On Universal D-Semifaithful Coding for Memoryless Sources with Infinite Alphabets

Jorge Silva, Pablo Piantanida

On Universal D-Semifaithful Coding for Memoryless Sources with Infinite Alphabets

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

***Abstract*— The problem of variable length and fixed-distortion universal source coding (or D-semifaithful source coding) for stationary and memoryless sources on countably infinite alphabets (∞-alphabets) is addressed in this paper. The main results of this work offer a set of sufficient conditions (from weaker to stronger) to obtain weak minimax universality, strong minimax universality, and corresponding achievable rates of convergences for the worst-case redundancy for the family of stationary memoryless sources whose densities are dominated by an envelope function (or the envelope family) on ∞-alphabets. An important implication of these results is that universal D-semifaithful source coding is not feasible for the complete family of stationary and memoryless sources on ∞-alphabets. To demonstrate this infeasibility, a sufficient condition for the impossibility is presented for the envelope family. Interestingly, it matches the well-known impossibility condition in the context of lossless (variable-length) universal source coding. More generally, this work offers a simple description of what is needed to achieve universal D-semifaithful coding for a family of distributions Λ. This reduces to finding a collection of quantizations of the product space at different block-lengths — reflecting the fixed distortion restriction — that satisfy two asymptotic requirements: the first is a universal quantization condition with respect to Λ, and the second is a vanishing information radius (I-radius) condition for Λ reminiscent of the condition known for lossless universal source coding.**

***Index Terms*— Lossy compression, variable length source coding, D-semifaithful code, universal source coding, infinite alphabets, strong minimax universality, information radius, universal quantization, envelope families.**

I. Introduction

Universal Source Coding (USC) has a long history [2]–[11]. This research topic started with a series of important papers on the late sixties and early seventies by Fitingof [7], [12], Lynch [8], Davisson [13], Shtarkov and Babkin [9], Babkin [10], and Cover [11]. The main focus was proposing constructive coding methods for variable-length lossless coding under different finite-alphabet assumptions for discrete-time sources. On this early stage, the seminal work by Davisson [5] is the first that introduced an information-theoretic viewpoint for USC, formalizing different forms of redundancies and universal coding with respect to those redundancies.[1]

In lossless variable-length source coding, it is well-known that if we know the statistics of a stationary and memoryless source, the Shannon entropy of the 1D marginal of the process characterizes the minimum achievable rate [3]. However, when the statistics of the source are not known but the source belongs to a family of stationary and memoryless distributions Λ, the problem reduces to characterizing the worst-case expected overhead (or *worst-case redundancy*) that an encoder-decoder pair exhibit due to the lack of knowledge about true distribution [2], [14]. In fact, 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 [2].

The information radius (I-radius) has been richly studied by the community, and there are numerous contributions [15]–[19]. In particular, it is well-known that the I-radius grows sub-linearly for the family of finite alphabet stationary and memoryless sources [2], which implies the existence of a universal source code that achieves Shannon entropy for every distribution in this family provided that the block length tends to infinity. Unfortunately, this positive result does not extend to the case of stationary and memoryless sources on *countably infinite alphabets* (∞-alphabets) [4], [6], [15]. From an information complexity perspective, this infeasibility result means that the I-radius of this family is unbounded for any finite block-length; consequently, lossless universal source coding for ∞-alphabet stationary and memoryless sources is an intractable problem.

There has been renewed interest in USC with infinite alphabets in recent year [15], [16], [20]–[22]. Restricting the study to the case of memoryless sources with marginal densities dominated by an envelope function f (or the envelope family Λ_f), a series of new results have been presented in [15], [16], [20], [22]. Remarkably, [15, Theorems 3 and 4] show that f being summable (over the infinite alphabet) is a necessary and sufficient condition to guarantee strong minimax universality for the envelope family Λ_f. Consequently, universality can be achieved for a non-trivial (infinite dimensional) collection of distributions with infinite support. Furthermore, the specific rate of convergence for the worse-case redundancy (i.e., the information radius of Λ_f) has been derived for exponential and power law (envelope) families in ∞-alphabets as well the construction of coding schemes that achieve optimal worst-case redundancies (information limits) [16], [20], among other interesting results.

The material in this paper was partially published in the Proceedings of the 2019 IEEE International Symposium on Information Theory (ISIT) [1].

J. F. Silva is with the Information and Decision Systems (IDS) Group, University of Chile, Av. Tupper 2007 Santiago, 412-3, Room 508, Chile, Tel: 56-2-9784090, Fax: 56-2 -6953881, (email: josilva@ing.uchile.cl).

P. Piantanida is with the Laboratoire des Signaux et Systèmes (L2S), CentraleSupélec, CNRS, Université Paris-Saclay, 91190 Gif-sur-Yvette, France (email: pablo.piantanida@centralesupelec.fr).

[1] A nice presentation of the early history of universal coding can be found in [5, Section II] and reference therein.

Complementing the previous results on infinity alphabet sources and using ideas from weak source coding by Han [23], almost lossless universal source coding was introduced in [22], [24]. The general idea of this approach is to relax the lossless assumption by introducing a non-zero distortion that tends to zero with the block-length (asymptotic zero distortion), with the intention of achieving weak universality over the entire collection of memoryless sources on ∞-alphabets [21], [22]. Results in this weak setting demonstrate that almost lossless USC is feasible for the entire family of stationary and memoryless distributions [22, Th. 4] on ∞-alphabets, and the sensitive role that the vanishing distortion plays on the analysis of the problem when moving from a point-wise to a uniform convergence to zero [22, Th. 5].

A. Contributions

In this paper, we investigate the problem introduced by Ornstein and Shields in [25] of fixed-distortion and variable length universal source coding—or universal D-semifaithful coding—for ∞-alphabet sources. Following the line of work of the seminal paper by Boucheron *et al.* [15], among others [16], [20], [22], we study the family of stationary and memoryless sources whose probability mass functions are dominated by an envelope function f by adopting the criterion of strong minimax universality [2]. The redundancy in this case should be measured with respect to the rate-distortion function lower bound [3], [26], [27]. Our main results (cf. Theorems 2 and 3) parallel the results presented in the lossless problem [15, Theorems 3 and 4] and offer a set of conditions on the envelope function to obtain weak minimax universality, strong minimax universality as well as an achievable rate of convergence for the worse-case redundancy. Conversely, Theorem 2 shows that if the envelope function is not summable, then strong minimax universality is not feasible, i.e., an impossible result. Indeed, this result matches the infeasibility condition known for the case of lossless USC [15]. More generally, we present a simple result that captures what is needed (necessary and sufficient conditions) to achieve universal D-semifaithful source coding in terms of some asymptotic properties imposed on a collection of partitions of the source alphabet (Lemma 1).

A central technical contribution of this paper relies on the derivation of a lower bound for the minimax redundancy of a D-semifaithful code, operating at a given distortion level, which is obtained using a redefined expression of the I-radius for the family of sources. The resulting I-radius expression is based on the information divergence restricted to quantization cells (or bins) induced by the D-semifaithful code. This lower bound represents the central ingredient to derive the impossibility argument over envelope families. On the other hand, achievable results are obtained for summable envelope functions, similarly to the case of lossless source coding [15], [16]. For this, a two-stage constructive coding scheme is employed (operating at a fixed distortion) for which results are adopted from universal D-semifaithful coding on finite alphabets (Lemma 5) and universal lossless source coding on ∞-alphabets [15], [16]. To the best of our knowledge, our results are the first that explore universal D-semifaithful coding for stationary and memoryless sources on ∞-alphabets using the criterion of strong minimax universality. A preliminary version of this paper was presented in [1] where some of the results were introduced without a complete presentation of their proofs.

B. Related Work on Universal D-semifaithful for Finite Alphabet Sources

Relevant results on universal D-semifaithful coding have been presented for finite alphabet sources [25], [28], [29]. Ornstein and Shields [25] proposed a universal D-semifaithful code for finite alphabet ergodic sources deriving almost-sure convergence of the rate of the code to the rate-distortion function (a sample-wise analysis). Complementing this analysis, Yu and Speed [28] proposed a two-stage universal D-semifaithful code for the family of finite alphabet stationary and memoryless sources with some added regularity conditions. They showed that the average rate of this D-semifaithful code achieves (uniformly over this family) the rate-distortion function at a rate of convergence that is $O(n^{-1}\log n)$. On the optimality of this last constructive result, it is showed in [30] that the rate $O(n^{-1}\log n)$ is optimal for the *Hamming* distortion measure. This optimality was showed more generally in [31], [32] and they also presented new schemes that achieve the optimal rate of convergence of $O(n^{-1}(\log n + o(\log n)))$ for finite alphabet stationary and memoryless sources. Results of the same nature were obtained in [33].

Revisiting the sample-wise redundancy analysis of lossy source coding operating at a fixed distortion, Kontoyiannis [29] showed that the best (sample-wise) redundancy rate (in bits per sample) of a code that knows the model is $O(1/\sqrt{n})$ (a converse result). The analysis was then extended to a universal setting, where for finite alphabet memoryless sources the same redundancy rate (sample-wise) of $O(1/\sqrt{n})$ is shown. Surprisingly in terms of sample wise redundancy, this work showed that no penalization is observed when moving from an optimal code that knows the model to a universal setting for finite alphabet memoryless sources. This matching condition on the sample-wise redundancy is non-observed when the analysis is based on the average redundancy of a code [2].

Finally, adopting a refined angle for the sample-wise redundancy analysis, Kontoyiannis and Zhang [34] proposed a lossy version of the Kraft inequality to connect D-semifaithful codes (operating at a fixed distortion) with probability measures (in the reproduction alphabet) [34, Theorem 1]. This novel connection (codes-measures correspondence in lossy compression) leads to an alternative non-asymptotic formulation of the rate-distortion question in the form of an optimization problem. More precisely, they introduced the optimal code length valid for any finite block length [34, Eq.(2)], which can be interpreted as the best code length (probability on the reproduction alphabet) for the D-semifaithful coding task, see [29, Corollary 1]. Using this optimal code-length representation (non-asymptotic and sample-wise), the problem of universal D-semifaithful coding is addressed by looking at the redundancy of a universal D-semifaithful code with respect to the best code length (a sample-wise redundancy analysis).

Following ideas used in lossless compression [18], the authors explored the performance of a mixture codebook (generated by a mixture of i.i.d. distributions using a prior probability over the simplex). Sufficient conditions on the prior were developed to show that the proposed mixture codebook scheme is universal over the class of finite alphabet memoryless sources [34, Th.6]. Furthermore, some regularities on the prior were established in [34, Th.7] guaranteeing that the mixture codebook offers a universal D-semifaithful code with a sample-wise redundancy, relative to the mentioned optimal code length, that does not exceed $(d-1)/2\log(n)/n$ bits per sample. This last result is obatined almost surely when the block length tends to infinity where d denotes the cardinality of the reproduction alphabet. The achieved $1/2{\cdot}\log(n)/n$ sample-wise redundancy per degrees of freedom and per sample agrees with results known for lossless universal source coding: the minimax average redundancy is $O((d-1)\log(n)/n)$ for the class of finite alphabet memoryless sources [2], [14]. In addition, known converse results for D-semifaithful coding in [30]–[32] might indicate that this mixture codebook scheme has an optimal (sample-wise) redundancy per dimension.

C. Paper Organization

The rest of the paper is organized as follows. Section II introduces some definitions and basic elements for the formalization of the problem. Section III presents the universal D-Semifaithful source coding problem and introduces a general result (Lemma 1). Section IV presents results for the family of envelope distributions (Theorems 2 and 3). Final remarks and directions for future work are presented in Section V. The arguments used to prove the main results, Theorems 2 and 3 are presented in Section VII. Finally, supporting results and technical derivations are relegated to the Appendix sections.

II. Main Definitions and Preliminaries

Let us denote by $\mathbb{X}$ a countably infinite alphabet: the integers without loss of generality. The space is equipped with a distortion function $\rho : \mathbb{X} \times \mathbb{X} \longrightarrow \mathbb{R}^+$, and the non-trivial scenario is assumed where $\rho(x,\bar{x}) > 0$ if, and only if, $\bar{x} \neq x$. For any $n \geq 1$, we have $\rho_n : \mathbb{X}^n \times \mathbb{X}^n \longrightarrow \mathbb{R}^+$ of block length n to be the standard single letter construction obtained from ρ [26], [27], where for any $x^n = (x_1,..,x_n)$ and $\bar{x}^n = (\bar{x}_1,..,\bar{x}_n)$ in $\mathbb{X}^n$

$$\rho_n(x^n,\bar{x}^n) \equiv \frac{1}{n}\sum_{i=1}^{n}\rho(x_i,\bar{x}_i). \tag{1}$$

A D-semifaithful code of length n operating at a distortion $d > 0$ is a variable length coding scheme operating at a fixed distortion [25], [29]. More precisely, we consider the following definition:

Definition 1: A D-semifaithful code of length n operating at distortion $d > 0$ is defined/denoted by a triplet $\xi_n = (\phi_n, \mathcal{C}_n, \mathcal{D}_n)$, where

- $\phi_n : \mathbb{X}^n \longrightarrow \mathcal{B}_n \subset \mathbb{X}^n$ is a quantizer,
- $\mathcal{C}_n : \mathcal{B}_n \longrightarrow \{0,1\}^* \equiv \cup_{k\geq 1}\{0,1\}^k$ is a lossless binary (variable length) encoder, and
- $\mathcal{D}_n : \{0,1\}^* \longrightarrow \mathcal{B}_n$ is a binary decoder,

satisfying that for any $x^n \in \mathbb{X}^n$

$$\rho_n(x^n, \phi_n(x^n)) \leq d. \tag{2}$$

The set $\mathcal{B}_n = \{\phi_n(x^n), x^n \in \mathbb{X}^n\}$ contains the prototypes of ξ_n in $\mathbb{X}^n$. In this construction, the binary encoder $\mathcal{C}_n$, which is variable length, is prefix-free [3] implying that it satisfies the *Kraft-MacMillan inequality*:

$$\sum_{i\in\mathcal{B}_n} 2^{-\mathcal{L}(\mathcal{C}_n(i))} \leq 1, \tag{3}$$

where $\mathcal{L} : \{0,1\}^* \longrightarrow \mathbb{N}\setminus\{0\}$ is the function that returns the length (number of bits) of a vector in $\{0,1\}^*$.

Importantly for the analysis presented in this paper, the code $\xi_n = (\phi_n, \mathcal{C}_n, \mathcal{D}_n)$ induces a partition in $\mathbb{X}^n$ given/denoted by

$$\pi_{\phi_n} \equiv \left\{\mathcal{A}_{n,y^n} \equiv \phi_n^{-1}(\{y^n\}), y^n \in \mathcal{B}_n\right\} \subset 2^{\mathbb{X}^n}. \tag{4}$$

A. The Source Coding Problem

Let us consider an information source (a random sequence) $X = (X_n)_{n\geq 1}$ with values in $\mathbb{X}$ and process distribution denoted by $\mu = \{\mu_n \in \mathcal{P}(\mathbb{X}^n), n \geq 1\}$, where for any $n \geq 1$ $X^n = (X_1,..,X_n) \sim \mu_n$, and $\mathcal{P}(\mathbb{X}^n)$ denotes the collection of probabilities in $\mathbb{X}^n$. Then, the rate (in bits per sample) for encoding X^n with a D-semifaithful code ξ_n of length n operating at distortion $d > 0$ is given by

$$R(\xi_n,\mu_n) \equiv \frac{1}{n}\mathbb{E}_{X^n\sim\mu_n}\left\{\mathcal{L}(\mathcal{C}_n(\phi_n(X^n)))\right\}. \tag{5}$$

Using the source model μ, the variable length fixed distortion lossy source coding problem reduces to minimizing $R(\xi,\mu_n)$ in (5) over the family of D-semifaithful codes (operating at distortion d) for any $n \geq 1$ [6], [35]. It is well-known that for any D-semifaithful code ξ_n [2], [3]

$$nR(\xi,\mu_n) \geq H(v_{\mu_n}), \tag{6}$$

where v_{μ_n} denotes the probability induced by μ_n and ϕ_n in the reproducible alphabet $\mathcal{B}_n$, i.e., $v_{\mu_n}(y^n) = \mu_n(\phi_n^{-1}(\{y^n\}))$ for any $y^n \in \mathcal{B}_n$, and

$$H(v_{\mu_n}) \equiv -\sum_{y^n\in\mathcal{B}_n} v_{\mu_n}(y^n)\log(v_{\mu_n}(y^n)) \tag{7}$$

is the *Shannon entropy* of $v_{\mu_n} \in \mathcal{P}(\mathcal{B}_n)$ [3], [27] and the log function is base 2. Furthermore, fixing ϕ_n (the quantizer) and optimizing over the encoder-decoder pairs $(\mathcal{C}_n, \mathcal{D}_n)$ (the prefix-free mappings from $\mathcal{B}_n$ to $\{0,1\}^*$), we have that [3], [27]

$$\frac{H(v_{\mu_n})+1}{n} \geq \min_{\mathcal{C}_n} R((\phi_n,\mathcal{C}_n,\mathcal{D}_n),\mu_n) \geq \frac{H(v_{\mu_n})}{n}, \tag{8}$$

considering in this last optimization that $\mathcal{D}_n$ is a deterministic function of $\mathcal{C}_n$.

A convenient way to write the entropy of the induced distribution v_{m_n} in (8) is as the entropy of μ_n but projected over quantization (or a sub-sigma field of the measurable space $(\mathbb{X}^n, 2^{\mathbb{X}^n})$). Given a partition $\pi = \{A_i, i \in \mathcal{I}\}$ (countable or

finite) of $\mathbb{X}^n$ and a probability $\mu \in \mathcal{P}(\mathbb{X}^n)$, we introduce the entropy of μ restricted over the sub-sigma field $\sigma(\pi)$ by

$$H_{\sigma(\pi)}(\mu) \equiv -\sum_{i\in\mathcal{I}} \mu(A_i)\log\mu(A_i) \leq H(\mu) = -\sum_{x^n\in\mathbb{X}^n} \mu(x^n)\log\mu(x^n), \quad (9)$$

where the last inequality follows from basic information inequalities [3]. Then, $H(v_{\mu_n})$ is equal to $H_{\sigma(\pi_{\phi_n})}(\mu_n)$ and (8) can be re-written by

$$\frac{H_{\sigma(\pi_{\phi_n})}(\mu_n)+1}{n} \geq \min_{\mathcal{C}_n} R((\phi_n,\mathcal{C}_n,\mathcal{D}_n),\mu_n) \geq \frac{H_{\sigma(\pi_{\phi_n})}(\mu_n)}{n}. \quad (10)$$

From (10), the source coding (operational) problem is

$$R_n(d,\mu_n) \equiv \min_{\xi_n} R(\xi_n,\mu_n), \quad (11)$$

where ξ_n is running over the family of D-semifaithful codes of length n operating at distortion d (Def.1). This operational problem can be considered equivalent to solve[2]

$$\mathcal{R}_n(d,\mu_n) \equiv \min_{\pi\in\mathcal{Q}_n(d)} \frac{H_{\sigma(\pi)}(\mu_n)}{n}, \quad (12)$$

where $\mathcal{Q}_n(d)$ denotes the collection of partitions of $\mathbb{X}^n$ where any π in $\mathcal{Q}_n(d)$ satisfies that $\forall A\in\pi$, $\exists y^n \in A$ such that

$$\sup_{x^n\in A} \rho_n(x^n,y^n) \leq d,$$

i.e., any $\pi \in \mathcal{Q}_n(d)$ offers a d-covering of $\mathbb{X}^n$ with respect to ρ_n.

For stationary and memoryless sources, it is well known that $\mathcal{R}_n(d,\mu_n)$ converges (as n tends to infinity) to the celebrated rate-distortion function [3], [27], which is a function of $\mu_1 \in \mathcal{P}(\mathbb{X})$ [6], [35]. For completeness, we briefly revisit this result here.

B. The Source Coding Theorem

Let us consider $(X_n)_{n\geq 1}$ to be a stationary and memoryless source characterized by $\mu_1 \in \mathcal{P}(\mathbb{X})$. The rate distortion function of $\mu=\{\mu_n, n\geq 1\}$ relative to ρ is given by [6]:

$$\inf_{n\geq 1} \mathcal{R}^*(d,\mu_n) = \lim_{n\to\infty} \mathcal{R}^*(d,\mu_n), \quad (13)$$

where

$$\mathcal{R}^*(d,\mu_n) \equiv \frac{1}{n}\inf_{\mathbf{U},\mathbf{V}} I(\mathbf{U};\mathbf{V}). \quad (14)$$

The infimum in (14) is taken with respect to the collection of joint random vectors $(\mathbf{U},\mathbf{V})$ in $\mathbb{X}^n\times\mathbb{X}^n$ satisfying that $\mathbf{U}\sim\mu_n$ and $\mathbb{P}(\rho_n(\mathbf{U},\mathbf{V})\leq d)=1$ [6]. By the definitions of these objects, it is simple to verify that $R_n(d,\mu_n)\geq\mathcal{R}_n(d,\mu_n)\geq\mathcal{R}^*(d,\mu_n)$ for any $n\geq 1$. Importantly, Kieffer showed that

THEOREM 1: (Kieffer [6, Th. 4]) For a D-semifaithful coding problem operating at distortion $d>0$,

$$\lim_{n\to\infty}\mathcal{R}_n(d,\mu_n) = \lim_{n\to\infty}\mathcal{R}^*(d,\mu_n) = \mathcal{R}^*(d,\mu_1). \quad (15)$$

The last expression in (15) is the single letter information theoretic limit of this problem [6]. The expression in (15) is equal to the standard rate distortion function of the source that uses the expected distortion criterion [3], [26].

[2]Up to a discrepancy of at most $1/n$ in bits per sample: $\mathcal{R}_n(d,\mu_n)\leq R_n(d,\mu_n)\leq\mathcal{R}_n(d,\mu_n)+\frac{1}{n}$.

III. A Soft Result on Universal D-Semifaithful Coding

In universal source coding, the objective is to find a coding scheme that achieves the performance limit in (15) without knowledge of the underlying source distribution [2], [14]. To formalize this problem in the context of D-semifaithful coding, let $(X_n)_{n\geq 1}$ be a stationary and memoryless source with values in $\mathbb{X}$, where we impose that μ_1 belongs to $\Lambda\subset\mathcal{P}(\mathbb{X})$. Let $\{\xi_n=(\phi_n,\mathcal{C}_n,\mathcal{D}_n), n\geq 1\}$ be a D-semifaithful coding scheme operating at distortion $d>0$ with respect to the single letter distortions $\{\rho_n, n\geq 1\}$. Following the definitions used in universal lossless source coding [5], we say that

Definition 2: A coding scheme $\{\xi_n, n\geq 1\}$ (operating at distortion $d>0$) is strongly minimax universal for Λ at distortion d if

$$\lim_{n\to\infty} \underbrace{\sup_{\mu^n\in\Lambda^n}}_{\text{worst-case redundancy over } \Lambda^n \text{ of } \xi_n} [R(\xi_n,\mu^n)-\mathcal{R}_n(d,\mu^n)] = 0, \quad (16)$$

where $\Lambda^n \equiv \{\mu^n, \mu\in\Lambda\}\subset\mathcal{P}(\mathbb{X}^n)$, and μ^n is the product (i.i.d.) distribution induced by $\mu\in\mathcal{P}(\mathbb{X})$.

By definition of $\mathcal{R}_n(d,\mu^n)$ in (12), we have that $R(\xi_n,\mu^n)-\mathcal{R}_n(d,\mu^n)\geq 0$ and, consequently, this last expression can be interpreted as the redundancy (in bits per sample) we have to accept for not knowing the distribution of X^n and using a distribution independent lossy encoder. Therefore, if $\{\xi_n, n\geq 1\}$ is strongly minimax universal, it means that as the block length tends to infinity (and uniformly over the family of hypotheses in Λ), the scheme achieves the best performance obtained by a scheme that knows the distribution of the source previous to encoding. Similarly, we can use the following definition:

Definition 3: A scheme $\{\xi_n, n\geq 1\}$ (operating at distortion $d>0$) is weakly minimax universal for Λ at distortion d if [5],

$$\lim_{n\to\infty}[R(\xi_n,\mu^n)-\mathcal{R}_n(d,\mu^n)] = 0, \quad \forall\mu\in\Lambda. \quad (17)$$

In contrast to Definition 2, being weakly minimax universal imposes a point-wise convergence of the redundancy over the collection of hypotheses in Λ.

In this redundancy analysis, we use $\mathcal{R}_n(d,\mu^n)$ instead of the asymptotic limit $\mathcal{R}^*(d,\mu_1)$ stated in Theorem 1. Indeed, $\mathcal{R}_n(d,\mu^n)$ is a tighter finite-length lower bound (a non-asymptotic performance limit) of the optimal rate (in bits per sample) that can be achieved in the operational problem in (11) when the distribution of the source is available[3]. For the lossless case (i.e., $d=0$), $\mathcal{R}_n(d,\mu^n)$ reduces to the entropy of μ and, consequently, $R(\xi_n,\mu^n)-\mathcal{R}_n(d,\mu^n)$ can be seen also as an extension of the redundancy term used for the analysis of universal variable length lossless source coding [3], [5], [14].

Before we move to the presentation of the main context of study of this work, we present a general analysis for the worst-case redundancy in (16).

[3]$\mathcal{R}_n(d,\mu^n)\geq\mathcal{R}^*(d,\mu^n)\geq\inf_{n\geq1}\mathcal{R}^*(d,\mu_n)=\mathcal{R}^*(d,\mu_1)$, the last equality from Theorem 1.

A. Minimax Redundancy Analysis

Let $\xi_n = (\phi_n, \mathcal{C}_n, \mathcal{D}_n)$ be a D-semifaithful code of length n operating at distortion $d > 0$, and μ be a distribution in $\Lambda \subset \mathcal{P}(\mathbb{X})$. Then, the average redundancy of ξ_n (in bits per sample) can be expressed by

$$R(\xi_n, \mu^n) - \mathcal{R}_n(d, \mu^n) = \left[R(\xi_n, \mu^n) - \frac{H_{\sigma(\pi_{\phi_n})}(\mu^n)}{n}\right] + \left[\frac{H_{\sigma(\pi_{\phi_n})}(\mu^n)}{n} - \mathcal{R}_n(d, \mu^n)\right], \tag{18}$$

where π_{ϕ_n} is the partition of $\mathbb{X}^n$ induced by ϕ_n (see Eq.(4)) for the n-fold distribution induced by μ. In particular, the first term on the right-hand side (RHS) of (18) is non-negative from (10) and the second term is non-negative from the definition in (12).

1) The Projected Information Radius of Λ^n with Respect to π_{π_n}: For the moment, let us concentrate on the analysis of $\left[R(\xi_n, \mu^n) - H_{\sigma(\pi_{\phi_n})}(\mu^n)/n\right]$ in (18). From a well-known correspondence between distributions and prefix-free codes [3], the encoder $\mathcal{C}_n$ can be associated with a distribution $v_{\mathcal{C}_n} \in \mathcal{P}(\mathcal{B}_n)$[4] and $R(\xi_n, \mu^n) - H_{\sigma(\pi_{\phi_n})}(\mu^n)/n$ is lower bounded by

$$\frac{1}{n} D(v_{\mu^n} \| v_{\mathcal{C}_n}) = \frac{1}{n} \sum_{y^n \in \mathcal{B}_n} v_{\mu^n}(y^n) \log \frac{v_{\mu^n}(y^n)}{v_{\mathcal{C}_n}(y^n)} \geq 0,$$

where $v_{\mu^n} \in \mathcal{P}(\mathcal{B}_n)$ is a short-hand for the distribution induced by μ^n and ϕ_n in the reproducible space $\mathcal{B}_n$. Then, we can adopt a lower bound of the worst-case (over Λ) discrepancy by the expression

$$R_n^+(\Lambda, \xi_n) \equiv \frac{1}{n} \sup_{\mu \in \Lambda} D(v_{\mu^n} \| v_{\mathcal{C}_n}) \geq 0, \tag{19}$$

where $\xi_n = (\phi_n, \mathcal{C}_n, \mathcal{D}_n)$. For the rest of the analysis, we fix the quantization ϕ_n (i.e., $\mathcal{B}_n$ and its associated partition π_{ϕ_n}), and we optimize the prefix-free mapping (denoted by $\mathcal{C}_n$) from $\mathcal{B}_n$ to $\{0,1\}^*$ with respect to the lower bound divergence term in (19)[5]. The solution of this problem introduces the information radius of the family Λ^n projected over the sigma field induced by the partition π_{ϕ_n} [2]. More precisely, we obtain the following:

$$\min_{\mathcal{C}_n} \sup_{\mu \in \Lambda} \left[R(\xi_n = (\phi_n, \mathcal{C}_n, \mathcal{D}_n), \mu^n) - \frac{H_{\sigma(\pi_{\phi_n})}(\mu^n)}{n}\right]$$

$$\geq \min_{\mathcal{C}_n} R_n^+(\Lambda, \xi_n = (\phi_n, \mathcal{C}_n, \mathcal{D}_n)) \tag{20}$$

$$\geq \frac{1}{n} R^+(\Lambda^n, \sigma(\pi_{\phi_n})), \tag{21}$$

where

$$R^+(\Lambda^n, \sigma(\pi_{\phi_n})) \equiv \min_{v \in \mathcal{P}(\mathcal{B}_n)} \sup_{\mu^n \in \Lambda^n} D(v_{\mu^n} \| v) = \min_{v \in \mathcal{P}(\mathbb{X}^n)} \sup_{\mu^n \in \Lambda^n} D_{\sigma(\pi_\phi)}(\mu^n \| v). \tag{22}$$

The inequality in (21) follows from the observation that $\{v_{\mathcal{C}_n}; \mathcal{C}_n \text{ is prefix-free mapping}\} \subset \mathcal{P}(\mathcal{B}_n)$. The expression in (22) is the information radius of Λ^n projected on π_{ϕ_n} that is written in terms of the divergence between distributions on the original sample space $\mathbb{X}^n$ but restricted over the cells of π_{ϕ_n} using that

$$D_{\sigma(\pi)}(\mu \| v) \equiv \sum_{A \in \pi} \mu(A) \log \frac{\mu(A)}{v(A)} \leq D(\mu \| v), \tag{23}$$

for any π partition of $\mathbb{X}^n$ and $\mu, v \in \mathcal{P}(\mathbb{X}^n)$. Finally, the lower bound in (21) is tight (up to a discrepancy of $1/n$). More precisely, we have that[6]

$$\min_{\mathcal{C}_n} \sup_{\mu \in \Lambda} \left[R((\phi_n, \mathcal{C}_n, \mathcal{D}_n), \mu^n) - \frac{H_{\sigma(\pi_{\phi_n})}(\mu^n)}{n}\right] \leq \frac{1}{n}(R^+(\Lambda^n, \sigma(\pi_{\phi_n})) + 1). \tag{24}$$

In summary for a fixed quantizer ϕ_n, optimizing the second-stage (over the collection of prefix-free encoder-decoder pairs) reduces to the information radius problem in (22). This problem finds the distribution that is closest to the entire family Λ^n (or the centroid of the family) using the divergence restricted over the sub-sigma field $\sigma(\pi_{\phi_n})$ in (23). Naturally, this is the same information radius characterization used in universal (variable length) lossless source coding [2].

2) Universal Quantization over Λ^n: Let us now concentrate on the analysis of the other term in (18)

$$\left[H_{\sigma(\pi_{\phi_n})}(\mu^n)/n - \mathcal{R}_n(d, \mu^n)\right],$$

which depends exclusively on the quantizer ϕ_n (or equivalently on $\pi_{\phi_n} \in \mathcal{Q}_n(d)$, see (12)). Then considering the universal setting, we can optimize $\pi_{\phi_n} \in \mathcal{Q}_n(d)$ in the following worst-case sense:

$$\min_{\bar{\pi} \in \mathcal{Q}_n(d)} \sup_{\mu^n \in \Lambda^n} \left[H_{\sigma(\bar{\pi})}(\mu^n) - \min_{\pi \in \mathcal{Q}_n(d)} H_{\sigma(\pi)}(\mu^n)\right]. \tag{25}$$

This problem can be interpreted as the universal minimax counterpart of the problem presented in (12).

B. Strong-Minimax Universality

From the analysis made on the two terms in (18), it is observed that everything reduces to the selection of the first-stage of the encoding process (the quantization). The following result formalizes this observation:

LEMMA 1: A necessary and sufficient condition for the existence of a strongly universal D-semifaithful code operating at distortion $d > 0$ for Λ (Def. 2) is that there is a sequence of partitions $\{\pi_n, n \geq 1\}$ satisfying the following:

i) $\pi_n \in \mathcal{Q}_n(d)$ for all $n \geq 1$, (the fixed distortion requirement)
ii) $\lim_{n\to\infty} \frac{1}{n} R^+(\Lambda^n, \sigma(\pi_n)) = 0$, and
iii) $\lim_{n\to\infty} \frac{1}{n} \sup_{\mu^n \in \Lambda^n} \left[H_{\sigma(\pi_n)}(\mu^n) - \min_{\pi \in \mathcal{Q}_n(d)} H_{\sigma(\pi)}(\mu^n)\right] = 0$.

From this result achieving strong minimax universality for Λ at distortion d requires meeting two conditions: on

[4] $v_{\mathcal{C}_n}(y^n) = 2^{-\mathcal{L}(\mathcal{C}_n(y^n))}/C$ for all $y^n \in \mathcal{B}$, where $C = \sum_{y^n \in \mathcal{B}_n} 2^{-\mathcal{L}(\mathcal{C}_n(y^n))} \leq 1$ [3].

[5] $\mathcal{D}_n$ is a deterministic function of $\mathcal{C}_n$ and, therefore, it is omitted in the optimization in (20).

[6] For completeness, the argument to derive (24) is presented in Appendix VI-F.

the one hand, that a universal quantizer can be found that approximates the best performance stated in (12) as the block-length tends to infinity (the approximation criterion in iii), and, on the other hand, that the resulting information radius of the projected family grows sub-linearly with the block-length (the complexity criterion in ii). This result captures the information radius condition known in the lossless universal source coding problem but adds another component making the problem conceptually more difficult to address, which is the existence of a universal quantization for the family $\{\Lambda^n, n \geq 1\}$ in the sense of condition iii).

In this fixed-distortion setting, we could move to the extreme of asking for a zero distortion ($d = 0$), where for any reasonable distortion, the quantizer ϕ_n needs to be the identity to meet the distortion criterion in i). In this context, condition iii) is trivially met and minimax universality reduces to verifying the information radius condition of the un-projected family, i.e., $R^+(\Lambda^n) = \min_{v \in \mathcal{P}(\mathbb{X}^n)} \sup_{\mu^n \in \Lambda^n} D_{\sigma(\pi_\phi)}(\mu^n \| v)$. Then, in the zero distortion regime, Theorem 1 recovers the necessary and sufficient condition known for lossless minimax universal source coding [2], [14], [15].

In the next section, we will use these conditions implicitly and explicitly to study strong minimax universality for the family of envelope distributions on infinite alphabets.

C. Proof of Lemma 1

Proof: For the direct part, for any $n \geq 1$ and $d > 0$, let us consider a lossy code $\xi_n^* = (\phi_n^*, \mathcal{C}_n^*, \mathcal{D}_n^*)$ of length n such that ϕ_n^* is determined from π_n, i.e. $\pi_{\phi_n^*} = \pi_n$. From this, ξ_n is a D-semifaithful code operating at distortion d from the assumption that $\pi_n \in \mathcal{Q}_n(d)$.[7] For the second stage (the variable length encoder-decoder of $\mathcal{B}_n$), let us consider the pairs $(\mathcal{C}_n^*, \mathcal{D}_n^*)$ as a solution of the minimax operational problem presented in the LHS of (20), i.e.,

$$\min_{\mathcal{C}_n} \sup_{\mu \in \Lambda} \left[R((\phi_n^*, \mathcal{C}_n, \mathcal{D}_n), \mu^n) - \frac{H_{\sigma(\pi_{\phi_n^*})}(\mu^n)}{n} \right].$$

Then we know from (24) that

$$\sup_{\mu \in \Lambda} \left[R(\xi_n^*, \mu^n) - \frac{H_{\sigma(\pi_{\phi_n^*})}(\mu^n)}{n} \right] \leq \frac{1}{n} R^+(\Lambda^n, \sigma(\pi_{\phi_n^*})) + \frac{1}{n} = \frac{1}{n} R^+(\Lambda^n, \sigma(\pi_n)) + \frac{1}{n}. \quad (26)$$

[7] To achieve this, it is sufficient to have that $y^n \in \phi_n^{*-1}(\{y^n\})$ for any $y^n \in \mathcal{B}_n$.

Finally using (18), it follows that

$$\begin{aligned}
&\sup_{\mu \in \Lambda} [R(\xi_n^*, \mu^n) - \mathcal{R}_n(d, \mu^n)] \\
&\leq \sup_{\mu \in \Lambda} \left[R(\xi_n^*, \mu^n) - \frac{H_{\sigma(\pi_{\phi_n^*})}(\mu^n)}{n} \right] \\
&+ \sup_{\mu \in \Lambda} \left[\frac{H_{\sigma(\pi_{\phi_n^*})}(\mu^n)}{n} - \mathcal{R}_n(d, \mu^n) \right] \\
&\leq \frac{1}{n} \left(R^+(\Lambda^n, \sigma(\pi_n)) + 1 \right) \\
&+ \frac{1}{n} \sup_{\mu^n \in \Lambda^n} \left[H_{\sigma(\pi_n)}(\mu^n) - \min_{\pi \in \mathcal{Q}_n(d)} H_{\sigma(\pi)}(\mu^n) \right], \quad (27)
\end{aligned}$$

which concludes the proof from the assumptions on $\{\pi_n, n \geq 1\}$ in ii) and iii).

For the other implication (i.e., the necessary condition), let us assume that there is a D-semifaithful coding scheme $\{\xi_n^* = (\phi_n^*, \mathcal{C}_n^*, \mathcal{D}_n^*), n \geq 1\}$ operating at distortion $d > 0$ such that

$$\lim_{n \to \infty} \sup_{\mu^n \in \Lambda^n} [R(\xi_n^*, \mu^n) - \mathcal{R}_n(d, \mu^n)] = 0. \quad (28)$$

From (2), we have that $\pi_{\phi_n^*} \in \mathcal{Q}_n(d)$ for all $n \geq 1$ (condition i). Concerning the information radius, using (21) it follows that

$$\begin{aligned}
&\sup_{\mu^n \in \Lambda^n} \left[R((\phi_n^*, \mathcal{C}_n^*, \mathcal{D}_n^*), \mu^n) - \frac{H_{\sigma(\pi_{\phi_n^*})}(\mu^n)}{n} \right] \\
&\geq \min_{\mathcal{C}_n} \sup_{\mu^n \in \Lambda^n} \left[R((\phi_n^*, \mathcal{C}_n, \mathcal{D}_n), \mu^n) - \frac{H_{\sigma(\pi_{\phi_n^*})}(\mu^n)}{n} \right] \\
&\geq \frac{1}{n} R^+(\Lambda^n, \sigma(\pi_{\phi_n^*})). \quad (29)
\end{aligned}$$

Then using the decomposition of the average redundancy in (18), it follows that

$$\begin{aligned}
&\sup_{\mu^n \in \Lambda^n} (R(\xi_n^*, \mu^n) - \mathcal{R}_n(d, \mu^n)) \\
&\geq \sup_{\mu^n \in \Lambda^n} \left[R(\xi_n^*, \mu^n) - \frac{H_{\sigma(\pi_{\phi_n^*})}(\mu^n)}{n} \right] \\
&\geq \frac{1}{n} R^+(\Lambda_f^n, \sigma(\pi_{\phi_n^*})), \quad (30)
\end{aligned}$$

which proves that condition ii) is satisfied from (28). Using again (18), it follows that $\forall \mu^n \in \Lambda^n$

$$R(\xi_n^*, \mu^n) - \mathcal{R}_n(d, \mu^n) \geq \frac{H_{\sigma(\pi_{\phi_n^*})}(\mu^n)}{n} - \mathcal{R}_n(d, \mu^n). \quad (31)$$

Verifying condition iii) follows from (28) and the definition of $\mathcal{R}_n(d, \mu^n)$ in (12). □

IV. Results For Envelope Families

The results for envelope distributions on ∞-alphabets are presented in this section. Let us first introduce some definitions that will be needed for the statement of results. We begin introducing the family of models:

Definition 4: Let $f : \mathbb{X} \longrightarrow \mathbb{R}^+$ be a function. We define the envelope family induced by f as

$$\Lambda_f \equiv \{\mu \in \mathcal{P}(\mathbb{X}) : \mu(x) \leq f(x), \forall x \in \mathbb{X}\}, \quad (32)$$

where $(\mu(x))_{x\in\mathbb{X}}$ is a short-hand notation for the probability mass function (pmf) of μ.

Definition 5: Let $\mathcal{H}(\mathbb{X}) \subset \mathcal{P}(\mathbb{X})$ denote the set of all probabilities (source) with finite entropy in $\mathbb{X}$.

In addition, we need to introduce a notion of regularity for the distortion function. We consider the Euclidean norm between two points in $\mathbb{X}$ denoted by $|i-j|$ for any $i,j \in \mathbb{X}$. With this, the ball of radius ϵ and centered at i is denoted by $B_\epsilon(i) \equiv \{j \in \mathbb{X}, |i-j| < \epsilon\}$ for any $\epsilon > 0$ and $i \in \mathbb{X}$.

Definition 6: An unbounded distortion function $\rho : \mathbb{X} \times \mathbb{X} \longrightarrow \mathbb{R}^+$ is said to be consistent with respect to the Euclidean norm if for any $K > 0$, there exists $\epsilon > 0$ such that for any $i \in \mathbb{X}$ if $j \notin B_\epsilon(i)$ then $\rho(i,j) \geq K$.

The condition on ρ stated in Definition 6 implies that for any arbitrarily large $K > 0$, there is a sufficiently large $\epsilon > 0$ (only function of K) where the condition $\rho(i,j) < K$ implies that $|i-j| < \epsilon$. In other words, a fixed distortion condition $\rho(i,j) < d$ (for any $d > 0$) would be violated eventually by making i and j progressively far apart in terms of the Euclidean norm (restricted over $\mathbb{X}$). An implication of this condition is that any partition of $\mathbb{X}$ with cells that meet a fixed-distortion requirement (i.e., elements of $\mathcal{Q}_1(d)$ for any $d > 0$) requires an infinite number of cells[8]. In other words, a d-covering of $\mathbb{X}$ with cells that meet a fixed-distortion requirement retains the infinite cardinality of the lossless problem (the cardinality of $\mathbb{X}$). In summary, the condition in Definition 6 is relevant as we are interested in problems where D-semifaithful source coding do not reduce to an equivalent finite alphabet coding task.

A. Main Results

THEOREM 2: Let $\Lambda_f \subset \mathcal{P}(\mathbb{X})$ be induced by a non-negative function f and ρ be an unbounded distortion consistent with respect to the Euclidean norm (Def. 6). We have the following results:

i) If $f \notin \ell_1(\mathbb{X})$, then for any $d > 0$ and any D-semifaithful coding scheme $\{\xi_n, n \geq 1\}$ operating at distortion d:

$$\sup_{\mu\in\Lambda_f} [R(\xi_n, \mu^n) - \mathcal{R}_n(d, \mu^n)] = \infty, \ \forall n \geq 1.$$

ii) If $f \in \ell_1(\mathbb{X})$, then for any distortion $d > 0$, there exists a D-semifaithful coding scheme $\{\xi_n, n \geq 1\}$ operating at distortion d — with respect to $\{\rho_n, n \geq 1\}$ — that is weakly minimax universal, i.e.,

$$\lim_{n\to\infty} [R(\xi_n, \mu^n) - \mathcal{R}_n(d, \mu^n)] = 0,$$

for any $\mu \in \Lambda_f \cap \mathcal{H}(\mathbb{X})$.

iii) If $\Lambda_f \subset \mathcal{H}(\mathbb{X})$,[9] then the same construction presented in ii) is strongly minimax universal, i.e.,

$$\lim_{n\to\infty} \sup_{\mu\in\Lambda_f} [R(\xi_n, \mu^n) - \mathcal{R}_n(d, \mu^n)] = 0.$$

The proofs of these results are presented in Section VII.

Some remarks about Theorem 2:

1: The result in part i) implies that achieving strong-minimax universality is not feasible for the entire collection of stationary memoryless sources in ∞-alphabets. This is a direct implication of this result using $f(i) = 1$ for all $i \in \mathbb{X}$.

2: Interestingly, part i) matches the impossibility condition known for the lossless case in [15]. Therefore, in the context of infinite alphabet stationary and memoryless sources, a non-zero distortion does not help making feasible the task of universal source coding as we move from the lossless to the lossy (fixed-distortion) setting of the variable length coding problem.

3: The argument used for the impossibility part relies on the proof of Lemma 1 and in particular on bounding from below the worst-case redundancy by the I-radius of Λ_f projected over the cells induced by a D-semifaithful code (operating at distortion d). Then, the proof reduces to show that this redefined I-radius (see (22)) is unbounded for any partition of $\mathbb{X}$ that belongs to $\mathcal{Q}_n(d)$ and for any $d > 0$.

4: On the other hand, assuming that $f \in \ell_1(\mathbb{X})$, the result in part ii) shows that there is a D-semifaithful scheme that achieves weak minimax universality for any $d > 0$. This result is strengthened in part iii) showing that the same D-semifaithful construction is strongly minimax universal provided that $\Lambda_f \subset \mathcal{H}(\mathbb{X})$.

5: The constructive (achievability) argument used for the proof of Theorem 2 (part iii) is based on a two-stage (lossy-lossless) scheme (see Figure 1 in Section VII). The basic idea of this construction is to consider a specific two-stage lossy coding scheme. In the first-stage of this scheme, the problem is projected (loosely) to a finite alphabet task for which results for finite alphabet universal source coding are adopted (see Lemma 5 in Section VII). The second-stage, on the other hand, is addressed as a lossless source coding problem over a transformed infinite alphabet, where results from lossless universal source coding for envelope families are used (see Lemma 6 in Section VII).

6: An important result used in the proof of Theorem 2 (part iii) is that the so called *envelope distribution* $\tilde{\mu}_f$ derived from f by

$$\tilde{\mu}_f(x) \equiv \begin{cases} f(x) & \text{if } x \geq \tau_f \\ 1 - \sum_{x\geq\tau_f} f(x) & \text{if } x = \tau_f - 1 \\ 0 & \text{if } x < \tau_f - 1, \end{cases},$$

with $\tau_f \equiv \min\left\{k \geq 1, \sum_{x\geq k} f(x) \leq 1\right\}$, is the probability in Λ_f that achieves maximum entropy under some mild considerations. The statement of this result is presented in Lemma 7 (in Sect VII-B). Therefore, the condition $\Lambda_f \subset \mathcal{H}(\mathbb{X})$ reduces to the verification of $H(\tilde{\mu}_f) < \infty$ and, consequently, that the function $(f(x)\log 1/f(x))_{x\in\mathbb{X}}$ is summable[10]. Complementing this observation, we obtain the following implication:[11]

COROLLARY 1: $\Lambda_f \subset \mathcal{H}(\mathbb{X})$ is equivalent to $\sup_{\mu\in\Lambda_f} H(\mu) < \infty$.

7: Finally, Theorem 2 can be extended to the scenario of a bounded distortion if it is consistent with the Euclidean norm in the following sense:

[8] An extended version of this result is stated in Proposition 2. Its proof is presented in Appendix VI-C.

[9] This condition implies that $f \in \ell_1(\mathbb{X})$.

[10] $(f(x)\log 1/f(x))_{x\in\mathbb{X}}$ being summable implies that $f \in \ell_1(\mathbb{X})$.

[11] The proof is presented in Appendix VI-D.

Definition 7: A bounded distortion function $\rho : \mathbb{X} \times \mathbb{X} \longrightarrow [0, \rho_{max}]$, with $\rho_{max} > 0$, is said to be consistent with respect to the Euclidean norm if for any $K \in (0, \rho_{max})$, there is $\epsilon > 0$ such that for any $i \in \mathbb{X}$ if $j \notin B_\epsilon(i)$ then $\rho(i, j) \geq K$.

The statement of that result would be the same as the statement of Theorem 2 but restricting d to the range $(0, \rho_{max})$. The proof argument follows directly from the proof of Theorem 2, consequently, both the statement and the proof are omitted. Finally, it is worth noting that the *Hamming distance* satisfies Def. 7 as many other regular distortions, e.g., $\rho_M(i, j) \equiv K \min\{|i - j|, M\}$ for any $K \in \mathbb{R}^+ \setminus \{0\}$ and $M > 1$.

B. Rate of Convergence

The next result complements Theorem 2 by providing an upper bound on the rate of convergence for the worst-case overhead for the case of summable envelope families.

THEOREM 3: Under the setting of Theorem 2, if $\Lambda_f \subset \mathcal{H}(\mathbb{X})$, and we add the condition that

$$\limsup_{k\to\infty} \frac{\sum_{i\geq k} f(i) \log 1/f(i)}{\tilde{\mu}_f(\mathcal{T}_k) \log 1/\tilde{\mu}_f(\mathcal{T}_k)} < \infty$$

with $\mathcal{T}_k \equiv \{k, k+1, \ldots\} \subset \mathbb{X}$, then for any distortion $d > 0$, there is a D-semifaithful coding scheme $\{\xi_n^*, n \geq 1\}$ operating at distortion d — with respect to $\{\rho_n, n \geq 1\}$ — such that

$$\sup_{\mu\in\Lambda_f} [R(\xi_n^*, \mu^n) - \mathcal{R}_n(d, \mu^n)] \leq C_0 \frac{u_f(n) \log n}{n} + C_1 \frac{\log n}{n} + C_2 \frac{1}{n},$$

where C_0, C_1 and C_2 are constants and

$$u_f(n) \equiv \min\{k \geq 1 \text{ such that } \tilde{\mu}_f(\mathcal{T}_{k+1}) < 1/n\}. \quad (33)$$

The proof is presented in Section VII.

This last result adds a regularity assumption on the way the tail component of the entropy of $\tilde{\mu}_f$ tends to zero, which is sufficient to obtain a rate of convergence for the worst-case overhead that is $O(u_f(n) \log(n)/n)$. Importantly, it can be verified that polynomial envelope families (with $f_p(x) = K/x^p$ for some $p > 1$ and $K > 0$) and exponential envelope families (with $f_p(x) = Ke^{-\alpha x}$ with $K > 0$ and $\alpha > 0$) satisfy the tail conditions stated in this result, and, consequently, they are both strongly minimax universal. In fact, we have the following:

LEMMA 2: Let us consider a polynomial function given by $(f_p(i))_{i\geq 1} = (K/i^p)_{i\geq 1}$. For any $K > 0$ and $p > 1$ it follows that

$$\limsup_{k\to\infty} \frac{\sum_{i\geq k} f_p(i) \log 1/f_p(i)}{\tilde{\mu}_{f_p}(\mathcal{T}_k) \log 1/\tilde{\mu}_{f_p}(\mathcal{T}_k)} < \infty.$$

LEMMA 3: Let us consider an exponential function given by $(f_\alpha(i))_{i\geq 1} = (Ke^{-\alpha i})_{i\geq 1}$. For any $K > 0$ and $\alpha > 0$ it follows that

$$\limsup_{k\to\infty} \frac{\sum_{i\geq k} f_\alpha(i) \log(1/f_\alpha(i))}{\tilde{\mu}_{f_\alpha}(\mathcal{T}_k) \log(1/\tilde{\mu}_{f_\alpha}(\mathcal{T}_k))} < \infty.$$

The proofs of these Lemmas are presented in Appendices I and II, respectively.

Finally, the sequence $(u_f(n))_{n\geq 1}$ in (33) was introduced by Bontemps *et al.* in [16] for the lossless source coding problem, where the same rate $O(u_f(n) \log(n)/n)$ was obtained for the redundancy of the best (lossless) universal scheme with $f \in \ell_1(\mathbb{X})$, see [15, Th. 4] and [16, Th.2 and Prop. 5][12]. For the exponential envelope family of parameter α in Lemma 3, it was shown in [16, Prop. 6] and [22, Sect. VI.B.4)] that $\frac{1}{\alpha}\ln(Kn) + 1 < u_{f_\alpha}(n) \leq \frac{1}{\alpha}\ln(CKn)$ where $C = 1/(1 - e^{-\alpha})$. Consequently, from Theorem 3, there is $\{\xi_n^*, n \geq 1\}$ (operating at distortion d) such that $\sup_{\mu\in\Lambda_{f_\alpha}} [R(\xi_n^*, \mu^n) - \mathcal{R}_n(d, \mu^n)]$ is $O(\frac{(\log n)^2}{n})$. For the power law envelope family of parameter p in Lemma 2, the same method used to bound $u_{f_\alpha}(n)$ for the exponential family in [16, Prop.6] can be adopted to show that $(KS)^{1/p} \cdot n^{1/p} - 2 < u_{f_p}(n) \leq (KS)^{1/p} \cdot n^{1/p} + 1$, where $S = \sum_{k\geq 1} 1/k^p < \infty$ (as $(f_p(i))_{i\geq 1} \in \ell_1(\mathbb{X})$ for $p < 1$). Consequently, from Theorem 3, there is $\{\xi_n^*, n \geq 1\}$ (operating at distortion d) such that $\sup_{\mu\in\Lambda_{f_p}} [R(\xi_n^*, \mu^n) - \mathcal{R}_n(d, \mu^n)]$ is $O(\frac{(n^{1/p} \log n}{n})$.

V. Concluding Remarks and Future Work

On the analysis of universal D-semifaithful source coding on envelope families, Theorem 2 offers a necessary and sufficient condition to achieve minimax universality (in the sense introduced in Section III) for Λ_f in ∞-alphabets. Interestingly, the condition matches the summability condition over f known for the lossless (variable length) coding setting [15].

On future work, it remains an open problem to evaluate if the rate of convergence for the worst-case overhead obtained in Theorem 3 can be improved. It is intriguing that this result does not show a faster rate of convergence to zero with n (because of the non-zero distortion) with respect to its lossless counterpart that has the same rate [14]–[16], [20]. In fact, the result is insensitive to the value of d, which is something that requires a more careful analysis. In favor of the potential tightness of this part, we note that the non-zero distortion did not show an effect on the impossibility part (part i) of Theorem 2) with respect to its counterpart in the lossless problem [15]. On the other hand, it is clear that the distortion reduces the information radius of the projected family, in the sense that $R^+(\Lambda_f^n, \sigma(\pi_n)) \leq R^+(\Lambda_f^n)$ (see the definition in Eq.(22)). Then, the non-zero distortion does reduce this information radius complexity indicator. However, it is unclear that this gain in information radius translates into a gain in the overall minimax overhead expression in the lossy setting (with respect to its counterpart in the lossless setting) because the information radius captures only one of the two expressions of redundancy in (18). The other non-negative term is captured by the role of the universal quantization discrepancy mentioned in (35). To conclude this discussion, we realize (from the expression in (18) and the analysis in Section III-A.1) that a concrete way to prove that the result in Theorem 3 is optimal is to show that any sequence of partitions $\{\pi_n, n \geq 1\}$ such that $\pi_n \in \mathcal{Q}_n(d)$ satisfies that

$$\liminf_{n\to\infty} \frac{R^+(\Lambda_f^n, \sigma(\pi_n))}{R^+(\Lambda_f^n)} > 0. \quad (34)$$

[12] See also [22, Th.2] for a discussion of these results.

At a first glance, this result does not appear intuitive, but we could conjecture that it is true. Indeed, a related non-zero gain (information radius) result has been obtained by the authors of this work in [21], [22] but in a simpler context involving a tail-based scalar quantization and a distortion that is not fixed and tends to 0 with n. We believe that some of the tools used in this analysis can be adopted to derive (34), but the extension to analyze the object in (34) is not direct. This is a relevant direction for future work on universal source coding on infinite alphabets.

Finally, on the general analysis of universal D-semifaithful coding presented in Section III of this work, Lemma 1 tells us that meeting minimax universality for a given non-zero distortion $d > 0$ and a family of distributions Λ implies the existence of a universal sequence of D-semifaithful quantizers for Λ. Consequently, if the minimax redundancy criterion in (16) is met, for some $d > 0$, then there exists a sequence of partitions $\{\pi_n, n \geq 1\}$, such that $\pi_n \in \mathcal{Q}_n(d)$ (introduced in (12)), satisfying that

$$\lim_{n\to\infty} \frac{1}{n} \sup_{\mu^n \in \Lambda_f^n} \left[H_{\sigma(\pi_n)}(\mu^n) - \min_{\pi \in \mathcal{Q}_n(d)} H_{\sigma(\pi)}(\mu^n) \right] = 0, \tag{35}$$

where $H_{\sigma(\pi_n)}(\mu^n)$ is the entropy of μ^n restricted to the sub-sigma field induced by π_n (see Eq.(9)), and $\min_{\pi\in\mathcal{Q}_n(d)} H_{\sigma(\pi)}(\mu^n)$ is the quantizer in $\mathcal{Q}_n(d)$ that minimizes the entropy given the distribution μ^n and d. For obvious reasons, this representation dimension of the problem in (35) is not part of the lossless setting and requires a special treatment in this lossy case. In principle, it is not obvious that the criterion in (35) can be achieved for any family of stationary memoryless distributions in ∞-alphabets. Then, a direct implication of Theorem 2 for envelope families (the achievability part in iii)) is that there is a universal quantization scheme in the sense presented in (35) for Λ_f when $f \in \ell_1(\mathbb{X})$. The proof of Theorem 2 in Section VII-B offers a concrete construction for this universal quantization scheme based on the two-stage quantization approach illustrated in Figure 1.

VI. Entropy and Distribution Estimation for Infinite Alphabets: Discussion on Related Impossibility and Achievability Results

It is worth mentioning that a minimax universal (variable length) source code (Definition 3) can be used to determine (estimate) the rate-distortion function $\mathcal{R}^*(d, \mu^1)$ and the entropy of a distribution $\mu \in \mathcal{H}(\mathbb{X})$ as a special case ($d = 0$). Given this capability, it is interesting to mention some results for entropy estimation and distribution estimation for infinite alphabets. We focus on the lossless coding setting ($d = 0$) as this regime offers a connection with the problems of entropy estimation and distribution estimation in information divergence [4], [22], [36], [37].

In the lossless setting ($d = 0$), the existence of a weak minimax coding scheme $\left\{f_n : \mathcal{X}^n \to \{0,1\}^*\right\}$ (Def. 3) implies that $\sup_{\mu\in\Lambda} \lim_{n\to\infty}(r(f^n, \mu^n) - H(\mu)) = 0$ where $r(f_n, \mu^n) = \mathbb{E}(\mathcal{L}(f_n(X^n)))/n$ is the average rate in bits per sample of f_n. Consequently, under this (weak) minimax property, the average length of the code offers a consistent estimator of the entropy for all members of Λ. Unfortunately, for the family of finite entropy stationary and memoryless sources, there is no variable-length code that achieves this weak minimax criterion (point-wise convergence to the entropy over the members of $\mathcal{H}(\mathbb{X})$). This impossibility result was presented in [4, Theorem 3]. The authors of this result showed that for any block-length n and prefix-free code f_n, there exists a model $\mu \in \mathcal{H}(\mathbb{X})$ where $r(f_n, \mu^n) = \infty$.[13]. This negative result comes from the impossibility of estimating distributions in $\mathcal{H}(\mathbb{X})$ in the expected direct information divergence and the fact that the redundancy $r(f_n, \mu^n) - H(\mu)$ is lower bounded by the expected divergence between μ and an estimator of μ constructed from the code f_n [4, Theorem 2]. Consequently, the impossibility of distribution estimation for infinite alphabets implies the impossibility of estimating the entropy using the average length of a variable-length lossless (prefix-free) code. This impossibility result is consistent with Theorem 2 i), where based on the average length of a (D-semifaithful) code, we cannot estimate the rate-distortion function for the entire class of stationary and memoryless sources. Indeed, we show that for any $d > 0$, any $n \geq 1$, any (D-semifaithful) code ξ_n operating at distortion d and any $L > 0$, there is a distribution $\mu \in P(\mathbb{X})$ such that $R(\xi_n, \mu^n) - \mathcal{R}^*(d, \mu^1) > L$.

Breaking from the lossless coding structure, we presented in [22] an almost lossless (variable length) coding strategy for infinite alphabet memoryless sources with the capacity to estimate the entropy over the family $\mathcal{H}(\mathbb{X})$ by tolerating a non-zero but vanishing distortion. The proposed lossy scheme (with a vanishing distortion) does not induce an estimate of the distribution in expected information divergence. Consequently, the impossibility result in [4] does not contradict the fact that the average rate of this almost lossless source coding strategy provides a consistent estimation of the entropy distribution-free in $\mathcal{H}(\mathbb{X})$. More discussion on the use of this coding strategy for entropy estimation is presented in [22, Section IV.A].

Going beyond the use of source codes, there are other strategies and results for entropy estimation for infinite alphabets worth mentioning. These results, however, do not have an evident connection with the results presented in this paper and are relevant in their own merit. Regarding this, Antos and Kontoyiannis [36, Theorem 2 and Corollary 1] showed the remarkable result that the classical plug-in estimate of the entropy is strongly consistent[14] and consistent in the mean square error sense for any finite entropy distribution in $\mathcal{H}(\mathbb{X})$. On the analysis of the point-wise convergence of the estimation error, [36, Theorem 3] showed a finite length lower bound for this estimation error valid for any estimation scheme (impossibility result). This result implies (asymptotically) that no universal rate of convergence (to zero) can be achieved for the entropy estimation over the family of infinite alphabet memoryless sources. On the constructive side, constraining the problem to a family of distributions with specific power tail-

[13] Indeed, they show the stronger result that $\mathcal{L}(f_n(X^n)) = \infty$ almost surely, with $X^n \sim \mu^n$.

[14] Almost surely with respect to the empirical process distribution.

bounded conditions, Antos et al. [36, Theorem 7] presented a finite length expression for the rate of convergence of the estimation error of the classical plug-in estimator. Similar results – strong consistency distribution-free in $\mathcal{H}(\mathbb{X})$ and (almost sure) rate of convergence for the estimation error under some tail bounded conditions – have been obtained for a data-driven partition scheme in [37].[15].

VII. Proofs of the Main Results of Section IV

A. Theorem 2 — Part i): $f \notin \ell_1(\mathbb{X})$

Proof: Let us consider $d > 0$ and arbitrary D-semifaithful coding scheme $\{\rho_n = (\phi_n, \mathcal{C}_n, \mathcal{D}_n), n \geq 1\}$, such that

$$\rho_n(x^n, \phi_n(x^n)) \leq d, \tag{36}$$

for all $n \geq 1$ and $x^n \in \mathbb{X}^n$. We denote by $\mathcal{B}_n = \{\phi_n(x^n), x^n \in \mathbb{X}^n\}$ the range of ϕ_n and by π_{ϕ_n} the partition of $\mathbb{X}^n$ induced by ϕ_n (see Eq.(4)). From the decomposition in (18), for any $\mu^n \in \Lambda_f^n$

$$R(\xi_n, \mu^n) - \mathcal{R}_n(d, \mu^n) \geq \left[R(\xi_n, \mu^n) - \frac{H_{\sigma(\pi_{\phi_n})}(\mu^n)}{n}\right]. \tag{37}$$

From (37) and the analysis presented in Sec.III-A.1, the worst-case overhead over Λ_f is bounded by

$$\begin{aligned}
&\sup_{\mu^n \in \Lambda_f^n} R(\xi_n, \mu^n) - \mathcal{R}_n(d, \mu^n) \geq \frac{1}{n} \sup_{\mu \in \Lambda_f} D(v_{\mu^n} \| v_{\mathcal{C}_n}) \\
&\geq \frac{1}{n} \min_{v \in \mathcal{P}(\mathbb{X}^n)} \sup_{\mu^n \in \Lambda_f^n} D_{\sigma(\pi_\phi)}(\mu^n \| v) \\
&= \frac{1}{n} R^+(\Lambda_f^n, \sigma(\pi_{\phi_n})),
\end{aligned} \tag{38}$$

where $R^+(\Lambda_f^n, \sigma(\pi_{\phi_n}))$ is the information radius of the family Λ_f^n restricted to the sub-sigma field induced by π_{ϕ_n}.

The rest of the proof will show that $R^+(\Lambda_f^n, \sigma(\pi_{\phi_n})) = \infty$, for any $n \geq 1$. Using that $f \notin \ell_1(\mathbb{X})$, i.e., $\sum_{x \in \mathbb{X}} f(x) = \infty$, we can use the method presented in [15] to show that there is a countable collection of distributions $\tilde{\Lambda} = \{\tilde{\mu}_j, j \in \mathcal{J}\} \subset \Lambda_f$ with $|\mathcal{J}| = \infty$, where if we denote by

$$\mathcal{A}_j = \text{support}(\tilde{\mu}_j) \equiv \{x \in \mathbb{X}, \tilde{\mu}_j(x) > 0\},$$

then $|\mathcal{A}_j| < \infty$ for each $j \in \mathcal{J}$ and for any $i, j \in \mathcal{J}$ $i \neq j$ $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$. For completness, a construction of $\tilde{\Lambda}$ is presented in Appendix VI-B. Then, we can consider the n-fold family $\tilde{\Lambda}^n = \{\tilde{\mu}_j^n, j \in \mathcal{J}\}$ where $support(\tilde{\mu}_j^n) = \mathcal{A}_j^n = \mathcal{A}_j \times ... \times A_j \subset \mathbb{X}^n$. Using the consistency of ρ with respect to the Euclidean norm (in Def.6), Proposition 2 (in Appendix VI-C) shows that to achieve the fixed distortion criterion in (36), it is necessary that the range of ϕ_n has an infinite number of prototypes: i.e., we have that $|\mathcal{B}_n| = \infty$.

For any $j \in \mathcal{J}$, let us consider a covering of the support of $\tilde{\mu}_j^n$ by cells of π_{ϕ_n} by

$$\mathcal{C}(\mathcal{A}_j^n) \equiv \bigcup_{\mathcal{B} \in \pi_n(A_j^n)} \mathcal{B}, \tag{39}$$

where $\pi_n(\mathcal{A}_j^n) \equiv \{\mathcal{B} \in \pi_{\phi_n}, \mathcal{A}_j^n \cap \mathcal{B} \neq \emptyset\}$. At this point, we can show that $|\mathcal{C}(\mathcal{A}_j^n)| < \infty$, $\forall j \in \mathcal{J}$. This follows from the construction of $\{\mathcal{A}_j^n, j \in \mathcal{J}\}$ and the observation that any cell $\mathcal{B}$ in π_{ϕ_n} is a finite set from the hypothesis that $\pi_{\phi_n} \in \mathcal{Q}_n(d)$ and the consistency assumption on ρ_n (Def. 6). Therefore, we get that $\mathcal{C}(\mathcal{A}_j^n)$ in (39) is a finite set for any j.

Let us consider a countably infinite sub-collection of disjoint sets in $\{\mathcal{C}(\mathcal{A}_j^n), j \in \mathcal{J}\}$ by the following approach:

$$\begin{aligned}
&j_1 \equiv 1, \\
&j_2 \equiv \min\{j > j_1, \text{ such that } \mathcal{C}(\mathcal{A}_j^n) \cap \mathcal{C}(\mathcal{A}_{j_1}^n) = \emptyset\}, \\
&\dots \\
&j_k \equiv \min\left\{j > j_{k-1}, \text{ such that } \mathcal{C}(\mathcal{A}_j^n) \cap \bigcup_{l=1}^{k-1} \mathcal{C}(\mathcal{A}_{j_l}^n) = \emptyset\right\} \dots .
\end{aligned} \tag{40}$$

For any finite k, the solution in (40) is guaranteed to be achieved with a finite integer; then, we have an infinite new collection of probabilities $\hat{\Lambda}^n \equiv \{\tilde{\mu}_{j_k}^n, k \geq 1\} \subset \tilde{\Lambda}^n \subset \Lambda_f^n$. Based on the construction of $\tilde{\mu}_{j_k}^n$, the family $\hat{\Lambda}^n$ is composed of a collection of probabilities with disjoint support in $\mathbb{X}^n$. Then, we consider the following partition of $\mathbb{X}^n$

$$\eta_n \equiv \{\mathcal{C}(\mathcal{A}_{j_k}^n), k \geq 1\} \cup \left(\mathbb{X}^n \setminus \bigcup_{k=1}^{\infty} \mathcal{C}(\mathcal{A}_{j_k}^n)\right), \tag{41}$$

where it is clear that $\sigma(\eta_n) \subset \sigma(\pi_{\phi_n})$ and for any $\mu, v \in \mathcal{P}(\mathbb{X}^n)$, $D_{\sigma(\eta_n)}(\mu \| v) \leq D_{\sigma(\pi_{\phi_n})}(\mu \| v)$. The important point here is that $\hat{\Lambda}^n$ contains an infinite set of distributions with disjoint support when restricted to the cells of η_n in (41) and, thus, from the known connection between information radius and channel capacity [2], the following can be obtained:

LEMMA 4: $R^+(\hat{\Lambda}^n, \sigma(\eta_n)) = \infty$.

The proof of Lemma 4 is presented in Appendix V.

Therefore, we have that

$$R^+(\Lambda_f^n, \sigma(\pi_{\phi_n})) \geq R^+(\hat{\Lambda}^n, \sigma(\pi_{\phi_n})) \geq R^+(\hat{\Lambda}^n, \sigma(\eta_n)) = \infty, \tag{42}$$

from the fact that by construction $\hat{\Lambda}^n \subset \Lambda_f^n$ and $\sigma(\eta_n) \subset \sigma(\pi_{\phi_n})$. Finally (42) and (38) prove the impossibility part (Theorem 2 i)).[16] □

B. Theorem 2 — Part iii): $f \in \ell_1(\mathbb{X})$ and $\Lambda_f \subset \mathcal{H}(\mathbb{X})$

To organize the proof of this part, let us first introduce preliminary results and definitions that will be used in the main argument.

Definition 8: The distribution induced by the tail function f is given by

$$\tilde{\mu}_f(x) \equiv \begin{cases} f(x) & \text{if } x \geq \tau_f \\ 1 - \sum_{x \geq \tau_f} f(x) & \text{if } x = \tau_f - 1 \\ 0 & \text{if } x < \tau_f - 1, \end{cases} \tag{43}$$

where $\tau_f \equiv \min\left\{k \geq 1, \sum_{x \geq k} f(x) \leq 1\right\}$.

[15] As a side comment, the mentioned tail bounded conditions used in [36], [37] to obtain rate of convergence for entropy estimation are stronger than the condition used in our work to define the envelope families (Definition 4) — used in our achievability results in Theorems 2 and 3

[16] Alternatively, this result can be derived from (42) (Lemma 4) and Lemma 1.

Note that by construction, we have that $\tilde{\mu}_f(x) \in \Lambda_f$ and $\tau_f < \infty$ from the hypothesis that $f \in \ell_1(\mathbb{X})$.

Let us consider the finite set $\Gamma_k = \{1, .., k\}$ for any $k \geq 1$. Then we have the following result for finite alphabet sources:

LEMMA 5: For any $n \geq 1$, $k \geq 1$, distortion $d > 0$ and $\epsilon > 0$, there is a D-semifaithful code $\xi_n^{*k} = (\phi_n^{*k}, \mathcal{C}_n^{*k}, \mathcal{D}_n^{*k})$ on Γ_{k+1}, that operates at distortion $d > 0$ (w.r.t. $\tilde{\rho}_n$) and verifies that

$$\sup_{v \in \mathcal{P}(\Gamma_{k+1})} \left[\frac{1}{n} \mathbb{E}_{Y^n \sim v^n} \left\{ \mathcal{L}(\mathcal{C}_n^{*k}(\phi_n^{*k}(Y^n))) \right\} - \mathcal{R}_n(d, v^n) \right] \leq \frac{k \log(n+1)}{n} + \epsilon,$$

where $\mathcal{P}(\Gamma_{k+1})$ is the collection of probabilities on Γ_{k+1} (i.e., the simplex of dimension k).

The proof of Lemma 5 is presented in Appendix III.

For envelope families on infinite alphabets, we have the following remarkable result from Bontemps *et al.* [16]:

LEMMA 6: [16, Prop. 5] If $f \in \ell_1(\mathbb{X})$, then for any $n \geq 1$

$$(1 + o(1)) \frac{u_f(n) - 1}{4} \log n \leq R^+(\Lambda_f^n) \leq 2 + \log e + \frac{u_f(n) - 1}{2} \log n, \tag{44}$$

where

$$u_f(n) = \min \{k \geq 1 \text{ such that } \tilde{\mu}_f(\mathcal{T}_{k+1}) < 1/n\}. \tag{45}$$

Finally, let us consider a tail partition of $\mathbb{X}$ given by $\tilde{\pi}_k \equiv \{\Gamma_k, \{k+1\}, \{k+2\}, \ldots\}$ for any $k \geq 1$. The next result shows that the tail distribution $\tilde{\mu}_f$ (in Def. 43) achieves maximum entropy over the envelope family in the following sense:

LEMMA 7: If $H(\tilde{\mu}_f) < \infty$, it follows that eventually in k (i.e., for a sufficiently large k),

$$\sup_{\mu \in \Lambda_f} H_{\sigma(\tilde{\pi}_k)}(\mu) = H_{\sigma(\tilde{\pi}_k)}(\tilde{\mu}_f) < \infty.$$

Otherwise, if $H(\tilde{\mu}_f) = \infty$, then $\sup_{\mu \in \Lambda_f} H_{\sigma(\tilde{\pi}_k)}(\mu) = H_{\sigma(\tilde{\pi}_k)}(\tilde{\mu}_f) = \infty$ for any $k \geq 1$.

The proof of Lemma 7 is presented in Appendix IV. Consequently, we have that $\Lambda_f \subset \mathcal{H}(\mathbb{X})$ is equivalent to the condition $\sup_{\mu \in \Lambda_f} H(\mu) < \infty$.

Proof: The basic idea of the proof is to decompose the alphabet $\mathbb{X}$ into two segments and use a two-stage scheme. More precisely, let us consider the following mapping $S_k : \mathbb{X} \longrightarrow \Gamma_{k+1} = \{1, .., k+1\}$ where

$$S_k(x) \equiv \begin{cases} x & \text{if } x \in \Gamma_k = \{1, .., k\} \\ k+1 & \text{if } x > k \end{cases} \tag{46}$$

Applying this lossy mapping (letter by letter) to the source X^n, we create a truncated version of it:

$$Y_1^n(k) \equiv S_k(X^n) \equiv (S_k(X_1), \ldots, S_k(X_n)) \in \Gamma_{k+1}^n. \tag{47}$$

To retain the information lost from X^n in $Y_1^n(k)$, the following complementary mapping is used:

$$O_k(x) \equiv \begin{cases} 1 & \text{if } x \in \Gamma_k \\ x & \text{if } x > k \end{cases} \in \{1\} \cup \Gamma_k^c, \tag{48}$$

which induces

$$Z_1^n(k) \equiv O_k(X^n) \equiv (O_k(X_1), \ldots, O_k(X_n)) \in (\{1\} \cup \Gamma_k^c)^n. \tag{49}$$

It is clear that for any $k \geq 1$, $Y_1^n(k)$ and $Z_1^n(k)$ recover X^n with no loss. In this context, we propose a two-stage strategy where $Y_1^n(k)$ (a finite alphabet stationary memoryless source) is encoded with a D-semifaithful code (operating at distortion $d > 0$) and $Z_1^n(k)$ (an infinite alphabet stationary memoryless source) is encoded losslessly using a variable-length code. Let us consider a distortion $d > 0$ and a D-semifaithful triplet $\xi_n^k = (\phi_n^k, \mathcal{C}_n^k, \mathcal{D}_n^k)$ for the source $Y_1^n(k)$ on the alphabet Γ_{k+1}, operating at distortion $d > 0$ with respect to a distortion $\tilde{\rho}$ on $\Gamma_{k+1} \times \Gamma_{k+1}$, where we assume that $\tilde{\rho}$ coincides with ρ on $\Gamma_k \times \Gamma_k$ (the non-truncated symbols, see Eq. (46)). This means that for all $y^n \in \Gamma_{k+1}^n$

$$\tilde{\rho}_n(y^n, \phi_n^k(y^n)) \leq d. \tag{50}$$

On the other hand, we can consider a lossless variable-length encoder-decoder pair $(\tilde{\mathcal{C}}_n^k, \tilde{\mathcal{D}}_n^k)$ for the source $Z_1^n(k)$, where $\tilde{\mathcal{C}}_n^k : (\{1\} \cup \Gamma_k^c)^n \longrightarrow \{0,1\}^*$ and $\tilde{\mathcal{D}}_n^k : \{0,1\}^* \longrightarrow (\{1\} \cup \Gamma_k^c)^n$. Then, given an input $x^n \in \mathbb{X}^n$ the final output (after decoding) of this two-stage approach is

$$(\hat{y}^n, z^n) = (\phi_n^k(S_k(x^n)), O_k(x^n)) \in (\Gamma_{k+1})^n \times (\{1\} \cup \Gamma_k^c)^n. \tag{51}$$

Finally, we recover $\hat{x}^n$ from $(\hat{y}^n, z^n)$ by the following letter-by-letter mapping

$$\hat{x}^n = (\Psi_k(\hat{y}_1, z_1), \ldots, \Psi_k(\hat{y}_n, z_n)) \in \mathbb{X}^n,$$

where

$$\Psi_k(\hat{y}_i, z_i) \equiv \begin{cases} z_i & \text{if } z_i \in \Gamma_k^c \\ \hat{y}_i & \text{if } z_i = 1 \end{cases} \in \mathbb{X}. \tag{52}$$

Then, using the condition imposed on $\tilde{\rho}$, it follows that

$$\rho_n(x^n, \hat{x}^n) \leq \tilde{\rho}_n(y^n, \hat{y}^n) \leq d, \tag{53}$$

where $y^n = S_k(x^n)$ and $\hat{y}^n$ is defined in (51). The first inequality in (53) is verified in Appendix VI-A, and the second follows from the fact that ξ_n^k is a D-semifaithful code with respect to $\tilde{\rho}$. Therefore, this two-stage strategy produces a D-semifaithful code in $\mathbb{X}^n$ with respect to ρ. The encoding-decoding process is illustrated in Figure 1.

On the other hand, the length of this two-stage mapping (in bits per sample) that we denote by $\mathcal{T}_n^k$ is given by

$$\frac{1}{n} \mathcal{L}(\mathcal{T}_n^k(x^n)) = \frac{1}{n} \left[\mathcal{L}(\mathcal{C}_n^k(\phi_n^k(S_k(x^n)))) + \mathcal{L}(\tilde{\mathcal{C}}_n^k(O_k(x^n))) \right]. \tag{54}$$

Then if $X^n \sim \mu^n$, the average length is given by

$$\frac{1}{n} \mathbb{E}_{X^n} \left\{ \mathcal{L}(\mathcal{T}_n^k(X^n)) \right\} = \underbrace{\frac{1}{n} \mathbb{E}_{Y^n} \left\{ \mathcal{L}(\mathcal{C}_n^k(\phi_n^k(Y^n))) \right\}}_{\text{first-stage bit rate}} + \underbrace{\frac{1}{n} \mathbb{E}_{Z^n} \left\{ \mathcal{L}(\tilde{\mathcal{C}}_n^k(Z^n)) \right\}}_{\text{second-stage bit rate}}, \tag{55}$$

where $Y^n = S_k(X^n)$ and $Z^n = O_k(X^n)$.

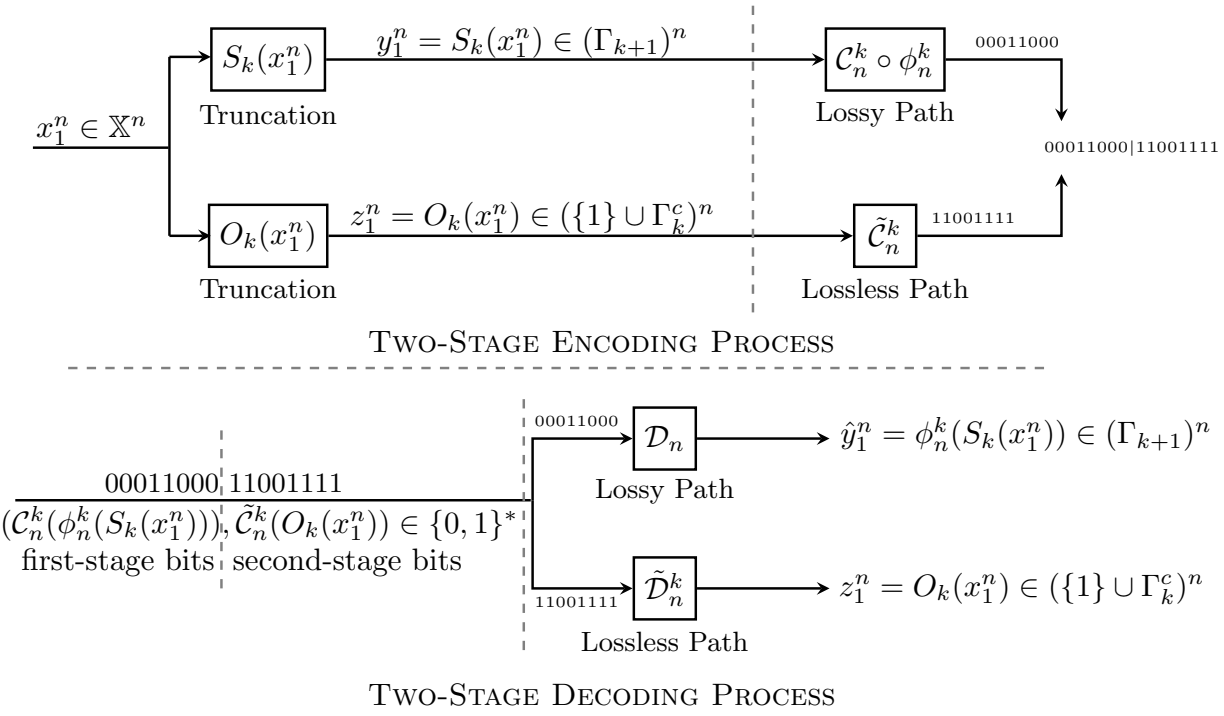

Fig. 1: Illustration of the two-stage scheme used in the achievability argument of Theorem 2 ($f \in \ell_1(\mathbb{X})$).

1) Analysis of the first-stage bits in (55): For the first term on the RHS of (55), it will be useful to consider the following truncated distortion ρ^k on $\mathbb{X} \times \mathbb{X}$, $(\forall x, \bar{x} \in \mathbb{X})$

$$\rho^k(x,\bar{x}) \equiv \begin{cases} \rho(x,\bar{x}) & \text{if } x,\bar{x} \in \Gamma_k \\ 0 & \text{if } x,\bar{x} \notin \Gamma_k \\ \min_{\tilde{x}>k} \rho(x,\tilde{x}) & \text{if } x \in \Gamma_k \text{ and } \bar{x} \notin \Gamma_k \\ \min_{\tilde{x}>k} \rho(\tilde{x},\bar{x}) & \text{if } x \notin \Gamma_k \text{ and } \bar{x} \in \Gamma_k \end{cases}, \tag{56}$$

to specify $\tilde{\rho}$ in $\Gamma_{k+1} \times \Gamma_{k+1}$, used in the first-stage of the construction. It follows that $\rho^k(x,\bar{x}) \leq \rho(x,\bar{x})$ and $\rho^k(x,\bar{x}) = \rho^k(S_k(x), S_k(\bar{x}))$. Consequently, we have that $\rho_n^k(x^n,\bar{x}^n) = \rho_n^k(S_k(x^n), S_k(\bar{x}^n))$ for any x^n and $\bar{x}^n$ in $\mathbb{X}^n$. For the rest of the argument, we fix $\tilde{\rho}_n(y^n,\bar{y}^n)$ to be $\rho_n^k(y^n,\bar{y}^n)$ for any $y^n, \bar{y}^n \in \Gamma_{k+1}^n \times \Gamma_{k+1}^n$. With this, let us introduce the counterpart of $\mathcal{R}_n(d,\mu_n)$ in (12) but using instead the induced distortion ρ_n^k, i.e.,

$$\mathcal{R}_n^k(d,\mu^n) \equiv \min_{\pi \in \mathcal{Q}_n^k(d)} \frac{H_{\sigma(\pi)}(\mu^n)}{n}, \tag{57}$$

where $\mathcal{Q}_n^k(d)$ is the collection of partitions of $\mathbb{X}^n$ such that any $\pi \in \mathcal{Q}_n^k(d)$ satisfies that $\forall A \in \pi$, $\exists y^n \in A$ such that $\sup_{x^n \in A} \rho_n^k(x^n,y^n) \leq d$. Then from the definition in (12), we have that

$$\mathcal{R}_n^k(d,\mu^n) \leq \mathcal{R}_n(d,\mu^n), \tag{58}$$

for any $d > 0$, any $n \geq 1$, any $k \geq 1$ and any $\mu \in \mathcal{P}(\mathbb{X})$.

On the other hand, if we consider the distribution of $Y^n = S_k(X^n) \in \Gamma_{k+1}$ (assuming that $X^n \sim \mu^n$ for some marginal $\mu \in \mathcal{P}(\mathbb{X})$) and in particular its marginal distribution v_μ in $\mathcal{P}(\Gamma_{k+1})$, we can use the operational finite-length rate-distortion function $\mathcal{R}_n(d,v_\mu^n)$ in (12) for v_μ. Using the fact that $\tilde{\rho}(S_k(x),S_k(x)) = \rho^k(x,\bar{x})$, it is simple to show that

$$\mathcal{R}_n(d,v_\mu^n) = \mathcal{R}_n^k(d,\mu^n), \tag{59}$$

for any $d > 0$, any $n \geq 1$, any $k \geq 1$ and any $\mu \in \mathcal{P}(\mathbb{X})$.

Finally, for any D-semifaithful code $\xi_n^k = (\phi_n^k, \mathcal{C}_n^k, \mathcal{D}_n^k)$ for Y^n operating at distortion $d > 0$ w.r.t. $\tilde{\rho}_n$, we have from (59), (10) and (12) that

$$\frac{1}{n}\mathbb{E}_{Y^n \sim v_\mu^n}\left\{\mathcal{L}(\mathcal{C}_n^k(\phi_n^k(Y^n)))\right\} \geq \mathcal{R}_n(d,v_\mu^n) = \mathcal{R}_n^k(d,\mu^n). \tag{60}$$

At this point, we can use the result in Lemma 5 for finite alphabet sources. In particular, from Lemma 5 (choosing $\epsilon = 1/n$) and the expressions in (60) and (59), we have that for any $n \geq 1$, $k \geq 1$ and distortion $d > 0$, there is a D-semifaithful code ξ_n^{*k} for the first-stage such that

$$\sup_{\mu \in \Lambda_f}\left[\frac{1}{n}\mathbb{E}_{X^n \sim \mu^n}\left\{\mathcal{L}(\mathcal{C}_n^{*k}(\phi_n^{*k}(S_k(X^n))))\right\} - \mathcal{R}_n^k(d,\mu^n)\right] \leq \frac{k\log(n+1)}{n} + \frac{1}{n}. \tag{61}$$

2) Analysis of the second-stage bits in (55): Considering the second term on the RHS of (55), let $m_\mu \in \mathcal{P}(\{1\} \cup \Gamma_k^c)$ be the distribution of $Z_i = O_k(X_i)$ induced by μ, then we have that

$$\frac{1}{n}\mathbb{E}_{Z^n \sim m_\mu^n}\left\{\mathcal{L}(\tilde{\mathcal{C}}_n^k(Z^n))\right\} \geq H(m_\mu), \tag{62}$$

because $\tilde{\mathcal{C}}_n^k$ is a variable length (prefix-free) lossless encoder of Z^n [3]. Furthermore, it is well-known that the redundancy of $\tilde{\mathcal{C}}_n^k$ is equal to (up to a discrepancy of $O(1/n)$) [3]

$$\frac{1}{n}\left[\mathbb{E}_{Z^n\sim m_\mu^n}\left\{\mathcal{L}(\tilde{\mathcal{C}}_n^k(Z^n))\right\} - H(m_\mu^n)\right] \approx \frac{1}{n} D(m_\mu^n \| m_{\tilde{\mathcal{C}}_n^k}), \tag{63}$$

where $m_{\tilde{\mathcal{C}}_n^k} \in \mathcal{P}((\{1\}\cup\Gamma_k^c)^n)$ is the distribution associated with the prefix-free code $\tilde{\mathcal{C}}_n^k$ [2], [3]. From this observation, the criterion for designing the second-stage in the context of universal source coding reduces to solving the following problem[17]:

$$R^+(\tilde{\Lambda}_f^n, k) \equiv \min_{m\in\mathcal{P}((\{1\}\cup\Gamma_k^c)^n)} \sup_{\mu\in\Lambda_f} D(m_\mu^n\|m), \tag{64}$$

which is the information radius of the projected family $\tilde{\Lambda}_f^n \equiv \left\{m_\mu^n, \mu\in\Lambda_f\right\}$. In particular, associated with the solution of (64) [2], there is a lossless code $\tilde{\mathcal{C}}_n^{*k}$ such that

$$\begin{aligned}\frac{R^+(\tilde{\Lambda}_f^n,k)}{n} &\le \sup_{\mu\in\Lambda_f}\left[\frac{1}{n}\mathbb{E}_{Z^n\sim m_\mu^n}\left\{\mathcal{L}(\tilde{\mathcal{C}}_n^{*k}(Z^n))\right\} - H(m_\mu)\right]\\ &\le \frac{R^+(\tilde{\Lambda}_f^n,k)+1}{n}.\end{aligned} \tag{65}$$

Importantly, using the information radius object introduced in (20), it is simple to check that

$$\begin{aligned}R^+(\tilde{\Lambda}_f^n,k) &= R^+(\Lambda_f^n, \sigma(\tilde{\pi}_k^{\times n}))\\ &= \min_{m\in\mathcal{P}(\mathbb{X}^n)}\sup_{\mu\in\Lambda_f} D_{\sigma(\tilde{\pi}_k^{\times n})}(\mu^n\|m),\end{aligned} \tag{66}$$

where

$$\tilde{\pi}_k^{\times n} \equiv \{\Gamma_k, \{k+1\}, \{k+2\}, \ldots\}^n$$

denotes the partition of $\mathbb{X}^n$ induced by the lossy mapping $(O_k(), O_k(), .., O_k()) : \mathbb{X}^n \longrightarrow (\{1\}\cup\Gamma_k^c)^n$. Then from (23) and (66)

$$R^+(\tilde{\Lambda}_f^n,k) \le R^+(\Lambda_f^n) \equiv \min_{m\in\mathcal{P}(\mathbb{X}^n)}\sup_{\mu\in\Lambda_f} D(\mu^n\|m). \tag{67}$$

This last expression is the information radius of the unconstrained family Λ_f^n [2]. The result by Bontemps et al. [16] (stated in Lemma 6) for summable envelope families comes in handy here. In fact, combining Lemma 6 with (65), for any $k \ge 1$ and $n \ge 1$, there exists a variable-length code $\tilde{\mathcal{C}}_n^{*k} : (\{1\}\cup\Gamma_k^c)^n \longrightarrow \{0,1\}^*$ satisfying that

$$\begin{aligned}&\sup_{\mu\in\Lambda_f}\left[\frac{1}{n}\mathbb{E}_{Z^n\sim m_\mu^n}\left\{\mathcal{L}(\tilde{\mathcal{C}}_n^{*k}(Z^n))\right\} - H(m_\mu)\right]\\ &\le \frac{u_f(n)-1}{2}\cdot\frac{\log n}{n} + \underbrace{\frac{2+\log e}{n}}_{O(1/n)}.\end{aligned} \tag{68}$$

It is important to note that the bound in the RHS of (68) is valid independent of (uniform over) k.

[17]Using the correspondence between prefix-free codes and perfect (dyadic) distributions.

3) Maximum Entropy analysis over the Envelope Family: For what follows, let us consider the assumption that[18]

$$\sup_{\mu\in\Lambda_f} H(\mu) < \infty. \tag{69}$$

Then from (68), we have that there is a coding scheme $\left\{\tilde{\mathcal{C}}_n^{*k}, n\ge 1\right\}$ satisfying that $\forall k \ge 1$:

$$\begin{aligned}&\sup_{\mu\in\Lambda_f}\frac{1}{n}\mathbb{E}_{X^n\sim\mu^n}\left\{\mathcal{L}(\tilde{\mathcal{C}}_n^{*k}(O_k(X^n)))\right\} \le\\ &\frac{u_f(n)-1}{2}\frac{\log n}{n} + O(1/n) + \sup_{\mu\in\Lambda_f} H(m_{\mu,k}),\\ &= \frac{u_f(n)-1}{2}\frac{\log n}{n} + O(1/n) + \sup_{\mu\in\Lambda_f} H_{\sigma(\tilde{\pi}_k)}(\mu),\ \forall n\ge 1,\end{aligned} \tag{70}$$

where in the first inequality $m_{\mu,k} \in \mathcal{P}(\{1\}\cup\Gamma_k^c)$ denotes the distribution of $Z = O_k(X)$ when $X \sim \mu \in \Lambda_f$, and in the second inequality, we use the tail partition $\tilde{\pi}_k = \{\Gamma_k, \{k+1\}, \{k+2\}, \ldots\}$. To continue with the argument, we use Lemma 7 that shows that $\tilde{\mu}_f$ in (43) achieves the maximum entropy of the problem stated in the right term of (70) (eventually in k). Then assuming (69), i.e., $H(\tilde{\mu}_f) < \infty$, and a sufficiently large k,

$$\begin{aligned}&\frac{1}{n}\sup_{\mu\in\Lambda_f}\mathbb{E}_{X^n\sim\mu^n}\left\{\mathcal{L}(\tilde{\mathcal{C}}_n^{*k}(O_k(X^n)))\right\}\\ &\le \frac{u_f(n)-1}{2}\frac{\log n}{n} + O(1/n)\\ &+ \tilde{\mu}_f(\Gamma_k)\log\frac{1}{\tilde{\mu}_f(\Gamma_k)} + \sum_{i\ge k+1}\tilde{\mu}_f(i)\log\frac{1}{\tilde{\mu}_f(i)},\ \forall n\ge 1.\end{aligned} \tag{71}$$

4) Concatenating the results in (55): From the expressions in (55), (61) and (71), we have that for any distortion $d > 0$ and threshold $k \ge 1$, there is a two-stage scheme $\left\{\mathcal{T}_n^{*k} = (\xi_n^{*k}, (\tilde{\mathcal{C}}_n^{*k}, \tilde{\mathcal{D}}_n^{*k})), n\ge 1\right\}$ where $\xi_n^{*k} = (\phi_n^{*k}, \mathcal{C}_n^{*k}, \mathcal{D}_n^{*k})$ is the D-semifaithful code of the first stage, operating at distortion d with respect to $\left\{\rho_n^k, n\ge 1\right\}$, and $(\tilde{\mathcal{C}}_n^{*k}, \tilde{\mathcal{D}}_n^{*k})$ is the variable-length lossless encoder-decoder pair of the second stage, such that for any $n \ge 1$ and a sufficiently large k:

$$\begin{aligned}&\sup_{\mu\in\Lambda_f}\left[\frac{1}{n}\mathbb{E}_{X^n\sim\mu^n}\left\{\mathcal{L}(\mathcal{T}_n^{*k}(X^n))\right\} - \mathcal{R}_n(d,\mu^n)\right]\\ &\le \frac{k\log(n+1)}{n} + \frac{u_f(n)-1}{2}\frac{\log n}{n} + O(1/n)\\ &+ \tilde{\mu}_f(\Gamma_k)\log\frac{1}{\tilde{\mu}_f(\Gamma_k)} + \sum_{i\ge k+1}\tilde{\mu}_f(i)\log\frac{1}{\tilde{\mu}_f(i)},\end{aligned} \tag{72}$$

assuming that $H(\tilde{\mu}_f) < \infty$. Finally it is clear in the above construction that we can take (k_n) function of n to achieve minimax universality using the fact that $(u_f(n)\cdot\log n/n)$ tends to zero with n [15], [16]. In fact, if (k_n) tends to ∞ with n

[18]This condition is equivalent to the condition $\Lambda_f \subset \mathcal{H}(\mathbb{X})$ used in statement of Theorem 2 — part iii). See Corollary 1.

and $\lim_{n\longrightarrow\infty} k_n \log(n)/n = 0$, from (72) this is sufficient to have that

$$\lim_{n\longrightarrow\infty} \sup_{\mu\in\Lambda_f} \left[\frac{1}{n}\mathbb{E}_{X^n\sim\mu^n}\left\{\mathcal{L}(\mathcal{T}_n^{*k_n}(X^n))\right\} - \mathcal{R}_n(d,\mu^n)\right] = 0. \tag{73}$$

Consequently, we achieve strong minimax universality with the construction $\left\{\mathcal{T}_n^{*k_n}, n\geq 1\right\}$ in the sense stated in (16). This concludes the proof of Part iii). □

C. Theorem 2 — Part ii): $f \in \ell_1(\mathbb{X})$ and $H(\tilde{\mu}_f) = \infty$

Proof: If we relax the finite entropy condition on the envelope distribution, i.e., we have that $H(\tilde{\mu}_f) = \infty$ from Corollary 1, the same arguments and in particular the two-stage construction presented in Section VII-B can be used to show that for any $\mu \in \Lambda_f$, such that $H(\mu) < \infty$, it follows that[19]

$$\underbrace{\left[\frac{1}{n}\mathbb{E}_{X^n\sim\mu^n}\left\{\mathcal{L}(\mathcal{T}_n^{*k_n}(X^n))\right\} - \mathcal{R}_n(d,\mu^n)\right]}_{\text{point-wise analysis}} \leq$$
$$\frac{k_n\log(n+1)}{n} + \frac{u_f(n)-1}{2}\frac{\log n}{n} + O(1/n)$$
$$+ \mu(\Gamma_k)\log\frac{1}{\mu(\Gamma_k)} + \sum_{i\geq k+1}\mu(i)\log\frac{1}{\mu(i)}. \tag{74}$$

Then under the conditions that (k_n) tends to infinity with n and (k_n) is $o(\log(n)/n)$, for any $\mu \in \Lambda_f \cap H(\mathbb{X})$ it follows that

$$\lim_{n\longrightarrow\infty}\left[\frac{1}{n}\mathbb{E}_{X^n\sim\mu^n}\left\{\mathcal{L}(\mathcal{T}_n^{*k_n}(X^n))\right\} - \mathcal{R}_n(d,\mu^n)\right] = 0, \tag{75}$$

which concludes the proof of Part ii). □

D. Theorem 3

Proof: Let us consider the assumption that

$$\limsup_{k\to\infty}\frac{\sum_{i\geq k}\tilde{\mu}_f(i)\log(1/\tilde{\mu}_f(i))}{\tilde{\mu}_f(\mathcal{T}_k)\log 1/\tilde{\mu}_f(\mathcal{T}_k)} < \infty. \tag{76}$$

Notice that the expression in the numerator of (76) is well defined as $H(\tilde{\mu}_f) < \infty$ from the hypothesis that $\Lambda_f \subset \mathcal{H}(\mathbb{X})$. Hence, the result in (72) for the worst-case overhead can be adopted. Using a sequence $(k_n)_n$ such that $k_n \to \infty$ then the term $H_{\sigma(\tilde{\pi}_{k_n})}(\tilde{\mu}_f)$ in (72) can be expressed (in the limit) by

$$\limsup_{n\to\infty} H_{\sigma(\tilde{\pi}_{k_n})}(\tilde{\mu}_f) =$$
$$\limsup_{n\to\infty}\tilde{\mu}_f(\mathcal{T}_{k_n})\log 1/\tilde{\mu}_f(\mathcal{T}_{k_n})\times$$
$$\left[1 + \frac{\sum_{i\geq k_n}\tilde{\mu}_f(i)\log(1/\tilde{\mu}_f(i))}{\tilde{\mu}_f(\mathcal{T}_{k_n})\log 1/\tilde{\mu}_f(\mathcal{T}_{k_n})}\right], \tag{77}$$

[19]For the sake of space, the steps to derive (74) are not presented as it follows directly from the argument presented in Section VII-B.

where from (76), there are two constants $K_0 > 0$ and $N > 0$, such that for any $n \geq N$:

$$H_{\sigma(\tilde{\pi}_{k_n})}(\tilde{\mu}_f) = \mu(\Gamma_{k_n})\log\frac{1}{\mu(\Gamma_{k_n})} + \sum_{i\geq k_n+1}\mu(i)\log\frac{1}{\mu(i)}$$
$$\leq \tilde{\mu}_f(\mathcal{T}_{k_n})\log 1/\tilde{\mu}_f(\mathcal{T}_{k_n})\cdot K_0. \tag{78}$$

In particular, choosing $(k_n^f)_n = (u_f(n))_n$ by the definition in (45) it follows that $\tilde{\mu}_f(\mathcal{T}_{k_n^f+1}) < 1/n$ and $\tilde{\mu}_f(\mathcal{T}_{k_n^f}) \geq 1/n$. Then for any $n \geq 1$:

$$\tilde{\mu}_f(\mathcal{T}_{k_n^f})\log 1/\tilde{\mu}_f(\mathcal{T}_{k_n^f}) \leq \frac{1}{n}\log n. \tag{79}$$

Therefore considering the two-stage scheme $\left\{\mathcal{T}_n^{*k_n^f}, n\geq 1\right\}$ driven by $(k_n^f)_{n\geq 1}$, from (72), (78) and (79), we have that eventually in n

$$\sup_{\mu\in\Lambda_f}\left[\frac{1}{n}\mathbb{E}_{X^n\sim\mu^n}\left\{\mathcal{L}(\mathcal{T}_n^{*k_n^f}(X^n))\right\} - \mathcal{R}_n(d,\mu^n)\right]$$
$$\leq \frac{u_f(n)\log(n+1)}{n} + \frac{u_f(n)-1}{2}\frac{\log n}{n}$$
$$+ O(1/n) + K_0\cdot\frac{\log n}{n}, \tag{80}$$

which concludes the proof. □

VIII. Acknowledgment

The work of J.F. Silva was supported by Fondecyt 1210315 CONICYT-Chile and the Advanced Center for Electrical and Electronic Engineering, Basal Project FB0008. This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 792464. We thank the Associate Editor and the two reviewers for the time devoted to reviewing our work. Their observations and comments were instrumental to improve this paper content. The authors thank Sebastian Espinosa for helping with the Figure. Finally, we thank Diane Greenstein for editing and proofreading all this material.

Appendix I
Proof of Lemma 2

Proof: First, it is simple to verify that if $p > 1$, then $(f_p(i)\log 1/f_p(i))_{i\geq 1} \in \ell_1(\mathbb{X})$, which implies that $\tilde{\mu}_{f_p} \in \mathcal{H}(\mathbb{X})$ (see Eq.(43)). Let us introduce the tail series:[20]

$$\mathcal{S}_p^k \equiv \sum_{i\geq k}\tilde{\mu}_{f_p}(i) = \sum_{i\geq k} f_p(i),$$

where the last equality is valid eventually (for k sufficiently large). Then it follows that

$$\mathcal{S}_p^k = k^{-p}\sum_{i\geq k}\frac{k^p}{i^p}$$
$$= k^{-p}\left(1 + \frac{1}{((k+1)/k)^p} + \ldots + \frac{1}{((k+K)/k)^p} + \ldots\right)$$
$$= k^{-p}\left(1 + \sum_{i\geq 1}\frac{1}{(1+i/k)^p}\right). \tag{81}$$

[20]For this analysis, we consider $K = 1$.

The term of the series in the brackets on the RHS of (81) is indexed by the fraction i/k, where k is fixed and i goes over the integers. Hence, this series decomposes in k-additive components as follows:

$$\underbrace{\left(1+\sum_{i\geq 1}\frac{1}{(i+1)^p}\right)}_{\text{term with 0 offset}}+\underbrace{\sum_{i\geq 1}\frac{1}{(i+1/k)^p}}_{\text{term with } 1/k \text{ offset}}+\ldots + \underbrace{\sum_{i\geq 1}\frac{1}{(i+(k-1)/k)^p}}_{\text{term with } (k-1)/k \text{ offset}}. \quad (82)$$

The 0-offset term in (82) equals $\sum_{i\geq 1}\frac{1}{i^p}=S_p^1$. The l/k-offset term is upper bounded by $\sum_{i\geq 1}\frac{1}{i^p}=S_p^1$ and lower bounded by $\sum_{i\geq 1}\frac{1}{(i+1)^p}=\sum_{i\geq 2}\frac{1}{i^p}=S_p^2$ for any $l\in\{1,..,k-1\}$. Therefore from (81) and (82), we have that

$$\frac{1}{k^{p-1}}S_p^1\geq S_p^k\geq\frac{1}{k^p}\left(S_p^1+(k-1)S_p^2\right)\geq\frac{1}{k^{p-1}}S_p^2, \quad (83)$$

which means that $S_p^k\sim\frac{1}{k^{p-1}}$. When $p>1$, this term tends to zero with k.

To continue with the proof, let us analyze the following information series:

$$I_p^k\equiv\sum_{i\geq k}\tilde{\mu}_{f_p}(i)\log(1/\tilde{\mu}_{f_p}(i))=\sum_{i\geq k}f_p(i)\log(1/f_p(i)),$$

where the last equality is valid eventually (for k sufficiently large). This last expression is equal to $p\sum_{i\geq k}\frac{1}{i^p}\log i$. Therefore, we can concentrate on the series:

$$\tilde{I}_p^k\equiv\sum_{i\geq k}\frac{1}{i^p}\log i=\frac{\log k}{k^p}\left[1+\sum_{i\geq 1}\frac{\log(k+i)/\log(k)}{((k+i)/k)^p}\right]=\frac{\log k}{k^p}\left[1+\sum_{i\geq 1}\frac{\log(k+i)/\log(k)}{(1+i/k)^p}\right]. \quad (84)$$

Similarly to (82), the series on the RHS of (84) can be decomposed in

$$\underbrace{\left[1+\sum_{i\geq 1}\frac{\log(k+ki)/\log(k)}{(1+i)^p}\right]}_{\text{0-term}}+\underbrace{\sum_{i\geq 1}\frac{\log(ik+1)/\log(k)}{(i+1/k)^p}}_{1/k\text{-offset term}}+\ldots+\underbrace{\sum_{i\geq 1}\frac{\log(ik+k-1)/\log(k)}{(i+(k-1)/k)^p}}_{(k-1)/k\text{-offset term}}. \quad (85)$$

For the 0-offset term, we have that

$$\left[1+\sum_{i\geq 1}\frac{\log(k+ki)/\log(k)}{(1+i)^p}\right]\leq 1+\sum_{i\geq 1}\left(\frac{1}{1+i}\right)^p+\frac{1}{\log k}\sum_{i\geq 1}\frac{\log(i+1)}{(i+1)^p}=S_p^1+\frac{1}{\log k}I_p^2, \quad (86)$$

while for the generic l/k-term in (85), we have that

$$\sum_{i\geq 1}\frac{\log(ik+l)/\log(k)}{(i+l/k)^p}\leq\sum_{i\geq 1}\frac{\log(ik+k)/\log(k)}{i^p}=\sum_{i\geq 1}\frac{1}{i^p}+\frac{1}{\log k}\underbrace{\sum_{i\geq 1}\frac{\log(i+1)}{i^p}}_{\bar{I}_p\equiv}=S_p^1+\frac{1}{\log(k)}\bar{I}_p. \quad (87)$$

Returning to (84), it follows from (85) and the posterior bounds that

$$I_p^k\leq\frac{p\log k}{k^{p-1}}\left[S_p^1+\frac{1}{\log k}\bar{I}_p\right]. \quad (88)$$

Then,

$$\limsup_{k\to\infty}\frac{I_p^k}{S_p^k\log(1/S_p^k)}\leq\limsup_{k\to\infty}\frac{\frac{p\log k}{k^{p-1}}S_p^1+\frac{p}{k^{p-1}}\bar{I}_p}{\frac{1}{k^{p-1}}S_p^2\log\frac{k^{p-1}}{S_p^1}} \quad (89)$$

$$=\frac{pS_p^1}{(p-1)S_p^2}<\infty, \quad (90)$$

which concludes the proof as $p>1$. □

Appendix II
Proof of Lemma 3

Proof: If we consider the information function $(i_\alpha(i))=(-f_\alpha(i)\log f_\alpha(i))$, it is clearly summable then $\tilde{\mu}_{f_\alpha}\in\mathcal{H}(\mathbb{X})$ (see Eq.(43)). Let us analyze the tail of $\tilde{\mu}_{f_\alpha}$, i.e., $S_\alpha^k\equiv\sum_{i\geq k}\tilde{\mu}_{f_\alpha}(i)$ for any $k\geq 1$. We have that $S_\alpha^k=e^{-\alpha k}\sum_{i\geq 1}Ke^{-\alpha i}=e^{-\alpha k}\cdot S_\alpha^1$. On the other hand, we need to analyze the tail fraction of the entropy of $\tilde{\mu}_{f_\alpha}$, i.e., $I_\alpha^k\equiv-\sum_{i\geq k}\tilde{\mu}_{f_\alpha}(i)\log\tilde{\mu}_{f_\alpha}(i)=-\sum_{i\geq k}f(i)\log f(i)$, the last equality holding eventually (for k sufficiently large). It is simple to show that

$$I_\alpha^k=\log(1/K)S_\alpha^k+K\alpha\log e\cdot\underbrace{\sum_{i\geq k}ie^{-\alpha i}}_{\bar{I}_\alpha^k\equiv} \quad (91)$$

where $\bar{I}_\alpha^k=ke^{-\alpha k}S_\alpha^0\,(1/K+1/k\cdot e^{-\alpha})$. Finally, we have from (91) that $I_\alpha^k=(\log(1/K)S_\alpha^1+S_\alpha^1)\cdot e^{-\alpha k}+(S_\alpha^0/K)\cdot ke^{-\alpha k}$. With this, it is simple to verify that

$$\limsup_{k\to\infty}\frac{I_\alpha^k}{S_\alpha^k\log(1/S_\alpha^k)}=\frac{S_\alpha^0}{KS_\alpha^1}\cdot\limsup_{k\to\infty}\frac{k}{k\alpha\log e+\log(1/S_\alpha^1)}=\frac{1}{Ke^{-\alpha}\alpha\log e}<\infty, \quad (92)$$

which proves the result. □

Appendix III
Proof of Lemma 5

Proof: Without loss of generality, let us consider the finite alphabet $\mathcal{A}=\{1,..,k\}$, a distortion $d>0$, and the

collection $\Lambda = \mathcal{P}(\mathcal{A})$. Using the non-asymptotic performance bound in (12), we are interested in the following object:

$$\min_{(\phi_n,\mathcal{C}_n,\mathcal{D}_n)} \sup_{\mu\in\Lambda} \left[\frac{1}{n}\mathbb{E}_{X^n\sim\mu^n}\left\{\mathcal{L}(\mathcal{C}_n(\phi_n(X^n)))\right\} - \mathcal{R}_n(d,\mu^n)\right], \tag{93}$$

where the minimum is carried over the collection of D-semifaithful codes on $\mathcal{A}$ operating at distortion d.

Let us fix an arbitrary $\epsilon > 0$. For any $x^n \in \mathcal{A}^n$, let p_{x^n} denote the type of x^n (the empirical distribution in $\mathcal{P}(\mathcal{A})$ induced by x^n), and $\tilde{P}_n \equiv \{p_{x^n}, x^n \in \mathcal{A}^n\}$ the collection of types obtained with sequences of length n. For any $p \in \tilde{P}_n$, the type class of p is given by $T_p \equiv \{x^n \in \mathcal{A}^n : p_{x^n} = p\}$, where it is clear that $\left\{T_p, p \in \tilde{P}_n\right\}$ offers a finite partition of $\mathcal{A}^n$. It is well known that $\left|\tilde{P}_n\right| \leq (n+1)^k$ [3]. For any member in the type class $p \in \tilde{P}_n$, let us choose a D-semifaithful code $\xi^{*k}_{n,p} = (\phi^{*k}_{n,p}, \mathcal{C}^{*k}_{n,p}, \mathcal{D}^{*k}_{n,p})$ indexed by p satisfying the condition:[21]

$$\frac{1}{n}\mathbb{E}_{Y^n\sim\bar{\mu}_p}\left\{\mathcal{L}(\mathcal{C}^{*k}_{n,p}(\phi^{*k}_{n,p}(Y^n)))\right\} \leq \mathcal{R}_n(d,\bar{\mu}_p) + \epsilon, \tag{94}$$

where $\bar{\mu}_p \in \mathcal{P}(\mathcal{A}^n)$ in (94) is a short-hand for the uniform distribution over $T_p \subset \mathcal{A}^n$.

With this, we consider a simple two-stage universal strategy, inspired by the two-stage scheme used in lossless universal source coding [2]. For encoding x^n there is fixed-rate function $f_n : \tilde{P}_n \to \{0,1\}^{k\lceil\log(n+1)\rceil}$ for indexing (encoding) the type of x^n, and conditioning on this information, the second-stage encodes x^n lossily with $\xi^{*k}_{n,p_{x^n}}$. Then the variable length representation of x^n operating at distortion d is given by $(f_n(p_{x^n}), \mathcal{C}^{*k}_{n,p_{x^n}}(\phi^{*k}_{n,p_{x^n}}(x^n))) \in \{0,1\}^*$. From this construction, it is simple to check that this scheme is a D–semifaithful code of $\mathcal{A}^n$ with respect to ρ_n.

Let us analyze its worst-case overhead in Λ. Let us consider $\mu \in \Lambda$, then if we denote by $\mathcal{T}^k_n = (f_n, (\xi^{*k}_{n,p}; p \in \tilde{P}_n))$ the two-stage scheme and (with small abuse of notation) we use $\mathcal{T}^k_n$ as a short-hand for the encoding mapping (from source symbols to binary sequences) then

$$\mathcal{L}(\mathcal{T}^k_n(x^n)) = \underbrace{k\log(n+1)}_{\text{first-stage bits}} + \underbrace{\mathcal{L}(\mathcal{C}^{*k}_{n,p_{x^n}}(\phi^{*k}_{n,p_{x^n}}(x^n)))}_{\text{second-stage bits}}, \tag{95}$$

$\forall x^n \in \mathcal{A}^n$. At this point, if we introduce $Y \equiv p_{X^n} \in \tilde{\mathcal{P}}_n$ (the type of $X^n \sim \mu^n$), it follows that for any $p \in \tilde{\mathcal{P}}_n$, $\mathbb{P}(Y = p) = \mu^n(T_p)$ and

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_{X^n\sim\mu^n}\left\{\mathcal{L}(\mathcal{T}^k_n(X^n))\right\} - \mathcal{R}_n(d,\mu^n)\\
&= \frac{1}{n}\mathbb{E}_{Y=p_{X^n}}\left\{\mathbb{E}_{X^n|Y}\left\{\mathcal{L}(\mathcal{T}^k_n(X^n))|Y\right\}\right\} - \mathcal{R}_n(d,\mu^n)\\
&= \frac{k\log(n+1)}{n} + \\
&\sum_{p\in\tilde{P}_n} \mu^n(T_p)\mathbb{E}_{X^n\sim\bar{\mu}_p}\left\{\mathcal{L}(\mathcal{C}^{*k}_{n,p}(\phi^{*k}_{n,p}(X^n)))\right\} - \mathcal{R}_n(d,\mu^n)
\end{aligned} \tag{96}$$

$$\begin{aligned}
&= \frac{k\log(n+1)}{n} + \\
&\sum_{p\in\tilde{P}_n} \mu^n(T_p)\left[\frac{1}{n}\mathbb{E}_{X^n\sim\bar{\mu}_p}\left\{\mathcal{L}(\mathcal{C}^{*k}_{n,p}(\phi^{*k}_{n,p}(X^n)))\right\} - \mathcal{R}_n(d,\bar{\mu}_p)\right]\\
&+ \underbrace{\sum_{p\in\tilde{P}_n} \mu^n(T_p)\mathcal{R}_n(d,\bar{\mu}_p) - \mathcal{R}_n(d,\mu^n)}_{\leq 0}
\end{aligned} \tag{97}$$

$$\leq \frac{k\log(n+1)}{n} + \epsilon. \tag{98}$$

The expression in (96) follows from (95) and the observation that conditioning to the event $Y = p$, for some $p \in \tilde{P}_n$, $X^n \sim \bar{\mu}_p$ independent of μ [3]. To obtain (97), we include the term $\sum_{p\in\tilde{P}_n}\mu^n(T_p)\mathcal{R}_n(d,\bar{\mu}_p)$ in (96) to then use the inequality in (94). Finally to obtain (98), we use the fact that $\mu^n(B) = \sum_{p\in\tilde{P}_n}\mu^n(T_p)\bar{\mu}_p(B)$ [3] and that $\mathcal{R}_n(d,\mu)$ is a concave function of the second argument from its construction in (12).[22] Finally, the inequality in (98) is valid distribution free (independent of μ), which concludes the proof. □

Appendix IV
Proof of Lemma 7

Proof: Let us assume that $H(\tilde{\mu}_f) < \infty$, where $\tilde{\mu}_f \in \Lambda_f$ is the tail distribution introduced in (43). Let us consider an arbitrary $\mu \in \Lambda_f$. Then we have that (assuming the regime

[21] This selection can be accomplished from (12).

[22] The concavity of $\mathcal{R}_n(d,\mu)$ is stated and proved in Proposition 3 at Appendix VI-E.

where $k > \tau_f$, see (43))

$$\begin{aligned}
&H_{\sigma(\tilde{\pi}_k)}(\tilde{\mu}_f) - H_{\sigma(\tilde{\pi}_k)}(\mu) = \\
&\tilde{\mu}_f(\Gamma_k)\log\frac{1}{\tilde{\mu}_f(\Gamma_k)} + \sum_{x\geq k+1}\mu(x)\log\frac{\mu(x)}{f(x)} \\
&+ \sum_{x\geq k+1}(f(x)-\mu(x))\log\frac{1}{f(x)} - \mu(\Gamma_k)\log\frac{1}{\mu(\Gamma_k)} \\
&\geq \tilde{\mu}_f(\Gamma_k)\log\frac{1}{\tilde{\mu}_f(\Gamma_k)} + \mu(\Gamma_k)\log\frac{\tilde{\mu}_f(\Gamma_k)}{\mu(\Gamma_k)} \\
&+ \mu(\Gamma_k)\log\mu(\Gamma_k) + \sum_{x\geq k+1}(f(x)-\mu(x))\log\frac{1}{f(x)} \qquad (99)\\
&= (\tilde{\mu}_f(\Gamma_k)-\mu(\Gamma_k))\cdot\log\frac{1}{\tilde{\mu}_f(\Gamma_k)} \\
&+ \sum_{x\geq k+1}(f(x)-\mu(x))\log\frac{1}{f(x)}, \\
&= \left(\sum_{x\geq k+1}\mu(x) - \sum_{x\geq k+1}\tilde{\mu}_f(x)\right)\cdot\log\frac{1}{\tilde{\mu}_f(\Gamma_k)} \\
&+ \sum_{x\geq k+1}(f(x)-\mu(x))\log\frac{1}{f(x)} \\
&= \sum_{x\geq k+1}(f(x)-\mu(x))\cdot\log\frac{1-\sum_{y\geq k+1}f(y)}{f(x)}. \qquad (100)
\end{aligned}$$

To obtain (99) we use that $\sum_{x\geq k+1}\mu(x)\log\frac{\mu(x)}{f(x)} \geq -\mu(\Gamma_k)\log\frac{\mu(\Gamma_k)}{\tilde{\mu}_f(\Gamma_k)}$ from the observation that $D_{\sigma(\tilde{\pi}_k)}(\mu\|\tilde{\mu}_f) \geq 0$. At this point, we use the fact that $f \in \ell_1(\mathbb{X})$, which means that $\lim_{k\to\infty}\sum_{x\geq k+1}f(x) = 0$. Therefore eventually (i.e., for a sufficiently large k) we have that $1-\sum_{x\geq k+1}f(x) > \sum_{x\geq k+1}f(x)$. Assuming this large k regime, it follows from (100) that

$$\begin{aligned}
&H_{\sigma(\tilde{\pi}_k)}(\tilde{\mu}_f) - H_{\sigma(\tilde{\pi}_k)}(\mu) \geq \\
&\sum_{x\geq k+1}(f(x)-\mu(x))\cdot\log\frac{\sum_{y\geq k+1}f(y)}{f(x)} \geq 0. \qquad (101)
\end{aligned}$$

The last inequality in (101) comes from the assumption that $\mu \in \Lambda_f$, which means that $\mu(x) \leq f(x)$ for all $x \in \mathbb{X}$.

On the second part of the result, we assume that $H(\tilde{\mu}_f) = \infty$. Here, it is clear that $H_{\sigma(\tilde{\pi}_k)}(\tilde{\mu}_f) = \infty$ for any $k \geq 1$, which is sufficient to obtain the unbounded result. □

Appendix V
Proof of Lemma 4

Proof: First, it is important to note that by the construction of $\hat{\Lambda}^n$ in (40) and the partition η_n in (41), $\hat{\Lambda}^n$ degenerates in the probability space $(\mathbb{X}^n, \sigma(\eta_n))$, in the sense that for any $k \geq 1$

$$H_{\sigma(\eta_n)}(\tilde{\mu}^n_{j_k}) = 0. \qquad (102)$$

Let us consider a distribution over the indices of the family $\hat{\Lambda}^n$ (i.e., over the integer set $\mathbb{N}$) $\rho \in \mathcal{P}(\mathbb{N})$, and with this we can construct a joint distribution $\rho \times \hat{\Lambda}^n$ in the product space $(\mathbb{N}, 2^{\mathbb{N}}) \times (\mathbb{X}^n, \sigma(\eta_n))$ in the standard way, i.e., $\rho \times \hat{\Lambda}^n(A \times B) = \sum_{a\in A}\rho(a)\cdot\tilde{\mu}^n_{j_a}(B)$ for any $A \subset \mathbb{N}$ and $B \in \sigma(\eta_n)$. Associated with this joint distribution, we can derive an expression for the mutual information of $\rho \times \hat{\Lambda}^n$ [2], [3]:

$$\begin{aligned}
\mathcal{I}(\rho;\hat{\Lambda}^n) &\equiv \sum_{a\in\mathbb{N}}\rho(a)\cdot D_{\sigma(\eta_n)}(\tilde{\mu}^n_{j_a}\|\bar{\mu}) \qquad (103)\\
&= H_{\sigma(\eta_n)}(\bar{\mu}) - \sum_{a\in\mathbb{N}}\rho(a)\cdot H_{\sigma(\eta_n)}(\tilde{\mu}^n_{j_a}), \qquad (104)
\end{aligned}$$

where $\bar{\mu}(B) \equiv \sum_{a\in\mathbb{N}}\rho(a)\tilde{\mu}^n_{j_a}(B)$ for any $B \in \sigma(\eta_n)$. Using (102), it is simple to show that $\mathcal{I}(\rho;\hat{\Lambda}^n) = H_{\sigma(\eta_n)}(\bar{\mu}) = H(\rho) = -\sum_{a\in\mathbb{N}}\rho(a)\log\rho(a)$. Finally it is well known, from the construction of the information radius of $\hat{\Lambda}^n$ [2], that $R^+(\hat{\Lambda}^n, \sigma(\eta_n)) \geq \mathcal{I}(\rho;\hat{\Lambda}^n) = H(\rho)$ for any $\rho \in \mathcal{P}(\mathbb{N})$. This last inequality proves the result as $\sup_{\rho\in\mathcal{P}(\mathbb{N})}H(\rho) = \infty$. □

Appendix VI
Auxiliary Results

A. Proposition 1

PROPOSITION 1: For all $x^n \in \mathbb{X}^n$, it follows that $\rho_n(x^n,\hat{x}^n) \leq \tilde{\rho}_n(y^n,\hat{y}^n)$.

Proof:

$$\begin{aligned}
\rho_n(x^n,\hat{x}^n) &= \frac{1}{n}\sum_{i=1}^n\rho(x_i,\hat{x}_i) \\
&= \frac{1}{n}\sum_{i=1}^n\left[\rho(x_i,\hat{x}_i)\mathbf{1}_{\Gamma_k}(x_i) + \rho(x_i,\hat{x}_i)\mathbf{1}_{\Gamma_k^c}(x_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^n\left[\rho(y_i,\hat{y}_i)\mathbf{1}_{\Gamma_k}(x_i) + \underbrace{\rho(x_i,z_i)}_{=0 \text{ as } z_i=x_i}\mathbf{1}_{\Gamma_k^c}(x_i)\right] \\
&\leq \frac{1}{n}\sum_{i=1}^n\tilde{\rho}(y_i,\hat{y}_i)\mathbf{1}_{\Gamma_k}(x_i) \qquad (105)\\
&\leq \frac{1}{n}\sum_{i=1}^n\tilde{\rho}(y_i,\hat{y}_i) \qquad (106)\\
&= \tilde{\rho}_n(y^n,\hat{y}^n). \qquad (107)
\end{aligned}$$

The first inequality in (105) follows from the construction of $\tilde{\rho}$ assuming that $\tilde{\rho}$ coincides with ρ in $\Gamma_k \times \Gamma_k$ and the mild assumption that $\tilde{\rho}(i,k+1) \leq \rho(i,k+1)$ for all $i \in \Gamma_k$. □

B. The construction of $\tilde{\Lambda} = \{\tilde{\mu}_j, j \in \mathcal{J}\}$

We know that $\sum_{j\geq 1}f(j) = \infty$. Let us introduce the bounded envelope function $\tilde{f} : \mathbb{X} \to [0,1]$ given by $\tilde{f}(i) \equiv \min\{1, f(i)\}$ for any $i \in \mathbb{X}$. It is simple to verify that $\sum_{j\geq 1}\tilde{f}(j) = \infty$ and that $\Lambda_{\tilde{f}} \subset \Lambda_f$. Using this bounded function, we introduce an iterative method to construct a countably infinite family $\tilde{\Lambda} = \{\tilde{\mu}_j, j \in \mathcal{J}\} \subset \Lambda_{\tilde{f}}$ as follows:

- **Iteration 1:** Select $\tilde{k}_1 = \min\left\{k \geq 1, \text{ st. } \sum_{l=1}^k\tilde{f}(l) \geq 1\right\}$[23]. Construct a

[23]This problem has a (finite) solution form the fact that $\sum_{j\geq 1}\tilde{f}(j) = \infty$.

probability $\tilde{\mu}_1 \in \mathcal{P}(\mathbb{X})$ by[24]

$$\begin{aligned}
&\tilde{\mu}_1(1) = \tilde{f}(1)\\
&....\\
&\tilde{\mu}_1(\tilde{k}_1 - 1) = \tilde{f}(\tilde{k}_1 - 1)\\
&\tilde{\mu}_1(\tilde{k}_1) = 1 - \tilde{\mu}_1(\left\{1, .., \tilde{k}_1 - 1\right\}) \leq \tilde{f}(\tilde{k}_1),
\end{aligned} \tag{108}$$

and $\tilde{\mu}_1(\left\{1, .., \tilde{k}_1\right\}^c) = 0$. By its construction in (108), we have that $\tilde{\mu}_1 \in \Lambda_{\tilde{f}}$ and $\mathcal{A}_1 = \text{support}(\tilde{\mu}_1) = \left\{1, .., \tilde{k}_1\right\}$.

- **Iteration 2:** Select $\tilde{k}_2 = \min\left\{k \geq 1, \text{ st. } \sum_{l=\tilde{k}_1+1}^{\tilde{k}_1+k} \tilde{f}(l) \geq 1\right\}$. Construct a probability $\tilde{\mu}_2 \in \mathcal{P}(\mathbb{X})$ by[25]

$$\begin{aligned}
&\tilde{\mu}_2(\tilde{k}_1 + 1) = \tilde{f}(\tilde{k}_1 + 1)\\
&....\\
&\tilde{\mu}_2(\tilde{k}_1 + \tilde{k}_2 - 1) = \tilde{f}(\tilde{k}_1 + \tilde{k}_2 - 1)\\
&\tilde{\mu}_2(\tilde{k}_1 + \tilde{k}_2) = 1 - \tilde{\mu}_2(\left\{\tilde{k}_1 + 1, .., \tilde{k}_1 + \tilde{k}_2 - 1\right\})\\
&\leq \tilde{f}(\tilde{k}_1 + \tilde{k}_2)
\end{aligned} \tag{109}$$

and $\tilde{\mu}_2(\left\{\tilde{k}_1 + 1, .., \tilde{k}_1 + \tilde{k}_2\right\}^c) = 0$. Then, we have that $\tilde{\mu}_2 \in \Lambda_{\tilde{f}}$ and $\mathcal{A}_2 = \text{support}(\tilde{\mu}_2) = \left\{\tilde{k}_1 + 1, .., \tilde{k}_1 + \tilde{k}_2\right\}$.

- **Iteration j: (for** $j > 2$**)** The construction follows from the same steps mentioned in iteration 2, where

$$\tilde{k}_j = \min\left\{k \geq 1, \text{ st. } \sum_{l=\tilde{k}_1+...+\tilde{k}_{j-1}+1}^{\tilde{k}_1+...+\tilde{k}_{j-1}+k} \tilde{f}(l) \geq 1\right\} \tag{110}$$

and $\tilde{\mu}_j(l) = \tilde{f}(l)$ for all $l \in \left\{\tilde{k}_1 + ... + \tilde{k}_{j-1} + 1, ..., \tilde{k}_1 + ... + \tilde{k}_j - 1\right\}$, $\tilde{\mu}_j(\tilde{k}_1 + ... + \tilde{k}_j) = 1 - \tilde{\mu}_j(\left\{\tilde{k}_1 + ... + \tilde{k}_{j-1} + 1, ..., \tilde{k}_1 + ... + \tilde{k}_j - 1\right\})$ and

$$\begin{aligned}
\mathcal{A}_j &= \text{support}(\tilde{\mu}_j)\\
&= \left\{\tilde{k}_1 + ... + \tilde{k}_{j-1} + 1, ..., \tilde{k}_1 + ... + \tilde{k}_j\right\}.
\end{aligned}$$

From the fact that $\sum_{i\geq 1} \tilde{f}(i) = \infty$, the key step in (110) is achieved in a finite value for any $j \geq 1$. In conclusion, this iterative (inductive) method shows the existence of an infinite collection of models $\{\tilde{\mu}_j, j \in \mathbb{N}\} \subset \Lambda_{\tilde{f}} \subset \Lambda_f$ with disjoint support.

C. Proposition 2

PROPOSITION 2: Let $\rho : \mathbb{X} \times \mathbb{X} \to \mathbb{R}^+$ be an unbounded distortion function consistent with respect to the Euclidean norm (Def. 6). For any $n \geq 1$ and $d > 0$, if $(\phi_n, \mathcal{C}_n, \mathcal{D}_n)$ is a D-semifaithful code of length n operating at distortion d (Def. 1), then $|B_n = \{\phi_n(x^n), x^n \in \mathbb{X}^n\}| = \infty$.

[24]Here, we assume that $\tilde{k}_1 > 1$, otherwise, we have the trivial case $\tilde{\mu}_1(1) = 1$.

[25]Here, we assume that $\tilde{k}_2 > 1$, otherwise, we have the trivial case $\tilde{\mu}_2(\tilde{k}_1 + 1) = 1$.

Proof: We prove this result by contradiction. Let us assume that $|B_n| < \infty$. We can consider the partition induced by $\phi_n(\cdot)$, i.e., $A^n_{y^n} \equiv \phi_n^{-1}(\{y^n\})$ for any $y^n \in B_n$, where $\pi_{\phi_n} = \left\{A^n_{y^n}, y^n \in B_n\right\}$ is a finite partition of $\mathbb{X}^n$. At least one of the cells of π_{ϕ_n} should have an infinite number of elements. Let us denote this infinite cell by $A^n_{\tilde{y}^n}$ with $\tilde{y}^n \in B_n$. If we index the elements of this cell, i.e., $A^n_{\tilde{y}^n} = \{x^n_\alpha, \alpha \in \mathcal{I}\}$ where $\mathcal{I}$ is a countably infinite set, we can create n (one dimensional) projections of $A^n_{\tilde{y}^n}$ by

$$\begin{aligned}
A^{n,1}_{\tilde{y}^n} &\equiv \{x^n_\alpha(1), \alpha \in \mathcal{I}\}\\
A^{n,2}_{\tilde{y}^n} &\equiv \{x^n_\alpha(2), \alpha \in \mathcal{I}\}\\
&...\\
A^{n,n}_{\tilde{y}^n} &\equiv \{x^n_\alpha(n), \alpha \in \mathcal{I}\} \subset \mathbb{X}
\end{aligned} \tag{111}$$

where $x^n_\alpha(i) \in \mathbb{X}$ denotes the i-component of the vector $x^n_\alpha \in A^n_{\tilde{y}^n}$. As we have that $\left|A^n_{\tilde{y}^n}\right| = \infty$, this implies that at least one of the $1D$ coordinate projections in (111) has infinite cardinality. Let us assume that $\left|A^{n,l^*}_{\tilde{y}^n}\right| = \infty$ for some $l^* \in \{1, .., n\}$ from this point on.

Let us consider $K > nd$ and its respective $\epsilon(K) > 0$ (from Def, 6) such that for any pair $i, j \in \mathbb{X}$, if $|i - j| \geq \epsilon(K)$ then $\rho(i, j) \geq K$ and $\rho(j, i) \geq K$. We can select $\alpha \in \mathcal{I}$ (i.e., $x^n_\alpha \in A^n_{\tilde{y}^n}$) such that $x^n_\alpha(l^*) \in A^{n,l^*}_{\tilde{y}^n} \cap B_{\epsilon(K)}(\tilde{y}^n(l^*))^c$ (as this last set is non-empty)[26]. By definition of $B_\epsilon(x)$, we have that $|x^n_\alpha(l^*) - \tilde{y}^n(l^*)| \geq \epsilon(K)$, which implies that $\rho(x^n_\alpha(l^*), \tilde{y}^n(l^*)) \geq K$. This implies that $\rho_n(x^n_\alpha, \tilde{y}^n) \geq \rho(x^n_\alpha(l^*), \tilde{y}^n(l^*))/n \geq K/n > d$ (see Eq.(1)). On the other hand, $x^n_\alpha \in A^n_{\tilde{y}^n}$ and, consequently, $\phi_n(x^n_\alpha) = \tilde{y}^n$, which implies that $\rho_n(x^n_\alpha, \phi_n(x^n_\alpha)) > d$. This last result contradicts the fact that $(\phi_n, \mathcal{C}_n, \mathcal{D}_n)$ is operating at distortion d. Therefore, our initial assumption that $|B_n| < \infty$ is incorrect, which concludes the proof. □

D. Proof of Corollary 1

Proof: It is sufficient to show that if $\Lambda_f \subset \mathcal{H}(\mathbb{X})$ then $\sup_{\mu\in\Lambda_f} H(\mu) < \infty$. We know that $\tilde{\mu}_f \in \Lambda_f$ (see Eq.(43)) and, consequently, $H(\tilde{\mu}_f) < \infty$. Let us consider $\mu \in \Lambda_f$ and an arbitrary large $K > 0$:

$$\begin{aligned}
H(\mu) &= \sum_{i=1}^{K} \mu(i) \log \frac{1}{\log \mu(i)} - \mu(\Gamma_K) \log \frac{1}{\mu(\Gamma_K)}\\
&\quad + H_{\sigma(\tilde{\pi}_K)}(\mu) \qquad (112)\\
&= \mu(\Gamma_K) \cdot \sum_{i=1}^{K} \frac{\mu(i)}{\mu(\Gamma_K)} \log \frac{\mu(\Gamma_K)}{\mu(i)} + H_{\sigma(\tilde{\pi}_K)}(\mu) \qquad (113)\\
&\leq \log K + H_{\sigma(\tilde{\pi}_K)}(\mu) \leq \log K + H_{\sigma(\tilde{\pi}_K)}(\tilde{\mu}_f)\\
&\leq \log K + H(\tilde{\mu}_f) < \infty. \qquad (114)
\end{aligned}$$

The first equality in (112) uses $\tilde{\pi}_k \equiv \{\Gamma_k, \{k+1\}, \{k+2\}, \ldots\}$ and the definition of $H_{\sigma(\tilde{\pi}_K)}(\mu)$ in (9). The inequalities in (114) follow from Lemma

[26]Considering that the ball $B_\epsilon(x) \subset \mathbb{X}$ (see Def.6) is a finite set (for any $x \in \mathbb{X}$ and $\epsilon > 0$), it follows that $\left|A^{n,l^*}_{\tilde{y}^n} \cap (B_\epsilon(\tilde{y}^n(l^*))^c\right| = \infty$ for any $\epsilon > 0$.

7 (using a sufficiently large K) and the fact that $H_{\sigma(\tilde{\pi}_K)}(\tilde{\mu}_f) \leq H(\tilde{\mu}_f) < \infty$. Finally the last expression in (114) is independent of μ (i.e., valid for any element of Λ_f), which concludes the proof. □

E. Proposition 3

PROPOSITION 3: $\mathcal{R}_n(d,\mu)$, defined in (12), is a concave function of its second argument in $\mathcal{P}(\mathbb{X}^n)$.

Proof: We assume that $d > 0$ and $n \geq 1$ are given. Let us consider a finite collection of models $\{\mu_i, i = 1,..,L\}$ in $\mathcal{P}(\mathbb{X}^n)$ and some weights $(w_i)_{i=1,..,L} \in [0,1]^L$ such that $\sum_{i=1}^L w_i = 1$. We need to verify that $\sum_{i=1}^L w_i \cdot \mathcal{R}_n(d,\mu_i) \leq \mathcal{R}_n(d,\mu)$ where $\mu = \sum_{i=1}^L w_i \cdot \mu_i \in \mathcal{P}(\mathbb{X}^n)$ denotes the convex combination of $\{\mu_i, i = 1,..,L\}$ using $(w_i)_{i=1,..,L}$. Let us consider an arbitrary partition $\pi \in \mathcal{Q}_n(d)$. Then we have that

$$\frac{H_{\sigma(\pi)}(\mu)}{n} \geq \sum_{i=1}^L w_i \cdot \frac{H_{\sigma(\pi)}(\mu_i)}{n} \tag{115}$$

$$\geq \sum_{i=1}^L w_i \cdot \mathcal{R}_n(d,\mu_i) \tag{116}$$

The first inequality in (115) is from the concavity of the entropy [3]. The second inequality comes by definition (see Eq.(12)) using the fact that $\pi \in \mathcal{Q}_n(d)$. Finally, the lower bound in (116) is valid for every $\pi \in \mathcal{Q}_n(d)$, which concludes the proof from the definition of $\mathcal{R}_n(d,\mu)$ in (12). □

F. Deriving the upper bound in (24)

We use the standard approach to construct a prefix-free (variable length) code from a probability in $\mathcal{P}(\mathcal{B}_n)$ [3]. For any model $v \in \mathcal{P}(\mathcal{B}_n)$, we can construct a code $\mathcal{C}_v$ such that for any $y^n \in \mathcal{B}_n$ it follows that $\mathcal{L}(\mathcal{C}_v(y^n)) = \lceil -\log(v(y^n)) \rceil$. This code satisfies the *Kraft-MacMillan inequality* in (3) considering that $-\log(v(y^n)) \leq \mathcal{L}(\mathcal{C}_v(y^n)) < -\log(v(y^n)) + 1$. Using these inequalities, we have that for any $v \in \mathcal{P}(\mathcal{B}_n)$ and $\mu \in \Lambda$ (considering that ϕ_n is given and fixed in this analysis)[27]

$$\left[R(\xi_n = (\phi_n, \mathcal{C}_v, \mathcal{D}_v), \mu^n) - \frac{H_{\sigma(\pi_{\phi_n})}(\mu^n)}{n} \right]$$
$$\leq \frac{1}{n}(D(v_{\mu^n}||v) + 1). \tag{117}$$

As the class of prefix-free mapping from $\mathcal{B}_n$ to $\{0,1\}^*$ contains the induced collection $\{\mathcal{C}_v, v \in \mathcal{P}(\mathcal{B}_n)\}$, from (117) it follows that

$$\min_{\mathcal{C}_n} \sup_{\mu\in\Lambda} \left[R((\phi_n, \mathcal{C}_n, \mathcal{D}_n), \mu^n) - \frac{H_{\sigma(\pi_{\phi_n})}(\mu^n)}{n} \right]$$
$$\leq \min_{v\in\mathcal{P}(\mathcal{B}_n)} \sup_{\mu\in\Lambda} \left[R((\phi_n, \mathcal{C}_v, \mathcal{D}_v), \mu^n) - \frac{H_{\sigma(\pi_{\phi_n})}(\mu^n)}{n} \right]$$
$$\leq \frac{1}{n} \cdot \min_{v\in\mathcal{P}(\mathcal{B}_n)} \sup_{\mu^n\in\Lambda^n} D(v_{\mu^n}\|v) + \frac{1}{n}. \tag{118}$$

Using the definition presented in Eq.(22), the last lower bound in (118) is $(R^+(\Lambda^n, \sigma(\pi_{\phi_n})) + 1)/n$.

[27] $\mathcal{D}_v$ denotes the decoder of $\mathcal{C}_v$.

REFERENCES

[1] J. F. Silva and P. Piantanida, "Universal d-semifaithful coding for countably infinite alphabets," in *ISIT*. IEEE International Symposium on Information Theory, 2019.

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

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

[4] 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.

[5] L. D. Davisson, "Universal noiseless coding," *IEEE Transactions on Information Theory*, vol. IT-19, no. 6, pp. 783–795, 1973.

[6] J. C. Kieffer, "Block coding for an ergodic source relative to a zero-one valued fidelity criterion," *IEEE Transactions on Information Theory*, vol. IT-24, no. 4, pp. 432–437, July 1978.

[7] B. M. Fitingof, "Optimal coding in the case of unknown and changing messsage statistics," *Prob. Inform. Transm.*, vol. 2, no. 2, pp. 3–11 (in Russian), 1–7 (English Transl.), 1966.

[8] T. J. Lynch, "Sequence time coding for data compression," *Proc. IEEE*, vol. 54, pp. 1490–1491, October 1966.

[9] Y. M. Shtarkov and V. F. Babkin, "Combinatorial methods of universal coding for discrete stationary sources," in *2nd Int. Information Theory Symp.*, 1971.

[10] V. F. Babkin, "Method of universal coding for an independent message source with non-exponential computational complexity," *Prob. Inform. Transm.*, vol. 7, no. 4, 1971.

[11] T. M. Cover, "Enumerative source coding," *IEEE Transactions on Information Theory*, vol. IT-19, pp. 73–76, January 1973.

[12] B. M. Fitingof, "The compression of discrete information," *Prob. Inform. Transm.*, vol. 3, no. 3, pp. 28–36 (in Russian), 22–29 (English Transl.), 1967.

[13] L. D. Davisson, "Comments on "sequence time coding for data compression"," *Proc. IEEE*, vol. 54, p. 2010, December 1966.

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

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] S. Boucheron, E. Gassiat, and M. Ohannessian, "About adaptive coding on countable alphabets: Max-stable envelope classes," *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 4948–4967, 2015.

[21] J. F. Silva and P. Piantanida, "The redundancy gains of almost lossless universal source coding over envelope families," in *IEEE International Symposium on Information Theory*, July 2017, pp. 1–5.

[22] ——, "Universal weak variable-length source coding on countably infinite alphabets," *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 649–668, January 2020.

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

[24] 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.

[25] D. S. Ornstein and P. C. Shields, "Universal almost sure data compression," *Annals of Probability*, vol. 18, no. 2, pp. 441–452, 1990.

[26] T. Berger, *Rate Distorsion Theory*, 1st ed., T. Kaiath, Ed. Prentice Hall, 1971.

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

[28] B. Yu and T. P. Speed, "A rate of convergence result for a universal d-semifaithful code," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 813–820, 1993.

[29] I. Kontoyiannis, "Pointwise redundancy in lossy data compression and universal lossy data compression," *IEEE Transactions on Information Theory*, vol. 46, no. 1, pp. 136–152, January 2000.
[30] N. Merhav, "A comment on "a rate of convergence result for a universal d-semifaithful code"," *IEEE Transactions on Information Theory*, vol. 41, no. 4, pp. 1200–1202, July 1995.
[31] Z. Zhang and E. Yang, "The redundancy of source coding with a fidelity criterion — part one: Known statistics," *IEEE Transactions on Information Theory*, vol. 43, no. 1, pp. 71–91, 1997.
[32] E. H. Yang and Z. Zhang, "The redundancy of source coding with a fidelity criterion — part ii: Coding at a fixed rate level with unkown statistics," *IEEE Transactions on Information Theory*, vol. 47, pp. 126–145, January 2001.
[33] D. Ishii and H. Yamamoto, "The redundancy of universal coding with a fidelity criterion," *IEICE Fundamentals*, vol. E80-A, no. 11, pp. 2225–2231, November 1997.
[34] I. Kontoyiannis and J. Zhang, "Arbitrary source models and bayesian codebooks in rate-distortion theory," *IEEE Transactions on Information Theory*, vol. 48, no. 8, pp. 2276–2290, August 2002.
[35] J. C. Kieffer, "Sample converses in source coding theory," *IEEE Transactions on Image Processing*, vol. 37, no. 2, pp. 263–268, March 1991.
[36] A. Antos and I. Kontoyiannis, "Convergence properties of functionals estimates for discrete distributions," *Random Structures and Algorithms*, vol. 19, no. 3-4, pp. 163–193, October-December 2001.
[37] J. F. Silva, "Shannon entropy estimation in ∞-alphabets from convergence results: Studying plug-in estimators," *Entropy*, vol. 20, no. 397, pp. 1–28, 2018.

PLACE PHOTO HERE

Jorge F. Silva (Senior Member, IEEE) is Associate Professor at the Department of Electrical Engineering, University of Chile, Santiago, Chile. He received the Master of Science (2005) and Ph.D (2008) in Electrical Engineering from the University of Southern California (USC). He is IEEE member of the Signal Processing and Information Theory Societies. Jorge F. Silva was research assistant at the Signal Analysis and Interpretation Laboratory (SAIL) at USC (2003-2008) and was also research intern at the Speech Research Group, Microsoft Corporation, Redmond (Summer 2005).

He is recipient of the Outstanding Thesis Award 2009 for Theoretical Research of the Viterbi School of Engineering, the Viterbi Doctoral Fellowship 2007-2008 and Simon Ramo Scholarship 2007-2008 at USC. He is an IEEE Senior Member and he was Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING (period February 2006 - February 2008). His research interests include: universal source coding, information measure estimation, learning and coding, representation learning, machine learning, statistical learning theory, compressed sensing, sparse and compresible information sources, wavelet and multi-resolution analysis.

PLACE PHOTO HERE

Pablo Piantanida (Senior Member, IEEE) received the B.Sc. degree in electrical engineering and the M.Sc. degree from the University of Buenos Aires, Argentina, in 2003, and the Ph.D. degree from Universit Paris-Sud, Orsay, France, in 2007. He is currently Professor with the Laboratoire des Signaux et Systmes (L2S), CentraleSuplec together with CNRS and Universit Paris-Saclay. He is also an associate member of Comte Inria research team (Lix - Ecole Polytechnique) and was visiting professor from 2018 to 2019 to the Montreal Institute for Learning Algorithms (Mila), Quebec.

His research interests include learning theory, information theory, machine learning, security of learning systems, privacy and applications to computer vision, health, natural language processing, among others. He has served as the General Co-Chair for the 2019 IEEE International Symposium on Information Theory (ISIT) and as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY 2019 - 2021 and Editorial Board of Section "Information Theory, Probability and Statistics" for Entropy. He is member of the IEEE Information Theory Society Conference Committee.