambda_{f_{Ce^{-\alpha}}} = \left\{\mu \in \mathcal{P}(\mathcal{X}), f_\mu(x) \leq f_{Ce^{-\alpha}}(x) = Ce^{-\alpha x}\right\},$$

where $C > 0$ and $\alpha > 0$. It has been shown in [20, Prop. 6] that

$$\frac{1}{\alpha}\ln(Cx) \leq U_{f_{Ce^{-\alpha}}}(x) + 1 \leq \frac{1}{\alpha}\ln(\kappa Cx)$$

where $\kappa = 1/(1 - e^{-\alpha})$ and $U_f(x)$ is defined in Def. 20 (see Section VIII-E for details). Importantly for our analysis, it follows that $u^*_{f_{Ce^{-\alpha}}}(n) - 1 \leq U_{f_{Ce^{-\alpha}}}(n) < u^*_{f_{Ce^{-\alpha}}}(n)$ (see Section VIII-E.2), therefore we have that for all $n \geq 1$

$$\frac{1}{\alpha} \ln(Cn) + 1 < u^*_{f_{Ce^{-\alpha}}}(n) \leq \frac{1}{\alpha}(\ln(\kappa) + \ln(Cn)).$$

Consequently, the critical dimension of this exponential family, which determines the regime of redundancy gain, scales like $\sim (\frac{1}{\alpha} \ln n)_n$. Similar analysis can be conducted on the power-law envelopes and sub-exponential envelopes classes studied in [9], [10], [20], [21]. See also an excellent exposition of these last results in [8].

VII. Summary and Concluding Remarks

The problem of almost lossless universal source coding for countably infinite alphabet sources is introduced in this work. Our main result shows that a weak notion of universal (variable length) source coding is feasible for the entire class of finite entropy stationary memoryless sources. This result is obtained by tolerating a (non-zero) single-letter distortion in the encoding process that vanishes asymptotically with the block-length. To this end, one key idea is an induced sequence of partitions of the ∞-alphabet, which offers a way to control the worst-case average redundancy associated with the i-radius of a class of distributions restricted to a subsigma-field. We have also studied a stronger almost losses condition, asking for uniform convergence of the distortion to zero (over the family of distributions), where it turns out that this variation of weak universality can be achieved for the same class of envelope distributions that is strong minimax universal in the lossless case. This last result suggests that asking for a non-zero distortion that convergence to zero point-wise (over the family of distributions) is the strongest relaxation from the lossless criterion that allows us to control the worst-case redundancy of the problem. Finally, we show that it is possible to obtain gains in the rate of convergence of the worst-case redundancy of an almost lossless scheme, with respect to the worst-case redundancy of the lossless case, by tolerating a non-zero distortion that tends to zero with the block-length. In this context, we fully characterize the regime of gains for a two-stage lossy scheme induced by tail based partitions.

VIII. Proofs of the Main Results

A. Theorem 3

First, we introduce a result and some definitions that will be used in the proof.

1) Preliminaries:

Definition 10: For $\mu \in \mathcal{P}(\mathcal{X})$ its rate distortion function is given by [4], [25], [26]:

$$R_\mu(d) := \inf_{P(\tilde{X}|X) \text{ st. } \mathbb{P}(\tilde{X} \neq X) \leq d} I(X; \tilde{X}).$$

Definition 11: For any $\mu \in \mathcal{P}(\mathcal{X})$ and $\theta > 0$, let us define $\tilde{\mu}_\theta \in \mathcal{P}(\mathcal{X})$ by: $\tilde{\mu}_\theta(\{i\}) := \min\{\theta, \mu(\{i\})\}$ for all $i > 1$ and $\tilde{\mu}_\theta(\{1\}) := 1 - \kappa_\theta$ where $\kappa_\theta := \sum_{i>1} \tilde{\mu}_\theta(\{i\})$.

Lemma 3: (Ho *et al.* [17, Th. 1]) Let us consider $\mu \in \mathcal{P}(\mathcal{X})$, where $\mathcal{X}$ is an ∞-alphabet, then there is $d_0 > 0$ such that $\forall d \leq d_0$

$$R_\mu(d) = H(\mu) - H(\tilde{\mu}_{\theta(d)}) \tag{30}$$

with $\tilde{\mu}_\theta$ introduced in Def.11 and $\theta(d) > 0$ being the solution of the condition $\kappa_\theta = \sum_{i>1} \tilde{\mu}_\theta(\{i\}) = d$.

2) Proof of Theorem 3: We consider the non-trivial case where μ has infinite support over $\mathcal{X}$, i.e., $\inf_{x \in \mathcal{X}} f_\mu(x) = 0$, otherwise the problem reduces to a finite alphabet scenario where this result is known [3], [4].

We begin with the converse argument. This reduces to prove that any lossy coding scheme with zero asymptotic distortion has a rate that convergences to a limit that is grater or equal to $H(\mu)$ (see Def. 4). For that, let us assume that we have a lossy scheme $\{(f_n, g_n) : n \geq 1\}$ such that

$$\lim_{n \to \infty} d(f_n, g_n, \mu^n) = 0 \Leftrightarrow \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{P}\{X_i \neq (g_n(f_n(X^n)))_i\} = 0. \tag{31}$$

If we denote by $\hat{X}^n := g_n(f_n(X^n))$ the reconstruction, from lossless variable length source coding it is well-known that [3]:

$$r(f_n, \mu^n) \geq \frac{1}{n} I(X^n, \hat{X}^n)$$

$$\geq \frac{1}{n} \sum_{i=1}^n I(X_i; \hat{X}_i) \geq \frac{1}{n} \sum_{i=1}^n R_\mu(\mathbb{P}\{X_i \neq \hat{X}_i\}) \geq R_\mu\left(\frac{1}{n} \sum_{i=1}^n \mathbb{P}\{X_i \neq \hat{X}_i\}\right) = R_\mu(d(f_n, g_n, \mu^n)), \tag{32}$$

where for the inequalities in (32), we use that $\mathbf{X}$ is memoryless, the non-negativity of the conditional mutual information [4], and the convexity of the rate-distortion function of μ [4], [25]. For the rest we assume that μ is organized in decreasing order

in the sense that $f_\mu(1) \geq f_\mu(2) \geq \cdots$ [11] and that we are in the regime where $d \leq d_o$ (introduced in Lemma 3). Using Lemma 3, we have that $R_\mu(d) = H(\mu) - H(\tilde{\mu}_{\theta(d)})$, where if we consider

$$K_\mu(d) := \min\{k > 1 : f_\mu(k+1) \leq \theta(d)\}, \tag{33}$$

it is simple to verify that:

$$H(\tilde{\mu}_{\theta(d)}) = (1 - \kappa_{\theta(d)}) \log \frac{1}{1 - \kappa_{\theta(d)}} + (K_\mu(d) - 1) \cdot \theta(d) \log \frac{1}{\theta(d)} + \sum_{i > K_\mu(d)} f_\mu(i) \log \frac{1}{f_\mu(i)}. \tag{34}$$

From (30) and (31), we focus on exploring $H(\tilde{\mu}_{\theta(d_n)})$ when $d_n \to 0$. First, it is simple to verify that $d \to 0$ implies that $\theta(d) \to 0$ by definition. Then, for a fix $\mu \in \mathcal{P}_\mathcal{H}(\mathcal{X})$ with infinite support, the problem reduces to chacaterize $\lim_{\theta \to 0} H(\tilde{\mu}_\theta)$. Note that $\tilde{\mu}_\theta$ converges point-wise to the degenerate probability $\mu^* = (1, 0, \cdots)$ as θ vanishes [12]. However, by the entropy discontinuity [18], [19], [30], the convergence of the measure to μ^* is not sufficient to guarantee that $\lim_{\theta \to 0} H(\tilde{\mu}_\theta) = H(\mu^*) = 0$.

First, it is simple to note that $\kappa_\theta \to 0$, as $\theta \to 0$ considering that $\tilde{\mu}_\theta(i) \to 0$ for all $i \geq 1$, $\mu \in \mathcal{P}_\mathcal{H}(\mathcal{X})$, and the *dominated convergence theorem* [31]. Then, $\lim_{\theta \to 0}(1 - \kappa_\theta) \log \frac{1}{1 - \kappa_\theta} = 0$, which is the limit of the first term in the RHS of (34). For the rest, we define the self-information function $i_{\tilde{\mu}_\theta}(i) := \tilde{\mu}_\theta(\{i\}) \log 1/\tilde{\mu}_\theta(\{i\}) > 0$, for all $i > 1$, and $i_{\tilde{\mu}_\theta}(1) := 0$. By definition $\lim_{\theta \to 0} i_{\tilde{\mu}_\theta}(i) = 0$ point-wise in $\mathcal{X}$, noting that $\lim_{\theta \to 0} K_\mu(\theta) = \infty$ and $H(\mu) < \infty$. Furthermore, there is $\theta_0 > 0$ such that for all $\theta < \theta_0$, $0 \leq i_{\tilde{\mu}_\theta}(i) \leq i_{\tilde{\mu}_{\theta_0}}(i)$ for all [13] $i \geq 1$, where from the assumption that $H(\mu) < \infty$, and the fact that $K_\mu(\theta_0) < \infty$, then $(i_{\tilde{\mu}_{\theta_0}}(i)) \in \ell_1(\mathcal{X})$. Again by the *dominated convergence theorem* [31], $\lim_{\theta \to 0} \sum_{i \geq 1} i_{\tilde{\mu}_\theta}(i) = 0$ and consequently, $\lim_{\theta \to 0} H(\tilde{\mu}_\theta) = 0$ from (34).

Returning to (32), it follows that for all $n \geq 1$,

$$r(f_n, \mu^n) \geq R_\mu\big(\underbrace{d(f_n, g_n, \mu^n)}_{d_n :=}\big)$$
$$= H(\mu) - H\big(\tilde{\mu}_{\theta(d_n)}\big). \tag{35}$$

Finally, as $d_n \to 0$,

$$\liminf_{n \to 0} r(f_n, \mu^n) \geq H(\mu) - \limsup_{d_n \to 0} H\big(\tilde{\mu}_{\theta(d_n)}\big)$$
$$= H(\mu) - \lim_{\theta \to 0} H(\tilde{\mu}_\theta)$$
$$= H(\mu). \tag{36}$$

Therefore, the inequality in (36) implies that $R_{al}(\mu) \geq H(\mu)$ from Definition 4.

The achievability part (i.e., $R_{al}(\mu) \leq H(\mu)$) follows from the proof of Proposition 1 in Appendix I-C. $\qquad\square$

B. Theorem 4

For the proof of Theorem 4, we first introduce some definitions and an achievability result:

1) Preliminaries: Regarding the distortion, we need the following definition:

Definition 12: A sequence of partitions $\{\pi_n : n \geq 1\}$ of $\mathcal{X}$ is asymptotically sufficient for $\Lambda \subset \mathcal{P}(\mathcal{X})$, if it is asymptotically sufficient for every measure $\mu \in \Lambda$ (cf. Definition 5).

Concerning the analysis of the worst-case average redundancy in a lossy context, it is instrumental to introduce the divergence restricted to a sub-sigma field [24].

Definition 13: Let π be a partition of $\mathcal{X}$ and $\sigma(\pi) \subset \mathcal{B}(\mathcal{X})$ its induces sigma-field. Then, for every $\mu, v \in \mathcal{P}(\mathcal{X})$, the divergence of μ with respect to v restricted to $\sigma(\pi)$ is [24]:

$$\mathcal{D}_{\sigma(\pi)}(\mu|v) := \sum_{\mathcal{A} \in \pi} \mu(\mathcal{A}) \log \frac{\mu(\mathcal{A})}{v(\mathcal{A})}. \tag{37}$$

Definition 14: Let $\Lambda \subset \mathcal{P}(\mathcal{X})$ and π be a partition of $\mathcal{X}$. Fot any $n \geq 1$, the information radius of $\Lambda^n \subset \mathcal{P}(\mathcal{X}^n)$ restricted to $\sigma(\pi)$ is given by:

$$R^+(\Lambda^n, \sigma(\pi)) := \min_{v^n \in \mathcal{P}(\mathcal{X}^n)} \sup_{\mu^n \in \Lambda^n} \mathcal{D}_{\sigma(\pi \times \cdots \times \pi)}(\mu^n|v^n), \tag{38}$$

where $\pi \times \cdots \times \pi$ denotes the product partition of $\mathcal{X}^n$, $\mathcal{P}(\mathcal{X}^n)$ is the set of probability measures in $(\mathcal{X}^n, \mathcal{B}(\mathcal{X}^n))$, and Λ^n denotes the collection of all i.i.d (product) probabilities measures in $(\mathcal{X}^n, \mathcal{B}(\mathcal{X}^n))$ induced by Λ.

Lemma 4: Let us consider $\Lambda \subset \mathcal{P}(\mathcal{X})$. If there is a sequence of partitions $\{\pi_n : n \geq 1\}$ of $\mathcal{X}$ such that:

---

[11] We note that for the charcaterization of $R_\mu(d)$ as $d \to 0$ this assumption implies no loss of generality.

[12] In the ∞-alphabet the point-wise convergence of probabilities to a limit is equivalent to the weak convergence and the convergence in total variations [29].

[13] This follows from the fact that the function $\theta \log \frac{1}{\theta}$ is monotonically increasing in the range of $(0, \theta_0)$ for some $\theta_0 > 0$.

- $\{\pi_n : n \geq 1\}$ is asymptotically sufficient for Λ (Def. 12), and
- $(R^+(\Lambda^n, \sigma(\pi_n)))_n$ is $o(n)$,

then the family of stationary and memoryless sources with marginal distribution in Λ admits an almost lossless source coding scheme.

The proof is presented in Section VIII-C.

2) Proof of Theorem 4: Let us consider a collection of finite size partitions $\{\pi_n : n \geq 1\} \subset \mathcal{B}(\mathcal{X})$ with $k_n = |\pi_n| < \infty$ for all n. We note that if

$$\bigcap_{m \geq 1} \bigcup_{l \geq m} \pi_l(x) = \{x\}, \quad \text{for all } x \in \mathcal{X} \tag{39}$$

then, this partition scheme is asymptotically sufficient for $\mathcal{P}(\mathcal{X})$. Concerning the information radius, we have that $k_n = |\pi_n| < \infty$, which reduces the analysis to the finite alphabet case. In this context, it is well-known that [3, Theorem 7.5]:

$$\frac{k_n - 1}{2} \log n - K_1 \leq R^+(\mathcal{P}(\mathcal{X}^n), \sigma(\pi_n)) \leq \frac{k_n - 1}{2} \log n + K_2, \tag{40}$$

for some universal constants K_1 and K_2. Then, provided that $(k_n)_n$ is $o(n/\log n)$ it follows that $(R^+(\mathcal{P}(\mathcal{X}^n), \sigma(\pi_n)))_n$ is $o(n)$. There are numerous finite partition sequences that satisfy the conditions stated in (39) and $(k_n)_n$ being $o(n/\log n)$. For example, the tail partition family given by $\bar{\pi}_{k_n} := \{\{1\}, \{2\}, \cdots, \{k_n - 1\}, \Gamma_{k_n - 1}^c\}$, where $\Gamma_k := \{1, \cdots, k\}$ and $\Gamma_0 := \emptyset$, considering that $(1/k_n)_n$ is $o(1)$ and $(k_n)_n$ is $o(n/\log n)$. Finally, for all $\Lambda^n \subset \mathcal{P}(\mathcal{X}^n)$ we have by definition that $R^+(\Lambda^n, \sigma(\pi_n)) \leq R^+(\mathcal{P}(\mathcal{X}^n), \sigma(\pi_n))$, which proves the result by applying Lemma 4. $\square$

C. Proof of Lemma 4

Proof: First note that if $\{\pi_n : n \geq 1\}$ is asymptotically sufficient for the family Λ, it means that for all $\mu \in \Lambda$, $\lim_{n \to \infty} \mu(\cup_{k \geq n} \pi_k(x) \setminus \{x\}) = 0$ (Def. 5). If we denote by $k_n = |\pi_n|$ and $\pi_n = \{\mathcal{A}_{n,i} : 1 \leq i \leq k_n\}$, then we can construct $\phi_n : \mathcal{X}^n \to \{1, \ldots, k_n\}$ such that $\phi_n^{-1}(i) = \mathcal{A}_{n,i}$ for all $1 \leq i \leq k_n$. On the other hand, we can choose an arbitrary $y_{n,i} \in \mathcal{A}_{n,i}$ for each $i \in \Gamma_{k_n}$, and the mapping $\psi : \Gamma_{k_n} \to \mathcal{X}$ in the way $\psi(i) = y_{n,i}$. At this point, we observe:

$$\begin{aligned} d(\phi_n, \psi_n, \mu^n) &= \mathbb{P}(X \neq \psi_n(\phi_n(X))) \\ &= \sum_{i=1}^{k_n} \sum_{x \in \mathcal{A}_{n,i}} f_\mu(x) \rho_{0,1}(x, y_{n,i}) \\ &= \sum_{i=1}^{k_n} \mu(\mathcal{A}_{n,i} \setminus \{y_{n,i}\}) \\ &= \mu\left(\mathcal{X} \setminus \bigcup_{i=1}^{k_n} \{y_{n,i}\}\right). \end{aligned} \tag{41}$$

Then, from the hypothesis that $\{\pi_n : n \geq 1\}$ is asymptotically sufficient for μ, and the use of *bounded convergence theorem*, it is simple to verify that the RHS of (41) goes to zero (see Section I-A). Therefore, this convergence happens point-wise $\forall \mu \in \Lambda$.

Remark 1: It worth mentioning that (41) tends to zero, if and only if, $\lim_{n \to \infty} \cup_{k \geq n} \pi_k(x) = \{x\}$ μ-almost surely (cf. Lemma 1). Hence, to achieve a point-wise convergence to zero of the distortion over Λ, for this two-stage scheme, it is necessary and sufficient that $\{\pi_n : n \geq 1\}$ is asymptotically sufficient for Λ.

Regarding the second coding stage, we ideally need to find a lossless code with the least worst-case average redundancy over the family

$$\Lambda^n/\sigma(\pi_n) := \{\mu^n/\sigma(\pi_n \times \cdots \times \pi_n) : \mu \in \Lambda\} \subset \mathcal{P}(\mathcal{X}^n, \sigma(\pi_n \times \cdots \times \pi_n)),$$

where $\mu/\sigma(\pi) \in \mathcal{P}(\mathcal{X}, \sigma(\pi))$ is a short hand for the probability μ restricted to the sub-sigma field induced by π [14], and $P(\mathcal{X}, \sigma)$ denotes the collection of probabilities restricted to the events of the sub-sigma field $\sigma \subset \mathcal{B}(\mathcal{X})$.

In fact, for a lossless prefix-free code $\mathcal{C}_n : \Gamma_{k_n}^n \to \{0,1\}^*$, associated to the first stage ϕ_n, its worst-case average redundancy over Λ^n is given by:

$$R(\Lambda^n, \phi_n, \mathcal{C}_n) := \sup_{\mu \in \Lambda}\left(r(\phi_n, \mathcal{C}_n, \mu^n) - H(\mu)\right). \tag{42}$$

For any fixed $\mu \in \Lambda$, it is clear that the entropy of $\Phi_n(X^n)$ is a lower bound for the average rate of the code, i.e., $r(\phi_n, \mathcal{C}_n, \mu^n) \geq \frac{1}{n}H(\Phi_n(X^n)) = H_{\sigma(\pi_n)}(\mu)$, then constraining to the events of $\sigma(\pi_n)$, we are interested in controlling the following stringer worst-case overhead:

$$\bar{R}(\Lambda^n, \phi_n, \mathcal{C}_n) := \sup_{\mu \in \Lambda}\left(r(\phi_n, \mathcal{C}_n, \mu^n) - H_{\sigma(\pi_n)}(\mu)\right). \tag{43}$$

---

[14]Note that if $\mu \in \mathcal{P}(\mathcal{X})$, then the restriction $\mu/\sigma(\pi) \in \mathcal{P}(\mathcal{X}, \sigma(\pi))$ reduces to the evaluation of μ over the cells of π. In fact, $\{\mu(B) : B \in \pi\}$ plays the role of the probability mass function of $\mu/\sigma(\pi)$ on the measurable space $(\mathcal{X}, \sigma(\pi))$.

Note that $H_{\sigma(\pi_n)} - H(\mu) \leq 0$ and thus, $\bar{R}(\Lambda^n, \phi_n, \mathcal{C}_n) \geq R(\Lambda^n, \phi_n, \mathcal{C}_n)$. Then, we can choose a code solution to the following mini-max problem:

$$\mathcal{C}_n^* := \arg \min_{\mathcal{C}^n : \Gamma_{k_n}^n \to \{0,1\}^*} \bar{R}(\Lambda^n, \phi_n, \mathcal{C}_n), \text{ for all } n \geq 1. \tag{44}$$

From the close connection between probabilities and prefix free codes [3], the performance of the optimal code in (44) is tightly related to the i-radius of the family $\Lambda^n / \sigma(\pi_n)$ in (38), in the sense that $\forall n \geq 1$:

$$\frac{R^+(\Lambda^n, \sigma(\pi_n))}{n} \leq \bar{R}(\Lambda^n, \phi_n, \mathcal{C}_n^*) \leq \frac{R^+(\Lambda^n, \sigma(\pi_n)) + 2}{n}. \tag{45}$$

Finally, from the hypothesis on the information radius and (45), we have that:

$$\lim_{n \to \infty} R(\Lambda^n, \phi_n, \mathcal{C}_n^*) \leq \lim_{n \to \infty} \bar{R}(\Lambda^n, \phi_n, \mathcal{C}_n^*) = 0. \tag{46}$$

Remark 2: The inequalities in (45) states that the sub-linear trend (with the block-length) on the i-radius of $\{\Lambda^n / \sigma(\pi_n) : n \geq 1\}$ is a necessary and sufficient condition for the existence of a strongly minimax universal code for $\{\Lambda^n / \sigma(\pi_n) : n \geq 1\}$.

$\square$

D. Theorem 5

Let us first introduce some notations, definitions and results that will be used in the proof of Theorem 5.

Definition 15: For $\Lambda \subset \mathcal{P}(\mathcal{X})$ its i-radius is given and denoted by:

$$R^+(\Lambda) := \inf_{v \in \mathcal{P}(\mathcal{X})} \sup_{\mu \in \Lambda} \mathcal{D}(\mu|v).$$

Definition 16: For a function $\phi : \mathcal{X} \to \mathcal{I}$ (where $\mathcal{I}$ is either a finite or a countably infinite set) and $\mu \in \mathcal{P}(\mathcal{X})$, let us denote by $v_\mu \in \mathcal{P}(\mathcal{I})$ the distribution induced by ϕ in $\mathcal{I}$ trough the standard construction[15]: $v_\mu(B) := \mu(\phi^{-1}(B))$ for all $B \subset \mathcal{I}$.

Lemma 5: Let $\phi : \mathcal{X} \to \mathcal{I}$ be a mapping where $\mathcal{I}$ is a countably infinite set. Then for any non-negative envelope function $f : \mathcal{X} \to \mathbb{R}^+$, there is $\tilde{f} : \mathcal{I} \to \mathbb{R}^+$ given by[16]

$$\tilde{f}(i) := \min \left\{ \sum_{x \in \mathcal{A}_i} f(x), 1 \right\}, \tag{47}$$

such that $\{v_\mu : \mu \in \Lambda_f\} = \tilde{\Lambda}_{\tilde{f}} := \left\{ v \in \mathcal{P}(\mathcal{I}) : f_v(i) \leq \tilde{f}(i) \text{ for all } i \in \mathcal{I} \right\}$.

The proof of this result is presented in Appendix I-D.

Lemma 5 implies that envelope families on $\mathcal{X}$ map to envelope families on $\mathcal{I}$ through the mapping ϕ. In this context, the result by Boucheron *et al.* [9] in Theorem 1 (in Section II) will be instrumental to prove Theorem 5.

Proof of Theorem 5:

1) Achievability: If $f \in \ell_1(\mathcal{X})$, the fact that Λ_f has a uniform bound on the tails of the distributions suggests that a family of tail truncating partitions should be considered to achieve the claim i). Let us define

$$\tilde{\pi}_{k_n} := \left\{ \{1\}, \cdots, \{k_n\}, \Gamma_{k_n}^c \right\}, \tag{48}$$

which resolves the elements of $\Gamma_{k_n} = \{1, \ldots, k_n\}$ and, consequently, there is a pair $(\tilde{\phi}_n, \tilde{\psi}_n)$ associated with $\tilde{\pi}_{k_n}$ such that $\forall \mu \in \Lambda_f$:

$$d(\tilde{\phi}_n, \tilde{\psi}_n, \mu^n) \leq \mu(\Gamma_{k_n}^c) \leq \sum_{x > k_n} f(x). \tag{49}$$

In fact, $\sup_{\mu \in \Lambda_f} d(\tilde{\phi}_n, \tilde{\psi}_n, \mu^n) \leq \sum_{x > k_n} f(x) < \infty$, and $(1/k_n)_n$ being $o(1)$ is a sufficient condition to satisfy the uniform convergence of the distortion to zero. Furthermore, from the proof of Lemma 4 (Eq.(45)) and (40), there is a lossless coding scheme $\left\{ \tilde{\mathcal{C}}_n : \Gamma_{k_n+1}^n \to \{0,1\}^* : n \geq 1 \right\}$ such that $\sup_{\mu \in \Lambda_f} \left(r(\phi_n, \tilde{\mathcal{C}}_n, \mu^n) - H(\mu) \right) \leq k_n \cdot \log \sqrt{n}/n + O(1/n)$. Therefore, we can consider $(k_n)_n$ being $O(n^\tau)$ with $\tau \in (0, 1)$ to conclude the achievability part.

---

[15]There is no question about the measurability of $\phi(\cdot)$ as we consider that $\mathcal{B}(\mathcal{X})$ is the power set.

[16]$\mathcal{A}_i := \phi^{-1}(\{i\})$ for any $i \in \mathcal{I}$.

2) Converse for Two-Stage Lossy Coding Schemes: [17] For the converse part, let us first consider an arbitrary two-stage lossy rule $(\phi_n, \psi_n, \mathcal{C}_n)$ with a finite partition $k_n = |\pi_n| < \infty$. If we denote its prototypes by $\mathcal{Y}_n := \{\psi(i) : i \in \Gamma_{k_n}\}$, it is clear that there exists $\mu \in \Lambda_f$ such that $\text{supp}(\mu) \subset \mathcal{Y}_n^c$, and consequently, $d(\phi_n, \psi_n, \mu^n) = 1$ for all n. Therefore, for any finite size partition rule $(\phi_n, \psi_n, \mathcal{C}_n)$ it follows that $\sup_{\mu \in \Lambda_f} d(\phi_n, \psi_n, \mu^n) = 1$ for all $n \geq 1$ and hence, when $f \notin \ell_1(\mathcal{X})$ no uniform convergence on the distortion can be achieved with a finite size lossy rule.

On the other hand, for the family of infinite size partition rules, i.e., $(\phi_n, \psi_n, \mathcal{C}_n)$ such that $|\pi_n| = \infty$, we focus our analysis on $R^+(\Lambda_f^n, \sigma(\pi_n))$ in (38). Let us fix a block-length $n > 0$ and a rule $(\phi_n, \psi_n, \mathcal{C}_n)$ of infinite size. For sake of clarity, we consider that $\phi_n : \mathcal{X} \to \mathcal{I}$, where $\mathcal{I}$ is a ∞-alphabet. For any $\mu \in \Lambda_f$, v_μ denotes the induced measure in $\mathcal{I}$ by the mapping ϕ_n trough the standard construction (see Def. 16). In addition, it is simple to verify that for any pair $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{X})$

$$\mathcal{D}_{\sigma(\pi_n)}(\mu_1|\mu_2) = \mathcal{D}(v_{\mu_1}|v_{\mu_2})$$
$$= \sum_{i \in \mathcal{I}} f_{v_{\mu_1}}(i) \frac{f_{v_{\mu_1}}(i)}{f_{v_{\mu_2}}(i)}, \quad (50)$$

where $\pi_n = \{\mathcal{A}_{n,i} = \phi_n^{-1}(\{i\}) : i \in \mathcal{I}\}$ and $f_{v_\mu}(i) := \mu(\mathcal{A}_{n,i}) \; \forall i$ denotes the pmf of v_μ on $\mathcal{I}$. Then,

$$R^+(\Lambda_f^n, \sigma(\pi_n)) = R^+(\{v_\mu^n : \mu \in \Lambda_f\}) \quad (51)$$

where v_μ^n denotes de product probability on $\mathcal{I}^n$ with marginal v_μ and $\mathcal{P}(\mathcal{I}^n)$ is the collection of probability measures on $\mathcal{I}^n$. Then, the i-radius of Λ_f^n restricted to the product sub-sigma field $\sigma(\pi_n \times \cdots \times \pi_n)$ is equivalent to the information radius of $\{v_\mu^n : \mu \in \Lambda_f\} \subset \mathcal{P}(\mathcal{I}^n)$ (Def. 15). From Lemma 5, $\{v_\mu^n : \mu \in \Lambda_f\}$ is an envelope family with envelope function given by (47). It is simple to verify that $f \notin \ell_1(\mathcal{X})$ implies that $\tilde{f} \notin \ell_1(\mathcal{I})$, Theorem 1 and (51) tell us that $R^+(\Lambda_f^n, \sigma(\pi_n)) = \infty$. Finally, since the i-radius in (51) tightly bounds the least-worst expected redundancy for the second lossless coding stage (see (44) and (45)), this implies that:

$$\sup_{\mu \in \Lambda_f} (r(\phi_n, \mathcal{C}_n, \mu^n) - H(\mu)) = \infty, \quad (52)$$

which concludes the argument.

3) Converse for general variable-length lossy codes: Let us consider a general lossy code (f_n, g_n) of length $n > 0$ introduced in Section III. Without loss of generality we can decouple f_n as the composition of a vector quantizer $\phi_n : \mathcal{X}^n \to \mathcal{I}_n$, where $\mathcal{I}_n$ is an index set, and a prefix-free losses mapping $\mathcal{C}_n : \mathcal{I}_n \to \{0, 1\}^*$, where $f(x^n) = \mathcal{C}_n(\phi_n(x^n))$ for all $x^n \in \mathcal{X}^n$. From this, we characterize the vector quantization induced by (f_n, g_n) as follows:

$$\pi_n := \{\phi_n^{-1}(\{i\}) : i \in \mathcal{I}_n\} \subset \mathcal{B}(\mathcal{X}^n). \quad (53)$$

Using this two-stage (vector quantization-coding) view, it is possible to show that[18]:

$$\bar{R}(\Lambda_f, f_n) := \sup_{\mu \in \Lambda_f} \left(r(f_n, \mu^n) - \frac{1}{n} H_{\sigma(\pi_n)}(\mu^n) \right) \geq \frac{1}{n} \inf_{v \in \mathcal{P}(\mathcal{X}^n)} \sup_{\mu \in \Lambda_f} \mathcal{D}_{\sigma(\pi_n)}(\mu^n|v), \quad (54)$$

which means that the worst-case overhead, expressed by $\bar{R}(\Lambda_f, f_n)$, is lower bounded by the i-radius of the n-fold family Λ_f^n projected into the sub-sigma field induced by π_n, i.e., a quantization of $\mathcal{X}^n$. Considering that $f \notin \ell_1(\mathcal{X})$, we follow the construction presented in [9] that shows that there is an infinite collection of distributions $\tilde{\Lambda} = \{\tilde{\mu}_j \in \Lambda_f, j \in \mathcal{J}\}$ with $|\mathcal{J}| = \infty$, where if we denote by

$$\mathcal{A}_{\tilde{\mu}_j} := \text{supp}(\tilde{\mu}_j) = \{x \in \mathcal{X} : f_{\tilde{\mu}_j}(x) > 0\},$$

then $|\mathcal{A}_{\tilde{\mu}_j}| < \infty$ for each $j \in \mathcal{J}$ and for any $j_1 \neq j_2$, $\mathcal{A}_{\tilde{\mu}_{j_1}} \cap \mathcal{A}_{\tilde{\mu}_{j_2}} = \emptyset$. In this context, for each $j \in \mathcal{J}$ $A_{\tilde{\mu}_j}^n := \mathcal{A}_{\tilde{\mu}_j} \times \ldots \times \mathcal{A}_{\tilde{\mu}_j} \in \mathcal{X}^n$ is the support of $\tilde{\mu}_j^n$.

At this point, let us use the assumption that: $\sup_{\mu \in \Lambda_f} d(f_n, g_n, \mu^n) < 1$. This implies that $\sup_{\mu \in \tilde{\Lambda}} d(f_n, g_n, \mu^n) < 1$. From the fact that $\tilde{\Lambda}$ is an infinite collection of probabilities with disjoint supports and the definition of the distortion, it is simple to verify that we need to allocate at least one prototype[19] per cell $A_{\tilde{\mu}_j}^n$, which implies that $|\mathcal{I}_n| = \infty$, because otherwise it follows that $\sup_{\mu \in \tilde{\Lambda}} d(f_n, g_n, \mu^n) = 1$.

Using (54), we will focus on evaluating the information radius of $\tilde{\Lambda}^n$ projected over the measurable space $(\mathcal{X}^n, \sigma(\pi_n))$ considering that by definition:

$$\inf_{v \in \mathcal{P}(\mathcal{X}^n)} \sup_{\mu \in \tilde{\Lambda}} \mathcal{D}_{\sigma(\pi_n)}(\mu^n|v) \leq \inf_{v \in \mathcal{P}(\mathcal{X}^n)} \sup_{\mu \in \Lambda_f} \mathcal{D}_{\sigma(\pi_n)}(\mu^n|v). \quad (55)$$

---

[17]We first present this preliminary converse argument, as it provides the ground to explore the redundancy gain analysis presented in Section VI.

[18]The proof of (54) is presented in Appendix I-E.

[19]The prototypes of (f_n, g_n) is the set $\mathcal{B}_n = \{g_n(f_n(x_1^n)) : x_1^n \in \mathcal{X}^n\}$.

For every $j \in \mathcal{J}$, let us define the covering of the support of $\tilde{\mu}_j^n \in \tilde{\Lambda}^n$ by

$$\pi_n(\mathcal{A}_{\tilde{\mu}_j}^n) := \left\{ B \in \pi_n : \mathcal{A}_{\tilde{\mu}_j}^n \cap B \neq \emptyset \right\} \text{ and } \mathcal{B}(\mathcal{A}_{\tilde{\mu}_j}^n) := \bigcup_{B \in \pi_n(\mathcal{A}_{\tilde{\mu}_j}^n)} B. \tag{56}$$

By construction, we note that $\left| \mathcal{A}_{\tilde{\mu}_j}^n \right| < \infty$ and consequently $\left| \pi_n(\mathcal{A}_{\tilde{\mu}_j}^n) \right| < \infty$ for all $j \in \mathcal{J}$. Considering that π_n has an infinite number of cells, we can choose an infinite subset of elements in $\tilde{\Lambda}^n = \left\{ \tilde{\mu}_j^n : j \in \mathcal{J} \right\}$ in the following way: We fix $j_1 = 1$ and $\bar{\mu}_1 = \tilde{\mu}_1 \in \tilde{\Lambda}$, then we consider

$$j_2 = \min \left\{ j > j_1, \text{ such that } \mathcal{B}(\mathcal{A}_{\tilde{\mu}_j}^n) \cap \mathcal{B}(\mathcal{A}_{\tilde{\mu}_1}^n) = \emptyset \right\} < \infty, \tag{57}$$

and we choose $\bar{\mu}_2 = \tilde{\mu}_{j_2} \in \tilde{\Lambda}$. Iterating this rule, at the k-stage ($k \geq 2$) we solve

$$j_k = \min \left\{ j > j_{k-1}, \text{ such that } \mathcal{B}(\mathcal{A}_{\tilde{\mu}_j}^n) \cap \left(\bigcup_{l=1}^{k-1} \mathcal{B}(\mathcal{A}_{\bar{\mu}_l}^n) \right) = \emptyset \right\} < \infty \tag{58}$$

and we take $\bar{\mu}_k = \tilde{\mu}_{j_k} \in \tilde{\Lambda}$, for all $k \geq 1$. Note that the solution of (58) is guaranteed from the fact that $|\mathcal{J}| = \infty$ and $\left| \pi_n(\mathcal{A}_{\tilde{\mu}_j}^n) \right| < \infty$ for all j. Finally, we define $\bar{\Lambda}^n := \{ \bar{\mu}_l^n : l \geq 1 \} \subset \tilde{\Lambda}^n$. Importantly (for the computation of the i-radius), this restricted family of distributions has the property that their support coverings in (56) are disjoint by its construction in (58). From $\bar{\Lambda}^n$, we can induce the following partition:

$$\xi_n := \left\{ \mathcal{B}(A_{\bar{\mu}_l}^n) : l \geq 1 \right) \right\} \cup \left(\mathcal{X}^n \setminus \bigcup_{l \geq 1} \mathcal{B}(A_{\bar{\mu}_l}^n) \right) \subset \sigma(\pi_n), \tag{59}$$

where the last identity is from the construction, as every cell of ξ_n is a finite union of cells of π_n (i.e., $\xi_n \ll \pi_n$). It is not difficult to check that for every $v \in \mathcal{P}(\mathcal{X}^n)$, we have that [20]

$$\sup_{\bar{\mu}_l \in \bar{\Lambda}} \mathcal{D}_{\sigma(\xi_n)}(\bar{\mu}_l^n | v) = \infty. \tag{60}$$

Consequently, we have that

$$\inf_{v \in \mathcal{P}(\mathcal{X}^n)} \sup_{\tilde{\mu}_j \in \tilde{\Lambda}} \mathcal{D}_{\sigma(\pi_n)}(\tilde{\mu}_j^n | v) \geq \inf_{v \in \mathcal{P}(\mathcal{X}^n)} \sup_{\tilde{\mu}_j \in \tilde{\Lambda}} \mathcal{D}_{\sigma(\xi_n)}(\tilde{\mu}_j^n | v)$$
$$\geq \inf_{v \in \mathcal{P}(\mathcal{X}^n)} \sup_{\bar{\mu}_l \in \bar{\Lambda}} \mathcal{D}_{\sigma(\xi_n)}(\bar{\mu}_l^n | v) = \infty, \tag{61}$$

the first inequality derives from $\xi_n \ll \pi_n$ and the second from $\bar{\Lambda} \subset \tilde{\Lambda}$. Finally (61) and the relationship between the worst-case redundancy and the information radius in (54) (Prop. 2 in Appendix I-E) imply that

$$\bar{R}(\tilde{\Lambda}, f_n) = \sup_{\tilde{\mu}_j \in \tilde{\Lambda}} \left(r(f_n, \tilde{\mu}_j^n) - \frac{1}{n} H_{\sigma(\pi_n)}(\tilde{\mu}_j^n) \right) = \infty. \tag{62}$$

In other words, from (62) there is $jo \in \mathcal{J}$ such that $r(f_n, \tilde{\mu}_{jo}^n) - \frac{1}{n} H_{\sigma(\pi_n)}(\tilde{\mu}_{jo}^n) = \infty$, where considering that by construction $\frac{1}{n} H_{\sigma(\pi_n)}(\tilde{\mu}_{jo}^n) \leq H(\tilde{\mu}_{jo}) < \log \left| \mathcal{A}_{\tilde{\mu}_{jo}} \right| < \infty$, this implies that $r(f_n, \tilde{\mu}_{jo}^n) - H(\tilde{\mu}_{jo}) = \infty$. Therefore, we have that

$$\sup_{\tilde{\mu}_j \in \tilde{\Lambda}} \left(r(f_n, \tilde{\mu}_j^n) - H(\tilde{\mu}_j^n) \right) = \infty, \tag{63}$$

which concludes the result considering that $\tilde{\Lambda} \subset \Lambda_f$. $\qquad\square$

E. Theorem 6

Without loss of generality, in this section we assume that $\mathcal{X}$ is the integer set $\mathbb{N} \setminus \{0\}$. To organize the proof, we first introduce some definitions and a series of important results that will be used in the main argument.

---

[20]This results follows from the fact that the elements of $\bar{\Lambda}^n$ projected into the sub-sigma field ξ_n degenerate, i.e., $H_{\sigma(\xi_n)}(\bar{\mu}_l^n) = 0$ for all $l \geq 1$.

1) Preliminaries:

Definition 17: [10] For a non-negative function $f : \mathcal{X} \to \mathbb{R}^+$, the *hazard function* of Λ_f is given by

$$h_f(u) := -\ln \bar{F}_f(u)$$

for all $u \in \mathcal{X}$.

Definition 18: [10] The continuous extension of $(h_f(u))_{u \in \mathcal{X}}$ to the positive real line $\mathbb{R}^+$ is defined by means of the following linear interpolation[21]:

$$\tilde{h}_f(k\lambda + (1-\lambda)(k+1)) := \lambda h_f(k) + (1-\lambda)h_f(k+1)$$

for $k \in \mathcal{X}$ and for all $\lambda \in [0,1)$.

Consistently with $(\tilde{h}_f(x))_{x \geq 0}$, in Def.18, it is possible to extend $(F_f(u))_{u \in \mathcal{X}}$ to $\mathbb{R}^+$ using the relationship expressed in Def. 17:

Definition 19: Given $f : \mathcal{X} \to \mathbb{R}^+$, the continuous extension of $(F_f(u))_{u \in \mathcal{X}}$ using $(\tilde{h}_f(x))_{x \geq 0}$ (Def. 18) is denoted by $(\mathcal{F}_f(x))_{x \geq 0}$ and called the *smoothed envelope distributions* of Λ_f.

Definition 20: [10, Eq.(1)] Under the setting of Def. 19, a function $U_f : [1, \infty] \longrightarrow \mathbb{R}$ can be defined as the solution of:

$$U_f(t) := \mathcal{F}_f^{-1}(1 - 1/t), \tag{64}$$

for all $t \geq 1$.

Definition 21: Let $(f(x))_{x \in \mathcal{X}}$ be non-negative and in $\ell_1(\mathcal{X})$. A non-decreasing continuous function can be obtained as[22]:

$$l_f(1/\epsilon) := \int_1^{1/\epsilon^2} \frac{U_f(x)}{2x} \partial x,$$

for any $\epsilon > 0$.

Definition 22: Let $(\epsilon_{f,n})_{n \geq 1}$ be the sequence obtained as the solution (point-wise) of: $l_f(1/\epsilon) = \frac{n\epsilon^2}{8}$ for all n.

We are in the position to state two instrumental results:

Lemma 6: [9, Th. 4] Let $\{\Lambda_f^n, n \geq 1\}$ be the envelope collection of stationary and memoryless sources with $f \in \ell_1(\mathcal{X})$ and tail function $(\bar{F}_f(u))_{u \in \mathcal{X}}$. Then for any $n \geq 1$

$$R^+(\Lambda_f^n) \leq \inf_{u \geq 1} \left[n\bar{F}_f(u)\log(e) + \frac{u-1}{2} \cdot \log n \right] + 2.$$

Lemma 7: [10, Prop. 5] Under the setting of Lemma 6, there is a sequence $(\xi_n)_n$ being $o(1)$ (and function of f) such that:[23]

$$R^+(\Lambda_f^n) \geq (1 + \xi_n)\log(e) \int_1^n \frac{U_f(x)}{2x} \partial x, \text{ for all } n \geq 1. \tag{65}$$

We also use results from the seminal work of Haussler and Opper [11] that we summarize here:

Definition 23: For any $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{X})$, the *Hellinger distance* is given/denoted by: $d_h(\mu_1.\mu_2)^2 := \sum_{x \in \mathcal{X}}(\sqrt{f_{\mu_1}(x)} - \sqrt{f_{\mu_2}(x)})^2$.

Definition 24: [11] For $\Lambda \subset \mathcal{P}(\mathcal{X})$ and $\epsilon > 0$, let $\mathcal{D}_\epsilon(\Lambda)$ be the smallest cardinality of a partition of Λ, whose cells have a diameter smaller or equal then ϵ (with respect to d_h in $\mathcal{P}(\mathcal{X})$) or it is infinity if no finite partition satisfies the diameter condition. Then, the metric entropy of Λ is given by:

$$\mathcal{H}_\epsilon(\Lambda) := \ln(\mathcal{D}_\epsilon(\Lambda)).$$

Then the following important results can be stated:

Lemma 8: [11, Lemma 7] Let us assume that $\Lambda \subset \mathcal{P}(\mathcal{X})$ is totally bounded, i.e., $\mathcal{H}_\epsilon(\Lambda) < \infty$ for all $\epsilon > 0$. Then, for all $n \geq 1$,

$$R^+(\Lambda^n) \geq \log(e) \cdot \sup_{\epsilon > 0} \min \left\{ \mathcal{H}_\epsilon(\Lambda), \frac{n\epsilon^2}{8} \right\} - 1.$$

Corollary 1: From Lemma 8, if we let $\epsilon_{\Lambda,n}^* := \inf \left\{ \epsilon > 0 : \mathcal{H}_\epsilon(\Lambda) \leq \frac{n\epsilon^2}{8} \right\}$, we have that $\forall n \geq 1$:

$$R^+(\Lambda^n) \geq \log(e) \cdot \mathcal{H}_{\epsilon_{\Lambda,n}^*}(\Lambda) - 1, \tag{66}$$

and, consequently,

$$\liminf_{n \to \infty} R^+(\Lambda^n)/\log(e)\mathcal{H}_{\epsilon_{\Lambda,n}^*}(\Lambda) \geq 1.$$

---

[21]This idea was proposed by Bontemps *et al.* [10] following Anderson [32].

[22]See Eq.(67).

[23]Remarkably, it has been shown in [10, Th. 2] that this closed-form lower bound captures the precise asymptotic of the information radius of the envelope class, meaning that: $\lim_{n \to \infty} R^+(\Lambda_f^n)/\log(e) \int_1^n \frac{U_f(x)}{2x} \partial x = 1$.

It is worth noting that the metric entropy lower bound for the information radius in Corollary 1 can be asymptotically tight, e.g., under a *slowly variant condition* on the behaviour of $\mathcal{H}_\epsilon(\Lambda)$ as ϵ goes to zero [10].[24]

Importantly for envelope families, when $f \in \ell_1(\mathcal{X})$ the asymptotic of the metric entropy of Λ_f, i.e., $\lim_{\epsilon \to 0} \mathcal{H}_\epsilon(\Lambda_f)$, is known. More precisely, Bontemps *et al.* [10, Prop.4] have shown that

$$\mathcal{H}_\epsilon(\Lambda_f) = (1 + o_f(1)) \int_1^{1/\epsilon^2} \frac{U_f(x)}{2x} \partial x \tag{67}$$

as ϵ tends to 0.

2) Proof of Theorem 6— Regime of Gain in Minimax Redundancy: Let us assume that $(k_n)_n$ is $o(u_f^*(n))$. This part derives directly from the tight lower and upper bounds developed by Bontemps *et al.* [10, Th. 2] and Boucheron *et al.* [9, Th. 4] for the case of summable envelopes. In particular, from Lemmas 6 and 7, we have that

$$R^+(\Lambda_f^n) \leq \left[n\bar{F}_f(u_f^*(n)) \log(e) + \frac{u_f^*(n) - 1}{2} \log n \right] + 2 \leq 2 + \log(e) + \frac{u_f^*(n) - 1}{2} \log n.$$

On the other hand, it has been shown that [25]

$$\int_1^n \frac{U_f(x)}{2x} \partial x \geq \frac{U_f(n) \log n}{4} \geq \frac{(u_f^*(n) - 1)}{4} \log n. \tag{68}$$

Consequently, from Lemma 7 we have that eventually with n:

$$(1 + \xi_n) \frac{(u_f^*(n) - 1)}{4} \log n \leq R^+(\Lambda_f^n) \leq 2 + \log(e) + \frac{(u_f^*(n) - 1)}{2} \log n, \tag{69}$$

which means that $(R^+(\Lambda_f^n))_n \approx (u_f^*(n) \log n)$. Moreover, it is well-known that [3]:

$$R^+(\Lambda_f^n, \sigma(\tilde{\pi}_{k_n})) \leq R^+(\mathcal{P}^n(k_n)) \leq \frac{k_n - 1}{2} \log n + K, \tag{70}$$

for some $K > 0$, where $\mathcal{P}(k_n)$ is a short-hand for the collection all probabilities defined on the finite alphabet Γ_{k_n}, i.e., the simplex of dimension $k_n - 1$. Consequently, under the assumption that $(k_n)_n$ is $o(u_f^*(n))$, from (69) and (70) it follows that:

$$\lim_{n \to \infty} \frac{R^+(\Lambda_f^n, \sigma(\tilde{\pi}_{k_n}))}{R^+(\Lambda_f^n)} = 0.$$

$\square$

3) Proof of Theorem 6 — Regime of No-gain in Minimax Redundancy: Let us assume that $(k_n)_n \geq (u_f^*(n))_n$ eventually with n. Here we adopt results from the seminal work of Haussler and Opper [11] that offers a lower bound for the mutual information and consequently, the channel capacity that corresponds to the information radius of a family of distributions [3]. However in our problem, we have a dynamic collection of distributions, explained by the process of projecting Λ_f into the dynamic collection of sub-sigma fields $\{\sigma(\tilde{\pi}_{k_n}) : n \geq 1\}$. More precisely, and adopting the notation introduced in Section VIII-C, we have the collection of distributions:

$$\Lambda_f^n / \sigma(\tilde{\pi}_{k_n}) := \left\{ \mu^n / \sigma(\tilde{\pi}_{k_n} \times \cdots \times \tilde{\pi}_{k_n}) : \mu \in \Lambda_f \right\} \subset \mathcal{P}(\mathcal{X}^n, \sigma(\tilde{\pi}_{k_n} \times \cdots \times \tilde{\pi}_{k_n})), \tag{71}$$

for all $n \geq 1$, where $\mu/\sigma(\pi) := \{\mu(\mathcal{A}) : \mathcal{A} \in \sigma(\pi)\} \in \mathcal{P}(\mathcal{X}, \sigma(\pi))$ denotes the probability μ restricted to the sub-sigma field induced by π and $\mathcal{P}(\mathcal{X}, \sigma)$ denotes the collection of probabilities restricted to the events of the sub-sigma field $\sigma \subset \mathcal{B}(\mathcal{X})$. Furthermore, associated to $\tilde{\pi}_{k_n} = \{\mathcal{A}_{k_n,i} : i = 1, .., k_n\}$ there is a lossy mapping $\phi_n : \mathcal{X}^n \longrightarrow \Gamma_{k_n}$ where $\phi_n^{-1}(i) = \mathcal{A}_{k_n,i}$ for $i \in \Gamma_{k_n}$. Consequently through ϕ_n, every $\mu \in \mathcal{P}(\mathcal{X})$ induces a probability in Γ_{k_n}, which we denote by $\rho_{\mu,k_n} \in \mathcal{P}(\Gamma_{k_n})$, by the standard construction: $\rho_{\mu,k_n}(B) = \mu(\phi_n^{-1}(B))$ for all $B \subset \Gamma_{k_n}$. Note that ρ_{μ,k_n} is fully characterized by its pmf $f_{\rho_{\mu,k_n}}(i) = \rho_{\mu,k_n}(\{i\}) = \mu(\mathcal{A}_{k_n,i}), \forall i \in \Gamma_{k_n}$, where we have that $f_{\rho_{\mu,k_n}}(i) = f_\mu(i)$ if $i < k_n$ and $f_{\rho_{\mu,k_n}}(k_n) = 1 - \mu(\Gamma_{k_n-1})$. By letting

$$\tilde{\Lambda}_{f,k_n} := \{\rho_{\mu,k_n} : \mu \in \Lambda_f\} \subset \mathcal{P}(\Gamma_{k_n}),$$

from (51) we have that:

$$R^+(\Lambda_f^n, \sigma(\tilde{\pi}_{k_n})) = R^+(\tilde{\Lambda}_{f,k_n}^n). \tag{72}$$

---

[24]More details are presented in [11, Lem. 8, Th. 4 and Th. 5].

[25]Notice that: $\int_1^n \frac{U_f(x)}{2x} \partial x = \frac{1}{2} \int_0^{\ln n} U_f(e^y) \partial y \geq \frac{U_f(n) \ln n}{4}$, the last inequality from the concavity and positivity of $U_f(e^y)$ shown in [10, pp. 814]. On the other hand, from their definitions $\forall n \in \mathcal{X}, u_f^*(n) - 1 \leq U_f(n) < u_f^*(n)$.

Consequently, the problem reduces to characterize the information radius of a family of dynamic distributions $\left\{\tilde{\Lambda}^n_{f,k_n}, n \geq 1\right\}$ (defined on a dynamic alphabet whose size grows with the block-length). Using the envelope conditions of Λ_f and Lemma 5 in Section VIII-D, it is simple to show that $\tilde{\Lambda}_{f,k_n}$ satisfies an envelope condition on $\mathcal{P}(\Gamma_{k_n})$, more precisely:

$$\tilde{\Lambda}_{f,k_n} = \left\{\rho \in \mathcal{P}(\Gamma_{k_n}) : f_\rho(i) \leq f(i), \text{ for } i = [1 : k_{n-1}] \text{ and } f_\rho(k_n) \leq \sum_{l > k_n - 1} f(l) = \bar{F}_f(k_n - 1)\right\}. \tag{73}$$

Then, if we consider the extended (over the integer) finite size envelope function $\tilde{f}_{k_n} : \mathcal{X} \longrightarrow \mathbb{R}^+$ given by: $\tilde{f}_{k_n}(i) := f(i)$ for $i = [1 : k_n - 1]$, $\tilde{f}_{k_n}(k_n) := \bar{F}_f(k_n - 1)$ and $\tilde{f}_{k_n}(i) := 0$ for $i > k_n$, $\tilde{\Lambda}_{f,k_n}$ is equivalent to $\Lambda_{\tilde{f}_{k_n}} \subset \mathcal{P}(\mathcal{X})$ and thus,

$$R^+(\Lambda^n_f, \sigma(\tilde{\pi}_{k_n})) = R^+(\tilde{\Lambda}^n_{f,k_n}) = R^+(\Lambda^n_{\tilde{f}_{k_n}}), \ \forall n \geq 1. \tag{74}$$

Therefore, studying the minimax redundancy gain reduces to analyze the family of envelope distributions of finite size $\left\{\Lambda_{\tilde{f}_{k_n}} : n \geq 1\right\}$, where $\text{supp}(\tilde{f}_{k_n}) \subset \Gamma_{k_n}$ by construction. If we consider,

$$\epsilon^*_{n,k} := \inf \left\{\epsilon > 0 : \mathcal{H}_\epsilon(\Lambda_{\tilde{f}_{k_n}}) \leq \frac{n\epsilon^2}{8}\right\}, \tag{75}$$

the straight adoption of Lemma 8 in this dynamic context implies that

$$R^+(\Lambda^n_{\tilde{f}_{k_n}}) \geq \log(e) \cdot \mathcal{H}_{\epsilon^*_{n,k_n}}(\Lambda_{\tilde{f}_{k_n}}) - 1$$

for all n and, consequently,

$$\liminf_{n \longrightarrow \infty} R^+(\Lambda^n_{\tilde{f}_{k_n}}) \geq \log(e) \cdot \liminf_{n \longrightarrow \infty} \mathcal{H}_{\epsilon^*_{n,k_n}}(\Lambda_{\tilde{f}_{k_n}}) - 1. \tag{76}$$

Following the approach proposed by Haussler *et al.* [11], the idea is to obtain a tight approximation (ideally in closed-form) of the RHS of (76), assuming that the function $\mathcal{H}_{\epsilon^*_{n,k_n}}(\Lambda_{\tilde{f}_{k_n}})$ is asymptotically lower bounded by a continuous non-decreasing function. With that objective in mind, the following important result (Theorem 7 below) can be obtained. For the statement of this result, the following definition is needed:

Definition 25: Given $(f(x))_{x \in \mathcal{X}}$, non-negative and in $\ell_1(\mathcal{X})$, and a sequence of positive integers $(k_n)_n$, we say that $(\epsilon_n)_n \in (\mathbb{R}^+ \setminus \{0\})^\mathbb{N}$ is admissible for $(k_n)_n$ given Λ_f if

$$\bar{F}_f(k_n - 1) \leq \frac{\epsilon_n^2}{16} \tag{77}$$

holds eventually (with n).

Theorem 7: Let us consider $\Lambda_f \subset \mathcal{P}(\mathcal{X})$, with $f \in \ell_1(\mathcal{X})$ and $\text{supp}(f) = \mathcal{X}$, and a sequence of non-decreasing positive integers $(k_n)_n$ such that $(1/k_n)_n$ is $o(1)$. If $(\epsilon_{f,n})_n$ is admissible for $(k_n)_n$ given Λ_f (see Def. 25) then

$$\liminf_{n \to \infty} \frac{R^+(\Lambda^n_{\tilde{f}_{k_n}})}{\log(e) \cdot \int_1^n \frac{U_f(x)}{2x} \partial x} \geq 1$$

The proof of Theorem 7 is presented in Section VIII-F.

Remark 3: In general we have that

$$R^+(\Lambda^n_{\tilde{f}_{k_n}}) = R^+(\Lambda^n_f, \sigma(\tilde{\pi}_{k_n})) \leq R^+(\Lambda^n_f),$$

the last inequality from (74). On the other hand, it is known from [10, Th. 2] that there is a sequence $(a_n)_n$ being $o(1)$ where eventually in n

$$R^+(\Lambda^n_f) \leq (1 + a_n) \log(e) \cdot l_f(\sqrt{n}) = (1 + a_n) \log(e) \cdot \int_1^n \frac{U_f(x)}{2x} \partial x. \tag{78}$$

Consequently, under the assumptions of Theorem 7 it follows directly from this result and (78) that

$$\lim_{n \to \infty} \frac{R^+(\Lambda^n_{\tilde{f}_{k_n}})}{R^+(\Lambda^n_f)} = 1. \tag{79}$$

Returning to the proof, from Theorem 7, Definition 25 and Remark 3, a sufficient condition to obtain (79) (i.e., no gain in minimax redundancy) is that $(\sqrt{\bar{F}_f(k_n)})_n \ll (\epsilon_{f,n})_n$ (see some remarks about this in Lemma 10, Section VIII-F below). Furthermore from the proof of Theorem 7, we have that $(\epsilon_{f,n}) = (\sqrt{8/n \cdot l(1/\epsilon_{f,n})})_n \sim (\sqrt{8/n \cdot l_f(\sqrt{n})})_n = (\sqrt{8/n \cdot \int_1^n \frac{U_f(x)}{2x} \partial x})_n$, where it is known that $\int_1^n \frac{U_f(x)}{2x} \partial x \geq \frac{U_f(n) \ln n}{4}$ [10, pp.814]. From the main assumption, which consider that there is $N_o > 0$ such that $\forall n \geq N_o$, $k_n \geq u^*_f(n)$, we note that

$$\left(\frac{\bar{F}_f(k_n)}{(\epsilon_{f,n})^2}\right)_n \sim \left(\frac{n \cdot \bar{F}_f(k_n)}{\int_1^n \frac{U_f(x)}{2x} \partial x}\right)_n,$$

where for the second series we have that:

$$\frac{n \cdot \bar{F}_f(k_n)}{\int_1^n \frac{U_f(x)}{2x} \partial x} \leq \frac{4n \cdot \bar{F}_f(k_n)}{U_f(n) \ln n}$$

$$\leq \frac{4n \cdot \bar{F}_f(u_f^*(n))}{U_f(n) \ln n}$$

$$< \frac{4}{U_f(n) \ln n} \longrightarrow 0. \tag{80}$$

The strict inequality in (80) is by definition of $u_f^*(n)$ in (6), where $1/n \in (\bar{F}_f(u_f^*(n)), \bar{F}_f(u_f^*(n) - 1)]$. The last convergence in the RHS of (80) is from the fact that $U_f(n) \in [u_f^*(n) - 1, u_f^*(n)) \to \infty$ as n tends to infinity, this follows from (6) and the non-trivial assumption that $|\text{supp}(f)| = \infty$. In summary, from (80) we have that $\left(\sqrt{\bar{F}_f(k_n)}\right)_n \ll (\epsilon_{f,n})_n$, then Theorem 7 and its corollary in (79) implies that $\lim_{n\to\infty} \frac{R^+(\Lambda_{\tilde{f}_{k_n}}^n)}{R^+(\Lambda_f^n)} = 1$. This last limit and (74) conclude the proof. $\square$

F. Theorem 7

To organize the proof of Theorem 7, we present first two instrumental results:

The first result is a simple extension of [11, Lemma 8]:

Lemma 9: Let us consider the dynamic collection of distributions $\left\{\Lambda_{\tilde{f}_{k_n}} : n \geq 1\right\}$ presented in (73) where $f \in \ell_1(\mathcal{X})$, and let $(k_n)_n$ be a non-decreasing sequence of integers. In addition, let $l : \mathbb{R}^+ \to \mathbb{R}^+$ be a strictly increasing and unbounded continuous function. Let us denote by $(\epsilon_{l.n})_n$ the solutions to the expression: $l(1/\epsilon) = \frac{n\epsilon^2}{8}$ for all n. If there is a sequence $(\epsilon_n)_n$ such that:

1) $(\epsilon_n)_n \leq (\epsilon_{l.n})_n$ holds eventually with n, and
2) $\liminf_{n\to\infty} \frac{\mathcal{H}_{\epsilon_n}(\Lambda_{\tilde{f}_{k_n}})}{l(1/\epsilon_n)} \geq 1$,

then[26]

$$\liminf_{n\to\infty} \frac{R^+(\Lambda_{\tilde{f}_{k_n}}^n)}{\log(e) \cdot l(1/\epsilon_n)} \geq 1.$$

The proof is presented in Section VIII-G.

The second result characterizes a sufficient condition on $(\epsilon_n)_n$, function of $(k_n)_n$, i.e., the size sequence of tail based partitions, where the metric entropy of our collection of envelope distributions shares the same asymptotic than the unconstrained family determined in (67).

Lemma 10: Let us consider a sequence of non-negative integer $(k_n)_n$ and a sequence of non-negative reals $(\epsilon_n)_n$, where $(1/k_n)_n$ is $o(1)$ and $(\epsilon_n)_n$ is $o(1)$. If $f \in \ell_1(\mathcal{X})$ and $(\epsilon_n)_n$ is admissible for $(k_n)_n$ given Λ_f (see Def. 25) then

$$\mathcal{H}_{\epsilon_n}(\Lambda_{\tilde{f}_{k_n}}) = (1 + a_n) \int_1^{1/\epsilon_n^2} \frac{U_f(x)}{2x} \partial x$$

for a sequence $(a_n)_n$ being $o(1)$, and consequently,

$$\lim_{n\to\infty} \frac{\mathcal{H}_{\epsilon_n}(\Lambda_{\tilde{f}_{k_n}})}{l_f(1/\epsilon_n)} = \lim_{n\to\infty} \frac{\mathcal{H}_{\epsilon_n}(\Lambda_{\tilde{f}_{k_n}})}{\mathcal{H}_{\epsilon_n}(\Lambda_f)} = 1, \tag{81}$$

where $l_f(1/\epsilon) = \int_1^{1/\epsilon^2} \frac{U_f(x)}{2x} \partial x$ for $\epsilon > 0$ (see Def. 21).

The proof of this result is presented in Section VIII-H.

Comments on Lemma 10:

1) Lemma 10 establishes concrete sufficient conditions where $(\mathcal{H}_{\epsilon_n}(\Lambda_{\tilde{f}_{k_n}}))_n$ has the same asymptotic than the metric entropy of the unconstrained family $(\mathcal{H}_{\epsilon_n}(\Lambda_f))_n$, which is $\sim (\int_1^{1/\epsilon_n^2} \frac{U_f(x)}{2x} \partial x)_n$ from (67).
2) The proof of this result follows the *volume comparison arguments* proposed by Bontemps in [20, Lemmas 1 and 2].
3) Note that if $\lim_{n\to\infty} \bar{F}_f(k_n)/\epsilon_n^2 = 0$ implies that $(\epsilon_n)_n$ is admissible for $(k_n)_n$ given Λ_f.
4) Given $(k_n)_n$ and f, $\sqrt{\bar{F}_f(k_n)}$ offers a lower bound on the admissible regime for $(\epsilon_n)_n$ (see Def.25).
5) If $(\tilde{k}_n)_n \gg (k_n)_n$, i.e., $k_n/\tilde{k}_n \to 0$ as $n \to \infty$, then from Definition 25 $(\tilde{k}_n)_n$ offers a bigger admissible range for the $(\epsilon_n)_n$ than its counterpart for $(k_n)_n$.

Finally, as the asymptotic of the metric entropy in (67) offers a tight lower bound to the information radius of envelope families [10, Th.2], Lemma 10 in conjunction with Lemma 9 provide the mean to characterize a regime of no gain in minimax redundancy as presented in the proof of Theorem 7 below.

[26]In particular, if $\liminf_{n\to\infty} \frac{\mathcal{H}_{\epsilon_{l,n}}(\Lambda_{\tilde{f}_{k_n}})}{l(1/\epsilon_{l,n})} \geq 1$ then $\liminf_{n\to\infty} \frac{R^+(\Lambda_{\tilde{f}_{k_n}}^n)}{\log(e) \cdot l(1/\epsilon_{l,n})} \geq 1$.

Proof of Theorem 7: Using the hypothesis that $(\epsilon_{f,n})_n$ is admissible for $(k_n)_n$ given Λ_f, we have from Lemma 10 that as n goes to infinity:

$$\mathcal{H}_{\epsilon_{f,n}}(\Lambda_{\tilde{f}_{k_n}}) \geq (1 + o(1)) \cdot l_f(1/\epsilon_{f,n}), \tag{82}$$

which implies that

$$\liminf_{n \to \infty} \frac{\mathcal{H}_{\epsilon_{f,n}}(\Lambda_{\tilde{f}_{k_n}})}{l_f(1/\epsilon_{f,n})} \geq 1. \tag{83}$$

Note that $l_f(x) = \int_1^{x^2} \frac{U_f(\bar{x})}{2\bar{x}} \partial \bar{x}$ on $[1, \infty)$ (Def.21) is strictly increasing, continuous and unbounded, consequently applying Lemma 9 it follows that:

$$\liminf_{n \to \infty} \frac{R^+(\Lambda^n_{\tilde{f}_{k_n}})}{\log(e) \cdot l_f(1/\epsilon_{f,n})} \geq 1 \quad \Leftrightarrow \quad \liminf_{n \to \infty} \frac{R^+(\Lambda^n_{\tilde{f}_{k_n}})}{\log(e) \cdot n\epsilon_{f,n}^2/8} \geq 1, \tag{84}$$

where the last identity follows from the definition of $\epsilon_{f,n}$ in Def. 22.

At this point we use the result in [10, Proposition 3] that shows that the function $(l_f(x))$ is *very slowly variant* [10, Def. 4], in the sense that $\forall \eta \geq 0$ and $\kappa > 0$,

$$\lim_{x \to \infty} \frac{l_f(\kappa x l_f(x)^\eta)}{l_f(x)} = 1. \tag{85}$$

This slowly variant condition implies that $(\epsilon_{f,n})_n$, as a (point-wise) solution of the condition $l_f(1/\epsilon) = n\epsilon^2/8$, satisfies asymptotically (the argument is presented in the proof of [11, Theorem 5]) that:

$$\left(\epsilon_{f,n}^2/8\right)_n \sim \left(\frac{l_f(\sqrt{n})}{n}\right)_n = \left(\frac{1}{n} \int_1^n \frac{U_f(x)}{2x} \partial \bar{x}\right)_n. \tag{86}$$

Consequently (86) and (84) prove the result. $\qquad \square$

G. Proof of Lemma 9

Proof: From Lemma 8 it follows that for all n:

$$\frac{R^+(\Lambda^n_{\tilde{f}_{k_n}})}{l(1/\epsilon_{l,n})} \geq \log(e) \cdot \min\left\{\frac{\mathcal{H}_{\epsilon_{l,n}}(\Lambda_{\tilde{f}_{k_n}})}{l(1/\epsilon_{l,n})}, \frac{n\epsilon^2}{8 \cdot l(1/\epsilon_{l,n})}\right\} - \frac{1}{l(1/\epsilon_{l,n})},$$

$$= \log(e) \cdot \min\left\{\frac{\mathcal{H}_{\epsilon_{l,n}}(\Lambda_{\tilde{f}_{k_n}})}{l(1/\epsilon_{l,n})}, 1\right\} - \frac{1}{l(1/\epsilon_{l,n})}. \tag{87}$$

As $(\epsilon_{l,n}) \geq (\epsilon_n)$, without loss of generality we assume that there is a mapping $\tau : \mathbb{N} \longrightarrow \mathbb{N}$ such that $\epsilon_{l,n} = \epsilon_{\tau(n)}$ for every n, where $\tau(n) \leq n$ eventually with n. From construction $(\epsilon_{l,n})$ is $o(1)$ and thus, $l(1/\epsilon_{l,n}) \longrightarrow \infty$. Then,

$$\liminf_{n \longrightarrow \infty} \frac{R^+(\Lambda^n_{\tilde{f}_{k_n}})}{l(1/\epsilon_{l,n})} \geq \log(e) \cdot \min\left\{\liminf_{n \longrightarrow \infty} \frac{\mathcal{H}_{\epsilon_{\tau(n)}}(\Lambda_{\tilde{f}_{k_n}})}{l(1/\epsilon_{\tau(n)})}, 1\right\}$$

$$= \log(e) \cdot \min\left\{\liminf_{n \longrightarrow \infty} \frac{\mathcal{H}_{\epsilon_n}\left(\Lambda_{\tilde{f}_{k_{\tau^{-1}(n)}}}\right)}{l(1/\epsilon_n)}, 1\right\}$$

$$\geq \log(e) \cdot \left\{\liminf_{n \longrightarrow \infty} \frac{\mathcal{H}_{\epsilon_n}\left(\Lambda_{\tilde{f}_{k_n}}\right)}{l(1/\epsilon_n)}, 1\right\} \tag{88}$$

$$\geq \log(e), \tag{89}$$

where the inequality in (88) follows from the fact that $\mathcal{H}_\epsilon\left(\Lambda_{\tilde{f}_k}\right) \leq \mathcal{H}_\epsilon\left(\Lambda_{\tilde{f}_{\bar{k}}}\right)$ if $\bar{k} \geq k$ and that $\tau^{-1}(n) \geq n$, and (89) from the main hypothesis of Lemma 9. $\qquad \square$

H. Proof of Lemma 10

 Proof: Following the lower bound for $\mathcal{D}_\epsilon(\Lambda_f)$ elaborated in [20, Lemma 3], which is based on a covering argument and volume based inequality, we have that for all $m \geq 1$ and any arbitrary k such that $m + l_f \leq k$ (see Def. 2),

$$
\begin{aligned}
\mathcal{D}_\epsilon(\Lambda_{\tilde{f}_k}) &\geq \frac{\mathrm{Vol}\left(\prod_{i=l_f+1}^{l_f+m} [0, \sqrt{\tilde{f}_k(i)}]\right)}{\mathrm{Vol}\left(\mathcal{B}_m(\epsilon)\right)} \\
&= \frac{\prod_{i=l_f+1}^{l_f+m} \sqrt{\tilde{f}_k(i)}}{\epsilon^m \mathrm{Vol}(\mathcal{B}_m)},
\end{aligned}
\tag{90}
$$

where $\mathcal{B}_m(\epsilon)$ denotes the ball in $\mathbb{R}^m$ of radius ϵ, and $\mathcal{B}_m := \mathcal{B}_n(1)$. If we consider

$$
N_\epsilon := \inf \left\{ m \geq 1 : \bar{F}_f(m) < \epsilon^2/16 \right\}
$$

and let $\tilde{N}_\epsilon^k := \min \{N_\epsilon, k\}$, and we evaluate (90) with $m = \tilde{N}_\epsilon^k - l_f$ it follows that

$$
\begin{aligned}
\mathcal{H}_\epsilon(\Lambda_{\tilde{f}_k}) &= \ln \mathcal{D}_\epsilon(\Lambda_{\tilde{f}_k}) \\
&\geq \sum_{i=l_f+1}^{\tilde{N}_\epsilon^k} \ln \sqrt{\tilde{f}_k(i)} - \ln \mathrm{Vol}(\mathcal{B}_{\tilde{N}_\epsilon^k - l_f}) - (\tilde{N}_\epsilon^k - l_f) \ln 1/\epsilon.
\end{aligned}
\tag{91}
$$

On the other hand, we can adopt the upper bound in [20, Lemma 2] that is based on another volume comparison argument, leading to

$$
\begin{aligned}
\mathcal{D}_\epsilon(\Lambda_{\tilde{f}_k}) &\leq \frac{\mathrm{Vol}\left(\prod_{i=1}^{\tilde{N}_\epsilon^k} \left[-\epsilon/8, \sqrt{\tilde{f}_k(i)} + \epsilon/8\right]\right)}{\mathrm{Vol}\left(\mathcal{B}_{\tilde{N}_\epsilon^k}(\epsilon/8)\right)} \\
&= \frac{\prod_{i=1}^{\tilde{N}_\epsilon^k} \left(\sqrt{\tilde{f}_k(i)} + \epsilon/4\right)}{(\epsilon/8)^{\tilde{N}_\epsilon^k} \cdot \mathrm{Vol}(\mathcal{B}_{\tilde{N}_\epsilon^k})}.
\end{aligned}
\tag{92}
$$

This inequality reduces to [10, Eq. (6)]

$$
\begin{aligned}
\mathcal{H}_\epsilon(\Lambda_{\tilde{f}_k}) &\leq \sum_{i=1}^{l_f} \ln\left(\sqrt{\tilde{f}_k(i)} + \epsilon/4\right) + \sum_{i=l_f+1}^{\tilde{N}_\epsilon^k} \ln\left(\sqrt{\tilde{f}_k(i)}\right) \\
&\quad - \ln\left(\mathrm{Vol}\left(\mathcal{B}_{\tilde{N}_\epsilon^k}\right)\right) + \frac{\tilde{N}_\epsilon^k - l_f}{\sqrt{1 - e^{-b}}} + \tilde{N}_\epsilon^k \ln 8/\epsilon,
\end{aligned}
\tag{93}
$$

for $b = -\ln \bar{F}(l_f) > 0$.

 If we consider the regime where $N_\epsilon < k$, then it follows that $N_\epsilon = \tilde{N}_\epsilon^k$ and $\{l_f + 1, \ldots, N_\epsilon\} \subset \{1, \ldots, k-1\}$. Therefore $\tilde{f}_k(i) = f(i)$ for all $i \in \{l_f + 1, .., N_\epsilon\}$. In this scenario, we have that

$$
\mathcal{H}_\epsilon(\Lambda_{\tilde{f}_k}) \geq \sum_{i=l_f+1}^{N_\epsilon} \ln \sqrt{f(i)} - \ln \mathrm{Vol}(\mathcal{B}_{N_\epsilon - l_f}) - (N_\epsilon - l_f) \ln 1/\epsilon \text{ and}
\tag{94}
$$

$$
\mathcal{H}_\epsilon(\Lambda_{\tilde{f}_k}) \leq \sum_{i=1}^{l_f} \ln(\sqrt{f(i)} + \epsilon/4) + \sum_{i=l_f+1}^{N_\epsilon} \ln(\sqrt{f(i)}) - \ln\left(\mathrm{Vol}(\mathcal{B}_{N_\epsilon})\right) + \frac{N_\epsilon - l_f}{\sqrt{1 - e^{-b}}} + N_\epsilon \ln 8/\epsilon.
\tag{95}
$$

We point out that the RHS expression of (94) and (95) are the very same lower and upper bounds derived in [10, Eq.(7) and Eq.(6)] for $\mathcal{H}_\epsilon(\Lambda_f)$, respectively. Consequently in this regime, we obtain the lower and upper bound expressions of the unconstrained (i.e., lossless) problem.

By definition $N_\epsilon < k$ is equivalent to the condition that $\bar{F}_f(k-1) \leq \frac{\epsilon^2}{16}$. Using the hypothesis that (k_n) and (ϵ_n) are such that the condition in Eq.(77) (in Def.25) is satisfied eventually with n, it follows that

$$\liminf_{n\to\infty} \mathcal{H}_{\epsilon_n}(\Lambda_{\tilde{f}_{k_n}}) \geq \liminf_{n\to\infty} \left\{ \sum_{i=l_f+1}^{N_{\epsilon_n}} \ln \sqrt{f(i)} - \ln \text{Vol}\left(\mathcal{B}_{N_{\epsilon_n}-l_f}\right) - (N_{\epsilon_n} - l_f)\ln 1/\epsilon_n \right\}, \tag{96}$$

$$\limsup_{n\to\infty} \mathcal{H}_{\epsilon_n}(\Lambda_{\tilde{f}_{k_n}}) \leq \limsup_{n\to\infty} \left\{ \sum_{i=1}^{l_f} \ln\left(\sqrt{f(i)} + \epsilon/4\right) + \sum_{i=l_f+1}^{N_\epsilon} \ln\left(\sqrt{f(i)}\right) \right.$$
$$\left. - \ln\left(\text{Vol}(\mathcal{B}_{N_\epsilon})\right) + \frac{N_\epsilon - l_f}{\sqrt{1-e^{-b}}} + N_\epsilon \ln 8/\epsilon \right\}. \tag{97}$$

To conclude, it has been shown in [10, Prop. 4] that the RHS of both (94) and (95) behaves asymptotically as $(1 + o(1)) \int_1^{1/\epsilon^2} \frac{U_f(x)}{2x} \partial x$ when ϵ goes to zero. Consequently given that by hypothesis $\epsilon_n \longrightarrow 0$, this fact implies that $\mathcal{H}_{\epsilon_n}(\Lambda_{\tilde{f}_{k_n}}) = (1 + o(1)) \int_1^{1/\epsilon_n^2} \frac{U_f(x)}{2x} \partial x$ as n tends to infinity. Finally (81) follows from (96), (97) and [10, Prop. 4]. $\qquad\square$

IX. Acknowledgment

Appendix I
Supporting Results

A. Proof of Lemma 1

Proof: Let first prove the sufficient condition. Let us assume that $\{\pi_n : n \geq 1\}$ is asymptotically sufficient for μ. The induced distortion is given by

$$d(\phi_n, \psi_n, \mu^n) = \sum_{x\in\mathcal{X}} f_\mu(x) \cdot \underbrace{\rho(x, \psi_n(\phi_n(x)))}_{g_n(x):=}. \tag{98}$$

Considering that $\mu(\limsup_n \pi_n(x)) = f_\mu(x)$ for all $x \in \text{supp}(\mu)$, then it follows that $\lim_{n\to\infty} \psi_n(\phi_n(x)) = x$, μ-almost everywhere and $\lim_{n\to\infty} g_n(x) = 0$, μ-almost surely. Furthermore, $g_n(x)$ is a bounded function by definition, then the *bounded convergence theorem* [31] implies that $\lim_{n\to\infty} \int_{\mathcal{X}} g_n(x) d\mu(x) = 0 \Leftrightarrow \lim_{n\to\infty} d(\phi_n, \psi_n, \mu^n) = 0$.

For the converse, let us assume that $\cap_{n\geq 1} \cup_{m\geq n} \pi_m(x) \neq \{x\}$, μ-almost surely. In other words, $\exists x, x_o \in \text{supp}(\mu)$ with $x \neq x_o$ such that $\{x, x_o\} \subset \lim_{n\to\infty} \cup_{m\geq n}\pi_m(x)$. Consequently, there exists N such that for all $n \geq N$, $d(\phi_n, \psi_n, \mu^n) \geq \min\{f_\mu(x), f_\mu(x_o)\} \cdot \min\{\rho(x, x_o), \rho(x_o, x)\} > 0$. $\qquad\square$

B. Proof of Lemma 2

For the proof we need the following definitions:

Definition 26: Let us consider $\mu \in \mathcal{P}(\mathcal{X})$ and a function $g : \mathcal{X} \to \mathbb{R}$. g is said to be integrable with respect to μ if $\sum_{x\in\mathcal{X}} |g(x)| f_\mu(x) < \infty$. Finally, $\ell_1(\mu)$ denotes the collection of all integrable functions with respect to μ.

Proof: Let us consider:

$$H(\mu) - H_{\sigma(\pi_n)}(\mu) = \sum_{x\in\mathcal{X}} f_\mu(x) \cdot \underbrace{\log \frac{\mu(\pi_n(x))}{f_u(x)}}_{\tilde{g}_n(x):=}. \tag{99}$$

From the assumption that $H(\mu) < \infty$, then $\tilde{g}_n(x) \in \ell_1(\mu)$. Furthermore, under the assumption that $\{\pi_n : n \geq 1\}$ is asymptotically sufficient, we have that $\lim_{n\to\infty} \mu(\pi_n(x)) = f_\mu(x)$ for all $x \in \text{supp}(\mu)$ and thus, $\lim_{n\to\infty} \tilde{g}_n(x) = 0$ μ-almost everywhere. Finally considering that $\tilde{g}_n(x) \leq \log 1/f_\mu(x) \in \ell_1(\mu)$, the dominated convergence theorem implies that $\lim_{n\to\infty} \sum_{x\in\mathcal{X}} f_\mu(x) \cdot \tilde{g}_n(x) = 0$. $\qquad\square$

C. Proof of Proposition 1:

Proof: The argument reduces to verify the achievability of the entropy using a two-stage lossy construction. For that we consider the tail partition

$$\pi_n = \{\{1\}, \ldots, \{n\}, \{n+1, \ldots, \}\}, \tag{100}$$

associated to $\phi_n(x) = x$ if $x \in \{1, .., n\}$ and otherwise $\phi_n(x) = 0$. It is simple to verify that this scheme satisfies the zero distortion condition. For the lossless coding of $Y^n = \Phi_n(X^n) \in \{0, 1, \ldots, n\}^n$, we can consider the prefix-free *Shannon code* [4], whose rate is at most two bits away from the entropy of Y^n. Hence, there is $\mathcal{C}_n : \{0, 1, .., n\}^n \longrightarrow \{0, 1\}^*$ such that:

$$r(\phi_n, \mathcal{C}_n, \mu^n) \leq \frac{H(Y^n) + 1}{n} = \frac{H_{\sigma(\pi_n)}(\mu) + 1}{n}, \tag{101}$$

which suffices to show that

$$\limsup_{n \longrightarrow \infty} r(\phi_n, \mathcal{C}_n, \mu^n) \leq H(\mu), \tag{102}$$

and therefore $R_{al}(\mu) \leq \bar{R}_{al}(\mu) \leq H(\mu)$. $\qquad\square$

D. Proof of Lemma 5

Proof: First, it is direct to show that $\{v_\mu : \mu \in \Lambda_f\} \subset \tilde{\Lambda}_{\tilde{f}}$. Then, it remains to prove that for any $v \in \tilde{\Lambda}_{\tilde{f}}$ there is $\mu \in \Lambda_f$ such that $v_\mu = v$, in total variations. Let us fix an arbitrary $i \in \mathcal{I}$. If we first assume that $|\mathcal{A}_{n,i}| < \infty$, we propose the following approach:

$$\hat{x}_1 = \arg \min_{x \in \mathcal{A}_{n,i}} f(x), \qquad\qquad \hat{w}_{x_1} = \frac{f(\hat{x}_1)}{\sum_{x \in \mathcal{A}_{n,i}} f(x)},$$

$$\hat{x}_2 = \arg \min_{x \in \mathcal{A}_{n,i} \setminus \{\hat{x}_1\}} f(x), \qquad\qquad \hat{w}_{x_2} = \frac{f(\hat{x}_2)}{\sum_{x \in \mathcal{A}_{n,i}} f(x)},$$

$$\ldots \, \hat{x}_{|\mathcal{A}_{n,i}|} \in \mathcal{A}_{n,i} \setminus \{\hat{x}_i : i = 1, \ldots, |\mathcal{A}_{n,i}| - 1\} \text{ and} \qquad\qquad \hat{w}_{x_{|\mathcal{A}_{n,i}|}} = \frac{f(\hat{x}_{|\mathcal{A}_{n,i}|})}{\sum_{x \in \mathcal{A}_{n,i}} f(x)}. \tag{103}$$

With this we define $\mu(\{\hat{x}_i\}) = \hat{w}_{x_1} \cdot v(\{i\})$ for each $i \in \{1, \ldots, |\mathcal{A}_{n,i}|\}$. Note that $\mu(\{\hat{x}_i\}) \leq f(\hat{x}_i)$ and $\sum_{x \in \mathcal{A}_{n,i}} f_\mu(x) = v(\{i\})$ by construction. If $|\mathcal{A}_{n,i}| = \infty$ and $\sum_{x \in \mathcal{A}_{n,i}} f(x) < \infty$, we can follow the same inductive approach than in (103) to construct $\mu(\{x\})$ for all $x \in \mathcal{A}_{n,i}$. On the other hand, if $|\mathcal{A}_{n,i}| = \infty$ and $\sum_{x \in \mathcal{A}_{n,i}} f(x) = \infty$, then $\tilde{f}(i) = 1$ by definition, and we can always find $\mu_i \in \Lambda_f$ such that $\text{supp}(\mu_i) \in \mathcal{A}_{n,i}$. Then, we construct $\mu(\{x\}) = \mu_i(x) \cdot v(\{i\})$, where it is clear that $\mu(\mathcal{A}_{n,i}) = v(\{i\})$ and $f_\mu(x) \leq f(x)$ for all $x \in \mathcal{A}_{n,i}$ provided by $\mu_i \in \Lambda_f$. $\qquad\square$

E. Proposition 2

Proposition 2: Let us consider a lossy code (f_n, g_n) and a family of distributions $\Lambda \subset \mathcal{P}(\mathcal{X})$. If we denote by v_{μ^n} the probability in $\mathcal{I}_n$ induced by $\mu^n \in \Lambda^n$ (the n-fold distributions with marginal in Λ) and ϕ_n, by $v_{\mu^n}(\{i\}) = \mu^n(\phi_n^{-1}(\{i\}))$ $\forall i \in \mathcal{I}_n$, then

$$\begin{aligned}
\bar{R}(\Lambda, f_n) &= \sup_{\mu \in \Lambda} \left(r(f_n, \mu^n) - \frac{1}{n} H_{\sigma(\pi_n)}(\mu^n) \right) \\
&\geq \frac{1}{n} R^+(\{v_{\mu^n} : \mu \in \Lambda\}) \\
&= \frac{1}{n} \inf_{v \in \mathcal{P}(\mathcal{I}_n)} \sup_{\mu \in \Lambda} \mathcal{D}(v_{\mu^n}|v) \\
&= \frac{1}{n} \inf_{v \in \mathcal{P}(\mathcal{X}^n)} \sup_{\mu \in \Lambda} \mathcal{D}_{\sigma(\pi_n)}(\mu^n|v).
\end{aligned} \tag{104}$$

Proof: By definition $r(f_n, \mu^n) = \frac{1}{n} \mathbb{E}_{X_1^n \sim \mu^n} \{ \mathcal{L}(\mathcal{C}_n(\phi_n(X_1^n))) \}$. Consequently, if we let $Y_n = \phi_n(X^n)$ in $\mathcal{I}_n$ we have that $Y_n \sim v_{\mu^n}$, where v_{μ^n} denote the probability induced by μ^n and ϕ_n in $\mathcal{P}(\mathcal{I}_n)$. We will consider $r(\mathcal{C}_n, v_{\mu^n}) = \mathbb{E}_{Y_n \sim v_{\mu^n}} \{ \mathcal{L}(\mathcal{C}_n(Y_n)) \} = n \cdot r(f_n, \mu^n)$, and as $r(\mathcal{C}_n, v_{\mu^n}) \geq H(Y_n)$ [4], for the rest we focus on a refined worst-case redundancy, attributed to the second stage of f_n, given by

$$\bar{R}(\Lambda, \mathcal{C}_n) := \sup_{\mu \in \Lambda} \left(r(\mathcal{C}_n, v_{\mu^n}) - H(v_{\mu^n}) \right). \tag{105}$$

We note that $H(v_{\mu^n}) = H_{\sigma(\pi_n)}(\mu^n)$, therefore $\bar{R}(\Lambda, f_n) = \frac{1}{n}\bar{R}(\Lambda, \mathcal{C}_n)$. Considering (105) we have that

$$
\begin{aligned}
\bar{R}(\Lambda, \mathcal{C}_n) &\geq \min_{\tilde{\mathcal{C}}_n : \mathcal{I}_n \to \{0,1\}^*} \sup_{\mu \in \Lambda} \left(r(\tilde{\mathcal{C}}_n, v_{\mu^n}) - H(v_{\mu^n}) \right) \\
&\geq \inf_{v \in \mathcal{P}(\mathcal{I}_n)} \sup_{\mu \in \Lambda} \mathcal{D}(v_{\mu^n}|v) \\
&= R^+(\{v_{\mu^n} : \mu \in \Lambda\}) = \inf_{v \in \mathcal{P}(\mathcal{X}^n)} \sup_{\mu \in \Lambda} \mathcal{D}_{\sigma(\pi_n)}(\mu^n|v).
\end{aligned}
\tag{106}
$$

The first inequality in (106) is because we are solving the least worst-case redundancy (fixing the first stage of f_n), the second is from the tight connection between prefix-free mappings and probabilities in $\mathcal{P}(\mathcal{I}_n)$ and the role of the information divergence in lossless prefix-free coding [3], and the last equality is from definition of the induced probabilities in $\mathcal{P}(\mathcal{I}_n)$ and the identity in (50). We note that the expression in (106) is the information radius of our n-fold family Λ^n projected into the sub-sigma field induced by π_n, i.e., first stage of f_n. $\qquad\square$

REFERENCES

[1] J. F. Silva and P. Piantanida, "Almost lossless variable-length source coding on countably infinite alphabets," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 1–5.

[2] ——, "The redundancy gains of almost lossless universal source coding over envelope families," in *IEEE International Symposium on Information Theory*, July 2017, pp. 1–5.

[3] I. Csiszar and P. Shields, *Information Theory and Statistics: A Tutorial*. Now, 2004.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley Interscience, New York, 2006.

[5] L. Gyorfi, I. Pali, and E. van der Meulen, "There is no universal source code for an infinite source alphabet," *IEEE Transactions on Information Theory*, vol. 40, no. 1, pp. 267–271, 1994.

[6] L. D. Davisson, "Universal noiseless coding," *IEEE Transactions on Information Theory*, vol. IT-19, no. 6, pp. 783–785, November 1973.

[7] J. C. Kieffer, "A unified approach to weak universal source coding," *IEEE Transactions on Information Theory*, vol. IT-74, pp. 674–682, 1978.

[8] E. Gassiat, *Universal Coding and Order Identification by Model Selection Methods*. Springer Monographs in Mathematics, 2018.

[9] S. Boucheron, A. Garivier, and E. Gassiat, "Coding on countably infinite alphabets," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 358–373, 2009.

[10] D. Bontemps, S. Boucheron, and E. Gassiat, "About adaptive coding on countable alphabets," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 808–821, 2014.

[11] D. Haussler and M. Opper, "Mutual information, metric entropy and comulative relative entropy rsik," *The Annals of Statistics*, vol. 25, no. 6, pp. 2451–2492, 1997.

[12] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Transactions on Information Theory*, vol. 36, no. 3, pp. 453–471, May 1990.

[13] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 714–722, May 1995.

[14] ——, "Universal prediction," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2124–2147, Oct 1998.

[15] A. Barron, L. Györfi, and E. C. van der Meulen, "Distribution estimation consistent in total variation and in two types of information divergences," *IEEE Transactions on Information Theory*, vol. 38, no. 5, pp. 1437–1454, 1992.

[16] T. S. Han, "Weak variable-lenght source coding," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1217–1226, July 2000.

[17] S.-W. Ho and S. Verdú, "On the interplay between conditional entropy and error probability," *IEEE Transactions on Information Theory*, vol. 56, no. 12, pp. 5930–5942, 2010.

[18] J. F. Silva and P. Parada, "Shannon entropy convergence results in the countably infinite case," in *IEEE International Symposium on Information Theory*, IEEE, Ed., IEEE International Symposium on Information Theory. IEEE, 2012.

[19] S.-W. Ho and R. W. Yeung, "The interplay between entropy and variational distance," *IEEE Transactions on Information Theory*, vol. 56, no. 12, pp. 5906–5929, December 2010.

[20] D. Bontemps, "Universal coding on infinite alphabets: Exponentially decreasing envelopes," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1466–1478, March 2011.

[21] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh, "Poissonization and universal compression of envelope classes," in *ISIT 2014*, 2014.

[22] C. E. Shannon, "A mathematical theory of communication," *Bell System Tech, J.*, vol. 27, pp. 379–423; 623–656, July and October 1948.

[23] J. Rissanen and G. G. Langdon, "Aritmetic coding," *IBM J. Res. Develop.*, vol. 23, no. 2, pp. 149–162, 1979.

[24] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1958.

[25] T. Berger and J. D. Gibson, "Lossy source coding," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2693–2723, 1998.

[26] R. Gray, *Source Coding Theory*. Norwell, MA: Kluwer Academic, 1990.

[27] J. F. Silva, "Shannon entropy estimation in ∞-alphabets from covergence results: Studying plug-in estimators," *Entropy*, vol. 20, no. 397, pp. 1–28, 2018.

[28] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. van der Meulen, "Nonparametric entropy estimation: An overview," *Int. J. of Math. and Stat. Sci.*, vol. 6, no. 1, pp. 17–39, 1997.

[29] F. Piera and P. Parada, "On convergence properties of Shannon entropy," *Problems of Information Transmission*, vol. 45, no. 2, pp. 75–94, 2009.

[30] J. F. Silva and P. Parada, "On the convergence of Shannon differential entropy, and its connections with density and entropy estimation," *Journal of Statistical Planning and Inference*, vol. 142, no. 7, pp. 1716–1732, 2012.

[31] S. Varadhan, *Probability Theory*. American Mathematical Society, 2001.

[32] C. Anderson, "Extreme value theory for class of discrete distributions with applications to stochastic processes," *J. Appl. Probability*, vol. 7, pp. 99–113, 1970.