entropy

MDPI

Article

Fixed-Rate Universal Lossy Source Coding and Model Identification: Connection with Zero-Rate Density Estimation and the Skeleton Estimator

Jorge F. Silva [1,*] and Milan S. Derpich [2]

[1] Information and Decision System Group, Department of Electrical Engineering, Universidad de Chile, Av. Tupper 2007, Santiago 7591538, Chile

[2] Department of Electronic Engineering, Universidad Tecnica Federico Santa Maria, Valparaiso 2390123, Chile; milan.derpich@usm.cl

* Correspondence: josilva@ing.uchile.cl; Tel.: +56-2-978-4090

Received: 26 June 2018; Accepted: 22 August 2018; Published: 25 August 2018

Abstract: This work demonstrates a formal connection between density estimation with a data-rate constraint and the joint objective of fixed-rate universal lossy source coding and model identification introduced by Raginsky in 2008 (IEEE TIT, 2008, 54, 3059–3077). Using an equivalent learning formulation, we derive a necessary and sufficient condition over the class of densities for the achievability of the joint objective. The learning framework used here is the skeleton estimator, a rate-constrained learning scheme that offers achievable results for the joint coding and modeling problem by optimally adapting its learning parameters to the specific conditions of the problem. The results obtained with the skeleton estimator significantly extend the context where universal lossy source coding and model identification can be achieved, allowing for applications that move from the known case of parametric collection of densities with some smoothness and learnability conditions to the rich family of non-parametric L_1-totally bounded densities. In addition, in the parametric case we are able to remove one of the assumptions that constrain the applicability of the original result obtaining similar performances in terms of the distortion redundancy and per-letter rate overhead.

Keywords: fixed-rate lossy source coding; joint coding and modeling; universal source coding; learning with rate constraints; the skeleton estimator; L_1-totally bounded classes

1. Introduction

Universal source coding (USC) has a long history in information theory and statistics [1–5]. Davisson's seminal work [4] formalized the variable-length lossless coding problem and introduced important information quantities for performance analysis [1,2]. In this lossless setting, it is well-understood that the Shannon entropy provides the minimum achievable rate (in bits per sample) [2] to code a stationary and memoryless source when the probability (model) of the source is available. When the probability of the source is not known but belongs to a family of distributions $\mathcal{F}$ (the so called universal source coding problem), the focus of the problem is to characterize the penalty (or redundancy in bits per sample) that an encoder and decoder pair will experience due to the lack of knowledge about the samples' probability [1]. In the lossless case, a seminal result states that the least worst-case redundancy over $\mathcal{F}$ (or the minimax solution of the USC problem for $\mathcal{F}$) is determined by the information radius of $\mathcal{F}$ [1].

Building on this connection between least worse-case redundancy and information radius of $\mathcal{F}$, there are numerous important results developed for lossless USC [1,6–9]. In particular, it is known that the information radius grows sub-linearly (with the block-length) for the family of finite alphabet stationary and memoryless sources [1], which implies the existence of a universal source code that achieves Shannon entropy as the block length goes to a large value for every distribution in $\mathcal{F}$.

However universality is not possible for the family of alphabet stationary and memoryless sources because the information radius of this family is unbounded [3,5,7]. More recent results on lossless USC over countable infinite alphabets have looked at restricting the analysis to specific collections of distributions (with some tail bounded conditions) to achieve minimax universality [7–9] and also looked at weak variations of the lossless source coding setting [10–12].

In the fixed-rate lossy source coding problem, assuming first that the probability μ of a memoryless source is known, the performance limit of the coding problem is given by the Shannon distortion-rate function $D_\mu(R)$ [2,13]. Consequently, the universal lossy source coding problem reduces to compare the distortion of a coding scheme (satisfying a fixed-rate constraint) with the Shannon distortion-rate function assuming that the designer only knows that $\mu \in \mathcal{F}$. The literature on this problem is rich [3,5,14–18] with a first result dating back to Ziv [17] who showed the existence of weakly minimax fixed-rate universal lossy source code for the class of stationary sources under certain assumptions about the source, the alphabet, and the distortion measure. More refined results were presented in [5,16] one of which established necessary and sufficient conditions to achieve weakly minimax universality for the class of stationary and ergodic sources. To provide a more specific analysis of universal lossy source coding, Linder et al. [14] presented a lossy USC scheme with a distortion redundancy that goes to zero as $O(\sqrt{\frac{\log\log n}{\log n}})$ for the case of independent and identically distributed (i.i.d.) bound sources. Later Linder et al. [15] improved previous results showing a fixed-rate lossy construction with a distortion redundancy that vanishes as $O(n^{-1}\log n)$ and $O(\sqrt{n^{-1}\log n})$ with n for finite alphabet i.i.d. sources and bounded infinite alphabet i.i.d. sources, respectively. Similar convergence results were obtained using a nearest-neighbor vector quantization approach in [19].

It is also understood that universal variable length lossless-source coding is connected with the problem of distribution estimation [3,6,20] as there is a one-to-one correspondence between prefix-free codes and finite-entropy discrete distributions in the finite and countable alphabet case [1,2,21]. Building on this one-to-one correspondence in the lossless case, Györfi et al. ([3], Theorem 1) showed that the redundancy (in bits per sample) of a given code upper bounds the expected divergence between the true distribution of the source μ and the estimated distribution derived from the code. Therefore, the existence of a universal (lossless) source code for $\mathcal{F}$ implies the existence of a universal (distribution-free in $\mathcal{F}$) estimator of the distribution in expected (direct) information divergence [22]. This means that achieving lossless USC not only provides a lossless representation of the data, but it offers a consistent (error-free) estimator of the distribution at the receiver.

The connection between coding and distribution estimation that is evident in the lossless case is not, however, present in the (fixed-rate) lossy source coding problem. As argued in [18], a fixed-rate lossy source code does not offer a direct map with a probability distribution (model) for the source. In light of this gap between lossy codes and distributions (models) and motivated by some problems in adaptive control, where it is relevant to both compress data in a lossy way and identify the distribution of the source at the receiver [18,23], Raginsky explored the joint objective of fixed-rate universal lossy source coding and model (i.e., distribution) identification in [18].

Inspired by Rissanen's achievability construction in [6,20], Raginsky [18] proposed a new setting for the problem of fixed-rate universal lossy compression of continuous memoryless sources based on the idea of a two-stage joint coding and model or distribution identification framework. In this context, he proposed a two-stage scheme to consider two objectives: fixed-rate universal lossy source coding and source distribution (model) identification. The first objective of the scheme is to transmit the data (optimally) in the classical distortion-rate sense [24], while the second objective is to learn and transmit a description (quantized version) of the source distribution (model) [25,26]. Taking ideas from statistical learning, Raginsky proposed [18] splitting the data into training and testing samples. The training data is used in the first-stage of the encoding process to construct a quantized estimation of the source distribution and encode it (the first stage bits). Then in a second stage of the encoding process, the first-stage bits are used to pick a matched (with the estimated distribution) fixed-rate lossy source code to encode the test data (the second stage bits). In this joint coding and modeling setting, the existence of a zero-rate consistent estimator of the density (in expected total variation) is sufficient to show the existence of a weakly minimax universal fixed-rate source coding scheme [18] (Theorem 3.2), achieving the Shannon distortion-rate function [2,24,27,28], for any given rate. This result is obtained for a wide class of single-letter bounded distortion functions and for a family of source densities $\mathcal{F} = \{\mu_\theta : \theta \in \Theta\}$ indexed over a bounded finite dimensional space $\Theta \subset \otimes_{i=1}^{k}[-L, L] \subset \mathbb{R}^k$ (i.e., a parametric collection) with some needed smoothness and learnability conditions [18] (Theorem 3.2).

It is important to highlight that the joint coding and modeling achievability results in [18] did not degrade the performance of the source coding objective. In fact by restricting the analysis to the source coding objective alone, the joint coding and modeling framework in [18] showed the same state-of-the-art performance results as conventional two-stage universal source coding schemes (or universal vector quantizers) [14,15,19] in terms of distortion redundancy and per-letter rate overhead ($\mathcal{O}(\sqrt{\log(n)/n})$ and $\mathcal{O}(\log(n)/n)$, respectively) as the block length n tends to a large number. Importantly, the first-stage bits of this joint coding and modeling scheme are used to achieve model identification at the receiver with arbitrary precision in total variation (with a rate of convergence of $\mathcal{O}(\sqrt{\log(n)/n})$ as n goes to infinity), with no extra cost in bits per-letter compared with conventional fixed-rate lossy source coding methods.

Contributions of This Work

This work formally studies the interplay between density estimation under a data-rate constraint and the joint fixed-rate universal lossy source coding and modeling problem with training data or memory introduced in [18]. The first main result (Theorem 1) establishes a connection between zero-rate density-estimation and a universal joint coding and modeling scheme that achieves optimal lossy source coding (in a distortion-rate sense) and lossless model identification. This result is obtained for the general family of bounded single-letter distortions [13]. Remarkably, this connection implies that the construction of a joint coding and modeling scheme reduces to the construction of a zero-rate density estimator. From this result, the second main result (Theorem 2) stipulates a necessary and sufficient condition for the existence of a weakly minimax universal joint coding and modeling scheme. For the achievability part of this result, we used the skeleton estimator as our learning framework [29]. Using this learning framework we extend the parametric context explored in [18] to the rich non-parametric scenario of L_1-totally bounded densities [30].

Furthermore, revisiting the parametric case studied in [18], by using the skeleton estimator we are able to remove some of the assumptions that limit the applicability of the original result. We show that the skeleton estimator matches the best performance reported in [18] in terms of the distortion redundancy and (per-letter) rate overhead, in particular obtaining rates of convergence to zero of $\mathcal{O}(\sqrt{\log(n)/n})$ and $\mathcal{O}(\log(n)/n)$, respectively, as the block-length tends to infinity. To obtain this, our result relaxes the finite Vapnik and Chervonenkis (VC) dimension assumption considered in [18]. On the other hand, when the finite VC dimension assumption is added in the analysis, the skeleton learning scheme offers a convergence rate of $\mathcal{O}(1/\sqrt{n})$ for the distortion redundancy as the sample-length goes to infinity. Finally, the skeleton framework is implementable in the parametric case as its minimum-distance decision is carried out on a finite number of candidates and the oracle ϵ-skeleton (or the ϵ-covering in total variation of $\mathcal{F}$) [30] (Chapter 7) can be replaced by a practical uniform covering of the compact index set $\Theta \subset \mathbb{R}^k$ (Theorem 4). Finally, it is worth noting that a preliminary version of this work (in the context of density estimation under a data-rate constraint) was presented in [31].

The rest of the paper is organized as follows: Section 2 introduces the setting of the joint coding and modeling with training data. Section 3 elaborates the connections with zero-rate density estimation. Section 4 presents the main joint coding and modeling result (Theorem 2) and introduces the skeleton estimator. Finally, Section 5 revisits a special case where the distributions are indexed by finite dimensional bounded space (the parametric context). A summary of the results is presented in Sections 6 and 7. Finally, the proofs are presented in Section 8.

2. Preliminaries

The fixed-rate coding and modeling problem introduced in [18] is presented in this section. This joint coding and modeling problem will be the main focus of this work. In addition, notations and definitions used in the rest of the paper will be presented.

2.1. Basic Definitions

Let $\mathbb{X} \in \mathcal{B}(\mathbb{R}^d)$ be a separable and complete subset of $\mathbb{R}^d$ where $\mathcal{B}(\mathbb{R}^d)$ is the Borel sigma field. Let $\mathcal{P}(\mathbb{X})$ be the collection of probability measures on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$, with $\mathcal{B}(\mathbb{X})$ denoting the Borel sigma field restricted to $\mathbb{X}$, and let $\mathcal{AC}(\mathbb{X}) \subset \mathcal{P}(\mathbb{X})$ denote the set of probability measures absolutely continuous with respect to the Lebesgue measure λ [32]. For any $\mu \in \mathcal{AC}(\mathbb{X})$, $g_\mu(x) = \frac{d\mu}{d\lambda}(x)$ denotes its probability density function. The total variational distance [30] of v and μ in $\mathcal{P}(\mathbb{X})$ is given by (to avoid any confusion, if S is a set then $|S|$ denotes its cardinality).

$$V(\mu, v) = \sup_{A \in \mathcal{B}(\mathbb{X})} |\mu(A) - v(A)|. \tag{1}$$

For μ and v belonging to $\mathcal{AC}(\mathbb{X})$, if we define the Scheffé set for the pair (μ, v) by

$$A_{\mu,v} \equiv \{x \in \mathbb{X} : g_\mu(x) > g_v(x)\} \in \mathcal{B}(\mathbb{X}), \tag{2}$$

then $V(\mu, v) = \mu(A_{\mu,v}) - v(A_{\mu,v})$ [30,33].

2.2. Fixed-Rate Universal Lossy Source Coding with Memory or Training Data

Let $\{X_n : n \geq 1\}$ be an i.i.d. stochastic process (or stationary and memoryless source), where X_i takes values in $\mathbb{X} \subset \mathbb{R}^d$ and has a distribution μ in $\mathcal{F} = \{\mu_\theta : \theta \in \Theta\} \subset \mathcal{AC}(\mathbb{X})$. Θ is in general an index set for $\mathcal{F}$. The problem of lossy source-coding of a finite block of the process $X^n = (X_1, ..., X_n)$ reduces to find a mapping (or code) $\mathcal{C}^n(\cdot)$ from $\mathbb{X}^n$ to S_n, where S_n is a finite set. Given a cardinality constraint on S_n, the design objetive is to make $\mathcal{C}^n(X^n)$ as close as possible to X^n (in average) using for that a distortion function. The standard coding problem assumes the knowledge of μ for finding the optimal code (for any finite block n) [1,2,13], as well as for characterizing the fundamental performance limits of this task as n goes to infinity [2,24,28,34–36].

A more realistic scenario is the universal source coding (USC) problem [2], where the source distribution $\mu \in \mathcal{F}$ is unknown and a coding scheme needs to be designed optimally for the family $\mathcal{F}$. Here we focus on a specific learning variation of this task introduced by Raginsky in [18], where in addition to the data that needs to be compressed and recovered (with respect to a fidelity criterion), we have a finite number of i.i.d samples following the same distribution μ and that can be used to estimate μ in the encoding process (more details of this approach in Section 2.3). This additional data can be interpreted as memory, training data, or side information about μ available at the encoder because it is data that is not required to be compressed and recovered. The existence of this memory departs from the standard zero-memory setting considered in universal source coding [1]. However, this information can be seen as a realistic assumption in the context of a sequential block by block coding of an infinite sequence, where the data is partitioned into blocks of the same finite length and compressed sequentially block by block. Then in a given stage of this sequential process, the data from previous blocks are available at the encoder (lossless) for the process compressing the current block [18].

More specifically following the fixed-rate block coding and modeling setting introduced by Raginsky in [18], we consider an n-block coding scheme with finite memory m, where there is a distinction between the data $Z^m = (Z_1, ..., Z_m)$ that is available (as side information) to estimate the source distribution (training data) and the data X^n that needs to be encoded and recovered (source or test data), under the important assumption that both data sets are i.i.d. samples of the same unknown probability $\mu \in \mathcal{F}$. A systematic exposition of this coding setting and its connection with the classical setting of zero-memory block coding is presented in [18] (Section II). Formally, let us define an (m, n)-block code by the pair

$$\mathcal{C}^{m,n} \equiv \left(f : \mathbb{X}^m \times \mathbb{X}^n \to S_n, \qquad \phi : S_n \to \hat{\mathbb{X}}^n \right). \tag{3}$$

Then given a set of training samples $z^m \in \mathbb{X}^m$ and a finite block of the source $x^n \in \mathbb{X}^n$, $\mathcal{C}^{m,n}$ is the composition of: a encoding function $f(z^m, x^n)$ that maps x^n to an element in a finite set S_n conditioned on the training data (or memory) represented by z^m, and a decoding function $\phi(\cdot)$ that maps a symbol $s \in S_n$ into the reproduction points $\Gamma_{\mathcal{C}^{m,n}} \equiv \{\phi(s) : s \in S_n\}$ that we called the codebook of $\mathcal{C}^{m,n}$. In this context, $\hat{\mathbb{X}}$ denotes the reproduction space. As a short-hand, we denote by $\hat{x}^n = \mathcal{C}^{m,n}(x^n) = \phi(f(z^m, x^n))$ the reconstruction of x^n obtained by $\mathcal{C}^{m,n}$ and its memory z^m (for simplicity, the dependency of $\hat{x}^n$ or $\mathcal{C}^{m,n}(x^n)$ on the memory z^m will be implicit in the rest of the exposition.). The rate of $\mathcal{C}^{m,n}$ in bits-per-letter is given by $R(\mathcal{C}^{m,n}) \equiv \frac{\log_2 |S_n|}{n}$. In general, it is not possible to recover x^n from $\hat{x}^n$ given the cardinality constraint on S_n, and thus a single-letter distortion measure $\rho : \mathbb{X} \times \hat{\mathbb{X}} \to \mathbb{R}^+$ is used to quantify the n-block discrepancy by [24]

$$\rho^n(x^n, \hat{x}^n) \equiv \sum_{i=1}^{n} \rho(x_i, \hat{x}_i). \tag{4}$$

Finally considering $X^n \sim \mu^n$ and $Z^m \sim \mu^m$, the average distortion per-letter of $\mathcal{C}^{m,n}$ given Z^m is

$$D_\mu(\mathcal{C}^{m,n} | Z^m) \equiv \frac{1}{n} \mathbb{E}_{X^n \sim \mu^n} \left(\rho^n(X^n, \hat{X}^n) \right), \tag{5}$$

which is a function of Z^m and hence the average distortion per-letter of $\mathcal{C}^{m,n}$ is

$$D_\mu(\mathcal{C}^{m,n}) \equiv \mathbb{E}_{Z^m \sim \mu^m} \left(D_\mu(\mathcal{C}^{m,n} | Z^m) \right). \tag{6}$$

In universal source coding the performance of a code $D_\mu(\mathcal{C}^{m,n})$ is evaluated over a collection of distributions $\mu \in \mathcal{F}$ and is compared (point-wise) with the best code that can be obtained assuming that μ is known. For this analysis, we need the following definitions:

Definition 1 ([18]). *For a finite block length n and distribution $\mu \in \mathcal{F}$, the n-order operational distortion-rate function of μ at rate R is*

$$D^n_\mu(R) \equiv \inf_{m \geq 0} \inf_{\substack{\mathcal{C}^{m,n} \\ with\ R(\mathcal{C}^{m,n}) \leq R}} D_\mu(\mathcal{C}^{m,n}). \tag{7}$$

In this context, the operational distortion-rate function (DRF) [2,28] is given by

$$D_\mu(R) \equiv \lim_{n \to \infty} D^n_\mu(R) = \inf_{n \geq 1} D^n_\mu(R). \tag{8}$$

The celebrated Shannon lossy source-coding theorem [27] provides a single letter theoretical characterization for $D_\mu(R)$ in (8) (also known as the Shannon DRF). A nice exposition of this celebrated result can be found in [2,24,28].

It is worth noting that the operational distortion-rate function in (7) is equivalent to the classical zero-memory n-order operational distortion-rate function given by $\inf_{\mathcal{C}^{0,n}} \left\{ D_\mu(\mathcal{C}^{0,n}) : \text{such that } R(\mathcal{C}^{0,n}) \leq R \right\}$ [18] (Lemma 2.1). Then, allowing a nonzero memory (side information at the encoder) does not help in the minimization of the distortion when μ is known.

For the rest of the exposition, we will concentrate on the simple case studied in [18] where $n = m$ (i.e., the block-length is equal to the memory of the code). To be precise about the meaning of universality in this context, we resort to some standard definitions:

Definition 2 ([16])**.** *A coding scheme* $\{\mathcal{C}^{n,n} : n \geq 1\}$ *is weakly minimax universal for the class* $\mathcal{F}$ *at rate* R*, if* $\forall \mu \in \mathcal{F}$

$$\lim_{n\to\infty} D_\mu(\mathcal{C}^{n,n}) = D_\mu(R) \tag{9}$$

and $\limsup_{n\to\infty} R(\mathcal{C}^{n,n}) = \limsup_{n\to\infty} \frac{\log_2 |S_n|}{n} \leq R$*. Alternatively, the scheme is said to be strongly minimax universal for the class* $\mathcal{F}$ *at rate* R *if*

$$\lim_{n\to\infty} \sup_{\mu\in\mathcal{F}} \left[D_\mu(\mathcal{C}^{n,n}) - D_\mu(R)\right] = 0 \tag{10}$$

and $\limsup_{n\to\infty} R(\mathcal{C}^{n,n}) \leq R$*.*

Decomposing the distortion redundancy in two terms,

$$D_\mu(\mathcal{C}^{n,n}) - D_\mu(R) = \left[D_\mu(\mathcal{C}^{n,n}) - D^n_\mu(R)\right] + \left[D^n_\mu(R) - D_\mu(R)\right], \tag{11}$$

the first term $\left[D_\mu(\mathcal{C}^{n,n}) - D^n_\mu(R)\right]$ is the n-order distortion redundancy, which is the discrepancy that can be attributed exclusively to the goodness of the coding scheme. The second term in (11), i.e., $\left[D^n_\mu(R) - D_\mu(R)\right]$, has to do with how fast $D^n_\mu(R)$ converges to the Shannon DRF as the block length tends to infinity (see further details in [14] (Section III) and references therein). From this observation, we introduce the following definition:

Definition 3. *A coding scheme* $\{\mathcal{C}^{n,n} : n \geq 1\}$ *is strongly finite-block universal for the class* $\mathcal{F}$ *at rate* R *if*

$$\lim_{n\to\infty} \sup_{\mu\in\mathcal{F}} \left[D_\mu(\mathcal{C}^{n,n}) - D^n_\mu(R)\right] = 0 \tag{12}$$

and $\limsup_{n\to\infty} R(\mathcal{C}^{n,n}) \leq R$*.*

Note that if $\{\mathcal{C}^{n,n} : n \geq 1\}$ is strongly minimax universal then it is strongly finite-block universal, but the converse result is not true in general. The missing condition to make these two criteria equivalent is the uniform convergence of $D^n_\mu(R)$ to $D_\mu(R)$ in the class $\mathcal{F}$. More discussion about this point in Section 6.

2.3. Raginsky's Two-Stage Joint Universal Coding and Modeling

Motivated by the work of Rissanen [6], Raginsky [18] proposed a two-stage block code with finite memory (training data), with the objective of doing both fixed-rate lossy source coding, and identification of the source distribution at the receiver. More precisely, given $Z^n \sim \mu^n_\theta$ and $X^n \sim \mu^n_\theta$ (the training and the source-data samples, respectively), an (n,n)-joint coding and modeling rule is given by

$$\begin{aligned}\mathcal{C}^{n,n} \equiv \Big(& f_n : \mathbb{X}^n \to \tilde{S}_n, \phi_n : \tilde{S}_n \to \Theta, \\ & \left\{ f_{n,\tilde{s}} : \mathbb{X}^n \to S_n, \phi_{n,\tilde{s}} : S_n \to \hat{\mathbb{X}}^n ; \tilde{s} \in \tilde{S}_n \right\}\Big),\end{aligned} \tag{13}$$

where S_n and $\tilde{S}_n$ are finite-set functions of n. $\mathcal{C}^{n,n}$ processes (Z^n, X^n) in two stages. In the first stage, the pair (f_n, ϕ_n) in (13) uses Z^n to do density estimation and finite-rate encoding (quantization) by $f_n(Z^n)$, and $\phi_n(\cdot)$ decodes an estimated density in $\left\{\phi_n(s) : s \in \tilde{S}_n\right\} \subset \Theta$. At the end, the first stage provides a quantized estimation of $\mu_\theta \in \mathcal{F}$ given by

$$\hat{\theta}_n(Z^n) \equiv \phi_n(f_n(Z^n)) \in \Theta. \tag{14}$$

Using the index $\tilde{s} = f_n(Z^n) \in \tilde{S}_n$, the second stage of $\mathcal{C}^{n,n}$, represented by $\left\{(f_{n,s}, \phi_{n,s}); s \in \tilde{S}_n\right\}$ in (13), encodes and decodes the source data X^n by

$$\mathcal{C}^{n,n}(X^n) \equiv \phi_{n,\tilde{s}}(f_{n,\tilde{s}}(X^n)). \tag{15}$$

In summary, the outcome of the whole encoding process is the concatenation of the bits that represent $f_n(Z^n)$ (first-stage bits), and the bits that represent $f_{n,f_n(Z^n)}(X^n)$ (second-stage bits). The decoding process, on the other hand, reads the first-stage bits to recover $\hat{\theta}_n(Z^n)$ and then reads the second-stage bits to recover $\mathcal{C}^{n,n}(X_1^n)$. (see Figure 1 in which this process is illustrated). The rate (in bits per letter) of $\mathcal{C}^{n,n}$ is

$$R(\mathcal{C}^{n,n}) = \frac{\log_2 |\tilde{S}_n|}{n} + \frac{\log_2 |S_n|}{n}. \tag{16}$$

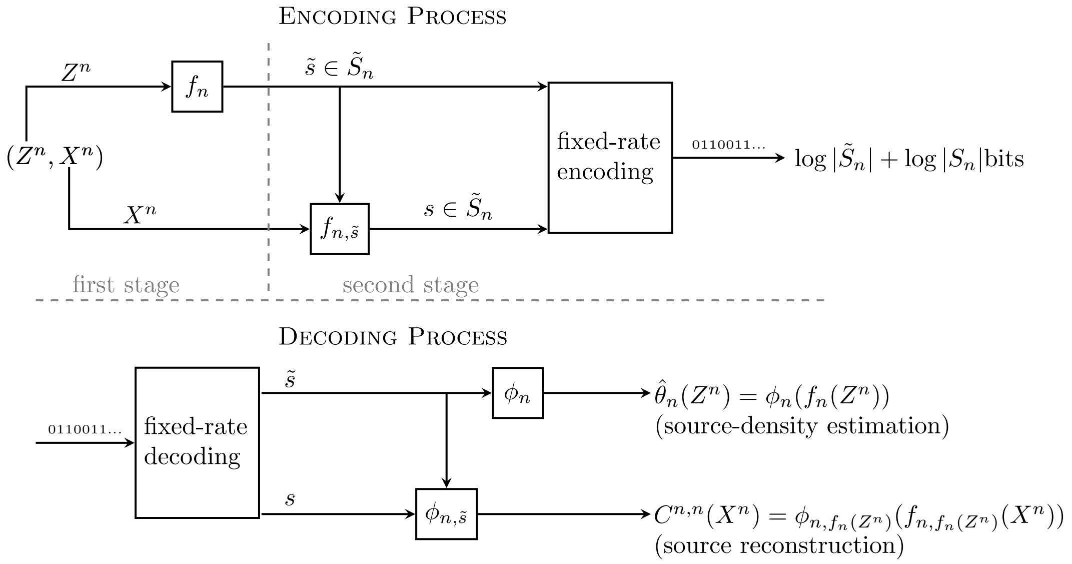

Figure 1. Illustration of Raginsky's two-stage joint source coding and modeling scheme. Top figure illustrates the coding process and the bottom figure shows the respective decoding process.

Based on this two-stage scheme, we could simultaneously achieve source coding and density estimation (modeling) at the decoder. This new joint coding and modeling objective motivates the introduction of the following definition:

Definition 4. *A joint coding and modeling scheme* $\{\mathcal{C}^{n,n} : n \geq 1\}$ *in (13) is strongly minimax universal for a class of distribution* $\mathcal{F} = \{\mu_\theta : \theta \in \Theta\} \subset \mathcal{AC}(\mathbb{X})$ *at the rate* $R > 0$*, if*

- $\lim_{n\to\infty} \sup_{\mu\in\mathcal{F}} \left[D_\mu(\mathcal{C}^{n,n}) - D_\mu^n(R)\right] = 0$*,*
- $\lim_{n\to\infty} \sup_{\mu\in\mathcal{F}} \mathbb{E}_{Z^n\sim\mu^n}(V(\mu_{\hat{\theta}_n(Z^n)}, \mu)) = 0$*, and*
- $\limsup_{n\to\infty} R(\mathcal{C}^{n,n}) \leq R$*.*

Consequently, if $\{\mathcal{C}^{n,n} : n \geq 1\}$ is strongly minimax universal for $\mathcal{F}$, it follows that as n tends to infinity, density estimation is achieved at the decoder (in expected total variations) and, from the source coding perspective, $\{\mathcal{C}^{n,n} : n \geq 1\}$ is strongly finite-block universal for $\mathcal{F}$ in the sense of Definition 3. For the rest of the paper, the strongly minimax universality of Definition 4 will be the main coding and modeling objective.

3. Connections with Zero-Rate Density Estimation

This section formalizes a connection between the objective of joint coding and modeling (declared in Definition 4) and a problem of zero-rate density estimation.

3.1. Density Estimation with a Rate Constraint

Let us first introduce the problem of rate constrained density estimation. Let $\mathcal{F} = \{\mu_\theta : \theta \in \Theta\} \subset \mathcal{AC}(\mathbb{X})$ be an indexed collection of densities as introduced in Section 2.2.

Definition 5. *An $(n, 2^{nR})$ learning rule of length n and rate R for $\mathcal{F}$ is a pair of functions (f, ϕ), with $f : \mathbb{X}^n \to S$ and $\phi : S \to \Theta$, where S is a finite set and*

$$\frac{1}{n} \log_2 |\{f(x^n) : x^n \in \mathbb{X}^n\}| = \frac{1}{n} \log_2 |S| = R. \tag{17}$$

The composition of these two functions $\pi = \phi \circ f : \mathbb{X}^n \to \Theta$ defines the rate-constrained learning rule for $\mathcal{F}$ taking values in the codebook $\{\phi(s) : s \in S\} \subset \Theta$, where $R(\pi) = \log_2(|S|)/n$ denotes its description complexity in bits per training sample.

Definition 6. *The rate $R \geq 0$ is achievable for $\mathcal{F}$, if a learning scheme $\Pi = \{(f_n, \phi_n) : n \geq 1\}$ exists such that*

$$\lim_{n\to\infty} \sup_{\mu \in \mathcal{F}} \mathbb{E}_{Z^n \sim \mu^n}(V(\mu_{\pi_n(Z^n)}, \mu)) = 0 \textit{ and } \limsup_{n\to\infty} R(\pi_n) \leq R, \tag{18}$$

where $Z_1, Z_2 \ldots$ in the left hand side (LHS) of (18) corresponds to i.i.d. realizations driven by $\mu \in \mathcal{F}$. In this case, we say that Π is an R-rate uniformly consistent scheme (or estimator) for the class $\mathcal{F}$.

3.2. Main Results

Proposition 1. *If for a given $R > 0$, $\{\mathcal{C}^{n,n} : n \geq 1\}$ is strongly minimax universal for the class $\mathcal{F}$ at the rate R (Definition 4), then its induced finite-description learning scheme obtained from the first stage in (13), i.e., $\Pi = \{(f_n, \phi_n) : n \geq 1\}$, is a zero-rate uniformly consistent estimator for $\mathcal{F}$ (Definition 6).*

The proof is presented in Section 8.1.

Interestingly, the existence of a zero-rate uniformly consistent scheme for $\mathcal{F}$ is also sufficient to achieve the joint coding and modeling objetive (Definition 4) if some mild conditions are adopted from the work in [18]. This is stated in the following result:

Theorem 1. *Let us assume that*

(i) $\rho : \mathbb{X} \times \hat{\mathbb{X}} \to \mathbb{R}^+$ can be expressed by $\rho(x, \hat{x}) = d(x, \hat{x})^p$ where $d(,)$ is a bounded metric in $\mathbb{X} \cup \hat{\mathbb{X}} \times \mathbb{X} \cup \hat{\mathbb{X}}$ with $p > 0$ and

(ii) for all $\mu \in \mathcal{F}$, for all $n \geq 1$, and for all $R > 0$, there exists a $(0, n)$-block code, say $\mathcal{C}_\mu^{*n}$, that achieves the n-order operational DRF $D_\mu^n(R)$ in (7).

Then the existence of a learning scheme $\Pi = \{(f_n, \phi_n) : n \geq 1\}$ that is zero-rate uniformly consistent for $\mathcal{F}$ implies that $\forall R > 0$ there exists a joint coding and modeling scheme $\{\mathcal{C}^{n,n} : n \geq 1\}$ that is strongly minimax universal for $\mathcal{F}$ at rate R (Definition 4).

The proof is presented in Section 8.2.

Remark 1. *The construction proposed for $\{\mathcal{C}^{n,n} : n \geq 1\}$ at any rate $R > 0$ (in Section 8.2) using the zero-rate density estimation scheme $\Pi = \{\pi_n = \phi_n \circ f_n : n \geq 1\}$ satisfies that:*

$$\sup_{\mu \in \mathcal{F}} \left[D_\mu(\mathcal{C}^{n,n}) - D^n_\mu(R) \right] \leq C \cdot \sup_{\mu \in \mathcal{F}} \mathbb{E}_{Z^n \sim \mu^n}(V(\mu_{\pi_n(Z^n)}, \mu)) \text{ and} \tag{19}$$

$$R(\mathcal{C}^{n,n}) - R \leq R(\pi_n), \tag{20}$$

$\forall n \geq 1$, where $C > 0$ is a constant. It is worth noting that these two inequalities summarize the result in Theorem 1 and, importantly, these two bounds are independent of R.

Remark 2. *An important consequence of the bounds in (19) and (20) is the fact that constructing a learning scheme $\Pi = \{\pi_n : n \geq 1\}$ with specific rates of convergence for $\sup_{\mu \in \mathcal{F}} \mathbb{E}(V(\mu_{\pi_n(Z^n)}, \mu))$ and $R(\pi_n)$ (as n goes to infinity) produces a joint coding and modeling scheme that achieves a uniform rate of convergence to zero (over $\mathcal{F}$) of the overhead in distortion by (19) and a uniform rate of convergence to zero of the overhead in rate by (20). This observation will be used in all the achievable results presented in Sections 4 and 5, where, consequently, the problem reduces to determine Π and expressions for $\sup_{\mu \in \mathcal{F}} \mathbb{E}(V(\mu_{\pi_n(Z^n)}, \mu))$ and $R(\pi_n)$.*

4. Joint Source Coding and Modeling Achievability Results

From the connection with zero-rate density estimation in Section 3, here we present a set of new results for the joint coding and modeling problem of Section 2.3. In these results, the general conditions (i) and (ii) stated in Theorem 1 are assumed.

4.1. Main Result: The Skeleton Density Estimator

Let us first introduce some notions from approximation theory [37].

Definition 7. *Let $\mathcal{F} \subset \mathcal{AC}(\mathbb{X})$ be a class of densities. We say that $\mathcal{F}$ is L_1-totally bounded if for every $\epsilon > 0$, there is a finite set of elements $\{\mu_i : i = 1, ..., N\}$ in $\mathcal{F}$ such that,*

$$\mathcal{F} \subset \bigcup_{i=1}^{N} B^V_\epsilon(\mu_i), \tag{21}$$

where $B^V_\epsilon(\mu) \equiv \{v \in \mathcal{AC}(\mathbb{X}) : V(\mu, v) < \epsilon\}$.

Definition 8. *For $\mathcal{F}$ L_1-totally bounded, let N_ϵ denote the smallest positive integer that achieves the condition in (21). N_ϵ is called the ϵ-covering number of $\mathcal{F}$ and $K(\epsilon) \equiv \log_2(N_\epsilon)$ is called the Kolmogorov's ϵ-entropy of $\mathcal{F}$ [30].*

Definition 9. *An ϵ-covering $\mathcal{G}_\epsilon$ of $\mathcal{F}$ such that $|\mathcal{G}_\epsilon| = N_\epsilon$ is called an ϵ-skeleton of $\mathcal{F}$ [29].*

Theorem 2. *There is a strongly minimax universal joint coding and modeling scheme for $\mathcal{F}$ at rate R for any rate $R > 0$ if, and only if, $\mathcal{F}$ is L_1-totally bounded.*

The proof is presented in Section 8.3.

The achievability part of the proof of Theorem 2 relies on the adoption of the skeleton estimator [29] (with its minimum distance learning principle in (42)), which is a zero-rate uniformly consistent density estimator for $\mathcal{F}$ (Definition 6). Furthermore, Theorem 2 can be complemented saying that the proposed construction $\{\mathcal{C}^{n,n} : n \geq 1\}$ derived from the skeleton estimator satisfies that ($\mathbb{P}_\mu$ is a short-hand for the process distribution of $(Z_n)_{n \geq 1}$ characterized by $\mu \in \mathcal{F}$ under the i.i.d. assumption.)

$$\lim_{n \to \infty} D_\mu(\mathcal{C}^{n,n} | Z^n) = D_\mu(R),\ \mathbb{P}_\mu - \textit{almost surely}, \tag{22}$$

$$\lim_{n \to \infty} V(\mu_{\pi_n(Z^n)}, \mu) = 0,\ \mathbb{P}_\mu - \textit{almost surely}, \tag{23}$$

$\forall \mu \in \mathcal{F}$. The argument is presented in Appendix A.

4.2. Examples of L_1-Totally Bounded Clases

Knowing specific expressions for $K(\epsilon) = \log_2 N_\epsilon < \infty$, the skeleton estimator can be optimized selecting its design parameter appropriately. In particular, the sequence $(\epsilon_n)_{n \geq 1}$ (see details in Section 8.3) is selected as the solution of the optimal balance between estimation and approximation errors (see (45) in Section 8.3), which is given by $\epsilon_n^* \equiv \inf \left\{ \epsilon > 0 : \log(2N_\epsilon^2) \leq \sqrt{n} \right\}$ [30] (Chapter 7.2). The details of this analysis are presented in Section 8.3 and [30] (Chapter 7). By doing so, an optimized zero-rate skeleton scheme $\Pi = \left\{ (f_{\epsilon_n^*}, \phi_{\epsilon_n^*}), n \geq 1 \right\}$, with concrete rate of convergence for $\sup_{\mu \in \mathcal{F}} \mathbb{E}_{Z^n \sim \mu^n}(V(\mu_{\pi_{\epsilon_n^*}(Z^n)}, \mu))$ and $R(\pi_{\epsilon_n^*})$, can be obtained. From Remarks 1 and 2, these results imply specific performance results for the induced joint coding and modeling scheme. To illustrate, we present three interesting examples below.

4.2.1. Finite Mixture Classes

Let $\mathcal{F} = \{\mu_\theta : \theta \in \Theta\}$ with $\Theta = \left\{ \theta \in [0,1]^d : \sum_{k=1}^d \theta_i = 1 \right\}$ be the class of measures which are a convex combination of $\{\mu_1, ..., \mu_d\} \subset AC(\mathbb{X})$, i.e., $\forall \theta \in \Theta, \forall A \in \mathcal{B}(\mathbb{X})$, $\mu_\theta(A) = \sum_{k=1}^d \theta_k \cdot \mu_k(A)$. $\mathcal{F}$ is L_1-totally bounded with $K(\epsilon)$ being $O(d \log(1/\epsilon))$ [30] (Chapter 7.4). From (45) the optimal sequence (ϵ_n^*) is $O(\sqrt{d/n})$ [30], which implies the following finite-rate performance bound [30] (Chapter 7.4):

$$\sup_{\theta \in \Theta} \mathbb{E} \left\{ V(\mu_{\pi_{\epsilon_n^*}(Z^n)}, \mu_\theta) \right\} \leq \sqrt{\frac{Cd \log n}{n}},$$

with C a universal non-negative constant. The rate in bits per-sample $R(\pi_{\epsilon_n^*}) = K(\epsilon_n^*)/n$ is $O(\log n/n)$.

4.2.2. Monotone Densities in $[0,1]^d$

Let $\mathcal{F}$ be the collection of densities with support on $[0,1]^d$, monotonically decreasing per coordinate and bounded by a constant $L > 0$. This class is known to be L_1-totally bounded, and furthermore $K(\epsilon) \leq \frac{CL^d}{\epsilon^d}$ [30] (Lemma 7.1), with the constant C depending only on d. From (45), (ϵ_n^*) being $O(L^{d/d+2}/n^{1/d+2})$ is optimal (please see details in [26,30]) with the following performance bound,

$$\sup_{\mu \in \mathcal{F}} \mathbb{E} \left\{ V(\mu_{\pi_{\epsilon_n^*}(Z^n)}, \mu) \right\} \leq \frac{CL^{d/d+2}}{n^{1/d+2}}.$$

In this case, the rate in bits per sample $R(\pi_{\epsilon_n^*}) = K(\epsilon_n^*)/n$ is $\mathcal{O}(1/n^{2/d+2})$.

4.2.3. r-Moment Smooth Class in $[0,1]$

Let $\mathcal{F}$ be the class of densities defined on the bounded support $[0,1]$, with r absolutely continuous derivatives (with r an integer greater than zero) and satisfying that: $\forall f \in \mathcal{F}$ $\int_{[0,1]} \left| f^{(r+1)} \right| dx \leq C$ for a constant $C > 0$. This class is L_1-totally bounded with $K(\epsilon)$ being $O(1/\epsilon^{r+1})$ [30] (Chapter 7.6). From (45), the optimal sequence (ϵ_n^*) is $O(1/n^{1/3+r})$, where $\sup_{\mu \in \mathcal{F}} \mathbb{E} \left\{ V(\mu_{\pi_{\epsilon_n^*}(Z^n)}, \mu) \right\}$ is $O(1/n^{1/3+r})$ and the rate in bits per sample $R(\pi_{\epsilon_n^*}) = K(\epsilon_n^*)/n$ is $O(1/n^{2/3+r})$.

Notably, the last two examples are fully non-parametric, where $K(\epsilon)$ is a polynomial function of $1/\epsilon$. Richer non-parametric examples of L_1-totally bounded clases of densities, where $K(\epsilon)$ is even exponentially in $1/\epsilon$, are presented in [30] (Chapters 7.6 and 7.8) and its references.

4.3. Yatracos Classes with Finite VC Dimension

Looking at the distortion redundancy bound in (19), when $\mathcal{F}$ is totally bounded the fastest rate of convergence that could be achieved with the skeleton estimator proposed in Theorem 2 is $O(\sqrt{1/n})$ (see Section 8.3 and the estimation error bound in (45)). In this section, more specific density collections are studied to achieve this best rate $O(\sqrt{1/n})$ for density estimation and distortion redundancy from (19). We follow the path proposed by Yatracos in [38], who explored families of distributions with a finite Vapnik and Chervonenkis (VC) dimension the so-called VC classes [39,40]. Let us first introduce some definitions:

Definition 10 ([38])**.** *Let $\mathcal{F} = \{\mu_\theta : \theta \in \Theta\} \subset \mathcal{AC}(\mathbb{X})$ be an indexed collection of densities. The Yatracos class for such a collection is given by*

$$\mathcal{A}_\Theta = \left\{ A_{\theta,\bar{\theta}} : \theta, \bar{\theta} \in \Theta, \theta \neq \bar{\theta} \right\}, \tag{24}$$

where $A_{\theta,\bar{\theta}} \equiv \left\{ x \in \mathbb{X} : g_{\mu_\theta}(x) > g_{\mu_{\bar{\theta}}}(x) \right\} \in \mathcal{B}(\mathbb{X})$ is the Scheffé set of μ_θ with respect to $\mu_{\bar{\theta}}$, as defined in (2).

Theorem 3. *Let us assume that*

(i) $\mathcal{F}$ is L_1-totally bounded,

(ii) the Yatracos class $\mathcal{A}_\Theta$ has a finite VC dimension (Definition A1 in Appendix B), and

(iii) the Kolmogorov's entropy of $\mathcal{F}$ associated with the sequence $\epsilon_n = 1/\sqrt{n}$ grows strictly sub-linearly, i.e., $\log_2(N_{1/\sqrt{n}})$ is $o(n)$,

then there is a zero-rate density estimator scheme $\Pi = \{(f_n, \phi_n) : n \geq 1\}$ for $\mathcal{F}$ such that

$$\sup_{\mu \in \mathcal{F}} \mathbb{E}_{Z^n \sim \mu^n} \left\{ V(\mu_{\pi_n(Z^n)}, \mu) \right\} \text{ is } O(1/\sqrt{n}),$$

where $\pi_n(Z^n) = \phi_n(f_n(Z^n))$ is the skeleton estimator in (42) with $\epsilon_n = 1/\sqrt{n}$. Furthermore, Π is also a zero-rate strongly consistent density estimator where $\forall \mu \in \mathcal{F}$

$$V(\mu_{\pi_n(Z^n)}, \mu) \text{ is } \mathcal{O}(\sqrt{\log n/n}),\ \mathbb{P}_\mu - \text{almost surely.}$$

The proof is presented in Section 8.4.

From Definition 7, $\log_2(N_\epsilon)$ is inversely proportional to ϵ. In fact, depending of how rich $\mathcal{F}$ is, $\log_2(N_\epsilon)$ can go from being $O(\log 1/\epsilon)$, passing from being polynomial in $1/\epsilon$, to being $O(e^{1/\epsilon})$ (see a number of examples in [30] (Chapter 7) and its references). Then the role of (iii) in the statement of Theorem 3 is to bound how fast N_ϵ should tend to infinity as ϵ goes to zero, to guarantee a zero-rate in the skeleton learning scheme. It is simple to show that N_ϵ being $O(e^{(1/\epsilon)^q})$ with $q \in [0, 2)$ is sufficient to achieve that $\log_2(N_{1/\sqrt{n}})$ is $o(n)$. This is a condition satisfied by a rich collection of L_1-totally bounded classes in $\mathcal{AC}(\mathbb{X})$. Concrete examples are presented in [30] (Chapter 7).

5. The Parametric Scenario

The results presented so far are of theoretical interest because they rely on the skeleton estimator that is constructed from the skeleton covering of $\mathcal{F}$ (see Definition 9), which is unknown in practice. Moving towards making the zero-rate skeleton learning scheme of practical interest, we revisit the important parametric scenario in which Θ, the index set of $\mathcal{F}$, is a compact set contained in a finite-dimensional Euclidean space $\mathbb{R}^k$. Interestingly, in this context we can consider a practical covering of $\mathcal{F}$ induced by the uniform partition of the parameter space Θ, as used in [18]. Unlike [18], where a minimum-distance estimate is first found and then quantized, here we first quantize the space Θ and then find the minimum-distance estimate among a finite collection of candidates (i.e., over a finte number of prototypes in Θ). Some assumptions will be needed.

Definition 11 ([18])**.** *Let $\mathcal{F} = \{\mu_\theta : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^k$. Let $I_{\mathcal{F}} : \Theta \to \mathcal{F}$ be the index function of $\mathcal{F}$ that maps θ to μ_θ. $I_{\mathcal{F}}$ is said to be locally uniformly Lipschitz, if there exists $r > 0$ and $m > 0$, such $\forall \theta \in \Theta$, $\forall \phi \in B_r(\theta)$,*

$$V(\mu_\theta, \mu_\phi) \leq m \, ||\theta - \phi|| \, , \tag{25}$$

where $B_r(\theta) \subset \Theta$ denotes the ball of radius r (with respect to the Euclidean norm in $\mathbb{R}^k$) centered at θ.

The following lemma shows that $\mathcal{F}$ is L_1-totally bounded under some parametric assumptions.

Lemma 1. *Let $\mathcal{F} = \{\mu_\theta : \theta \in \Theta\} \subset \mathcal{P}(\mathbb{X})$ with $\Theta \subset \mathbb{R}^k$. If Θ is bounded ($\exists L > 0$ such that $\Theta \subset \bigotimes_{i=1}^{k}[-L, L]$) and the mapping $I_{\mathcal{F}} : \Theta \to \mathcal{F}$ is locally uniformly Lipschitz (Definition 11), then $\mathcal{F}$ is L_1-totally bounded. Furthermore, N_ϵ is $O(1/\epsilon^k)$ for this family.*

The proof is presented in Section 8.5.

It is important to note that the ϵ-covering of $\mathcal{F}$ used in the proof of Lemma 1 to derive an upper bound for N_ϵ is practical (see Appendix C). This offers the possibility of implementing a practical skeleton estimator, which is the focus of the following result.

The Practical Skeleton Estimator

Under the assumptions of Lemma 1, let $(\tilde{f}_{n,\epsilon}, \tilde{\phi}_{n,\epsilon})$ denote the learning rule of length n associated with the minimum-distance principle in (42) with parameter ϵ (see details in Section 8.3), where instead of using the ϵ-skeleton $\mathcal{G}_\epsilon$ of $\mathcal{F}$ (in Definition 9), the implementable (see Appendix C) ϵ-covering of Θ presented in the proof of Lemma 1 is used. This practical ϵ-covering is denoted by $\tilde{\mathcal{G}}_\epsilon$ (by definition, $N_\epsilon = |\mathcal{G}_\epsilon| \leq \left|\tilde{\mathcal{G}}_\epsilon\right| = \tilde{N}_\epsilon \sim O(1/\epsilon^k)$, this last part from Lemma 1.). With this, let $\tilde{\Pi}((\epsilon_n)_{n\geq 1}) \equiv \left\{(\tilde{f}_{n,\epsilon_n}, \tilde{\phi}_{n,\epsilon_n}) : n \geq 1\right\}$ denote our practical learning scheme indexed by the precision numbers $(\epsilon_n)_{n\geq 1} \in (\mathbb{R}^+)^{\mathbb{N}}$. We are in a position to integrate Theorem 3 and Lemma 1 to state the following:

Theorem 4. *Under the assumptions of Lemma 1, the practical skeleton estimator $\tilde{\Pi}((\epsilon_n)_{n\geq 1})$ with $\epsilon_n^* = 1/\sqrt{n}$ satisfies that*

$$\sup_{\mu_\theta \in \mathcal{F}} \mathbb{E}_{Z^n \sim \mu^n} \left\{V(\mu_{\tilde{\pi}_{n,\epsilon_n^*}(Z^n)}, \mu_\theta)\right\} \text{ is } \mathcal{O}(\sqrt{\log n/n}), \text{ and } R(\tilde{\pi}_{n,\epsilon_n^*}) \text{ is } \mathcal{O}(\log n/n), \tag{26}$$

where $\tilde{\pi}_{n,\epsilon}(Z^n) \equiv \tilde{\phi}_{n,\epsilon}(\tilde{f}_{n,\epsilon}(Z^n))$.

In addition, if the Yatracos collection $\mathcal{A}_\Theta = \left\{A_{\theta,\bar{\theta}} : \theta, \bar{\theta} \in \Theta, \theta \neq \bar{\theta}\right\}$ has a finite VC dimension equal to J, then

$$\sup_{\mu_\theta \in \mathcal{F}} \mathbb{E}_{Z^n \sim \mu^n} \left\{V(\mu_{\tilde{\pi}_{n,\epsilon_n^*}(Z^n)}, \mu_\theta)\right\} \text{ is } \mathcal{O}(1/\sqrt{n}), \text{ and } R(\tilde{\pi}_{n,\epsilon_n^*}) \text{ is } \mathcal{O}(\log n/n). \tag{27}$$

The proof is presented in Section 8.6.

When $\mathbb{X} \subset \mathbb{R}^d$, Raginsky [18] showed that the finite VC dimension assumption of Theorem 4 is satisfied by the class of mixture families presented in Section 4.2.1 and a rich collection of exponential families of the form $\mathcal{F} = \{\mu_\theta : \theta \in \Theta\} \subset \mathcal{P}(\mathbb{X})$ with $\frac{d\mu_\theta}{d\lambda}(x) = f(x) \cdot e^{\sum_{i=1}^k \theta_i h_i(x) - g(\theta)}$, $\forall x \in \mathbb{X}$, where $f(x)$ is a reference density, $\{h_i(\cdot) : i = 1, ..., k\}$ is a set of arbitrary real-valued functions, $g(\theta)$ is a normalization constant ($g(\theta) = \ln \int_{\mathbb{X}} e^{\sum_{i=1}^k \theta_i h_i(x)} f(x)dx$ see details in [18] (Section V)), and Θ is a compact subset of $\mathbb{R}^k$ (see details in [18] (Section V)).

6. Summary of the Results

We summarize the results of the proposed zero-rate density estimation approach adopted for the problem of joint fixed-rate lossy source coding and modeling of continuous memoryless sources.

- Proposition 1 and Theorem 1 formalize the interplay between the two-stage joint fixed-rate coding and modeling objective and the problem of zero-rate uniformly consistent (in expected total variation) density estimation.

- Theorem 2 establishes a necessary and sufficient condition on a family of densities for the existence of a strongly minimax joint coding and modeling scheme achieving both source coding and model identification objectives (Definition 4). The result is obtained for the rich non-parametric collection of L_1-totally bounded densities.
- For the modeling stage, we propose using the skeleton estimator, which first quantizes the data and then finds the minimum-distance decision on this finite set of density candidates (42). This is a practical solution in the sense that the inference (minimization) is carried out over a finite set.
- By introducing combinatorial regularity conditions on the family of distributions $\mathcal{F} = \{\mu_\theta : \theta \in \Theta\}$, the skeleton scheme achieves $\mathcal{O}(1/\sqrt{n})$ rate of convergence in the n-order distortion redundancy, and the same rate in the expected total variational distance for the modeling part (Theorem 3).
- Finally, for a relevant parametric setting, a practical skeleton-based joint coding and modeling scheme is proposed that achieves a rate of $\mathcal{O}(1/\sqrt{n})$ for the n-order distortion redundancy (Theorem 4). This rate is slightly better than the $\mathcal{O}(\sqrt{\log n/n})$ achieved in [18] under the same rate overhead of $\mathcal{O}(\log(n)/n)$. Furthermore, Theorem 4 removes the finite-VC-dimension assumption over the Yatracos class $\mathcal{A}_\Theta$ considered in [18] (Theorem 3.2), while achieving the same performance rates in terms of n-order distortion redundancy $\mathcal{O}(\sqrt{\log n/n})$, uniform expected risk to learn the density $\mathcal{O}(\sqrt{\log n/n})$, and rate overhead $\mathcal{O}(\log n/n)$.

Concerning the last parametric result, we note that the result in [18] can be improved by the adoption of Dudley's entropy bound [41], which would yield the same asymptotic rate reported in this work for the n-order distortion redundancy.

A final remark is that under the bounded distortion metric assumption of Theorem 1 condition (i), Linder et al. [14] (Theorem 2) showed that $\forall \theta \in \Theta$, and for every $R > 0$ such that $D_{\mu_\theta}(R) > 0$, there is a constant $K_\theta(R) > 0$ such that

$$D^n_{\mu_\theta}(R) - D_{\mu_\theta}(R) \leq (K_\theta(R) + r_n)\sqrt{\frac{\log n}{n}}, \tag{28}$$

where (r_n) is a sequence that converges to zero ($o(1)$) uniformly in Θ. This result offers a rate of convergence of the n-order operational distortion-rate function to the Shannon DRF as the block length tends to infinity. In view of (11), we can adopt this result in Theorems 3 and 4, to say that the average distortion of the respective joint coding and modeling schemes at rate R, i.e., $D_\mu(\mathcal{C}^{n,n})$, convergences to the Shannon DRF $D_\mu(R)$ as $O(\sqrt{\frac{\log n}{n}})$ point-wise $\forall \mu \in \mathcal{F}$. Therefore in the process of comparing $D_\mu(\mathcal{C}^{n,n})$ with the Shannon DR function, we lose the $O(\sqrt{1/n})$ rate of convergence.

7. Conclusions

This work revisits the problem of fixed-rate universal lossy source coding and model identification with training data proposed in [18] from a learning perspective. Remarkably, we found that the problem is equivalent to the problem of density estimation of the source distribution with some concrete but non-conventional operational data-rate constraints in bits per sample. This learning problem can be seen as the task of estimating and encoding the distribution of samples with a zero-rate in bits per sample, while achieving a consistent estimation in expected total variations of the distribution after the decoding process. From our perspective, the rate-constraint density estimation problem is interesting in itself and can have relevant applications in other contexts such as distributed learning scenarios and sensor network problems.

Importantly for the joint coding and modeling problem, the connection with density estimation provides a context for the use of the skeleton estimator proposed by Yatracos in [29]. We highlight two important implications from its use. First, we extend results about minimax universality from the parametric context explored in [30] to the rich non-parametric family of L_1-totally bounded densities [26,30]. This result significantly expands the contexts where the joint model and coding objective can be achieved. We illustrated this with some examples in Section 4.2 and many more can be found in the literature of density estimation [26,30].

Second, in the parametric case studied in [18], we were able to remove some of the assumptions and obtain not only the same performance result in terms of rate of convergence of the n-order distortion redundancy but also slightly better convergence results. Therefore, the Skeleton estimator, though essentially a non-parametric learning scheme, is shown to be instrumental in enriching the applicability of the joint coding and modeling framework.

8. Proofs of Results

8.1. *Proposition 1*

Proof. The fact that Π is uniformly consistent for $\mathcal{F}$ is directly from Definition 4. On the other hand, the rate of $\pi_n = \phi_n \circ f_n$ is $R(\pi_n) = \frac{1}{n}\log_2 \left|\tilde{S}_n\right|$. From the definition of $D^n_\mu(R)$, it is simple to show from the strict monotonicity of $D_\mu(R)$ that in order for $\lim_{n\to\infty} \sup_{\mu\in\mathcal{F}} \left[D_\mu(\mathcal{C}^{n,n}) - D^n_\mu(R)\right] = 0$, it is required that $\limsup_{n\to\infty} \frac{1}{n}\log|S_n| > R - \epsilon$ for any $\epsilon > 0$. Then, from (16), and since $\log|\tilde{S}_n|/n = R(\pi_n)$, $\limsup_{n\to\infty} R(\mathcal{C}^{n,n}) \leq R$ implies that $\lim_{n\to\infty} R(\pi_n) = 0$. □

8.2. *Theorem 1*

Proof. The proof builds upon the ideas elaborated in [18] (Theorem 3.2, p. 3065). Let us consider an arbitrary $R > 0$ and let $\Pi = \{(f_n, \phi_n) : n \geq 1\}$ be the zero-rate learning scheme of the assumption. Using Π, let us construct the joint coding and modeling rule of length n by:

$$\begin{aligned} \mathcal{C}^{n,n} = \Big(&f_n : \mathbb{X}^n \to \tilde{S}_n, \phi_n : \tilde{S}_n \to \Theta, \\ &\left\{ f_{n,\tilde{s}} : \mathbb{X}^n \to S_n, \phi_{n,\tilde{s}} : S_n \to \hat{\mathbb{X}}^n : \tilde{s} \in \tilde{S}_n \right\}\Big). \end{aligned} \tag{29}$$

Concerning the first stage of $\{\mathcal{C}^{n,n} : n \geq 1\}$, it is induced directly from the coding-decoding rules of Π. For the second stage, $\forall n \geq 1$, $\forall \tilde{s} \in \tilde{S}_n$ the pair $(f_{n,\tilde{s}}, \phi_{n,\tilde{s}})$ is picked such that $\mathcal{C}^{*n}_{\mu_{\theta_{n,\tilde{s}}}} = \phi_{n,\tilde{s}} \circ f_{n,\tilde{s}}$, which is the optimal n-block code that achieves $D^n_{\mu_{\theta_{n,\tilde{s}}}}(R)$ (from the hypothesis in (ii)), with $\theta_{n,\tilde{s}} \equiv \phi_n(f_n(\tilde{s}))$ short-hand for the reproduction codeword induced from the first stage-pair (f_n, ϕ_n), and S_n satisfying the R-rate constraint, i.e., $|S_n| = 2^{nR}$. From construction and the fact that Π has zero-rate,

$$\lim_{n\to\infty} R(\mathcal{C}^{n,n}) = R + \lim_{n\to\infty} \log_2 \left|\tilde{S}_n\right|/n = R,$$

then $\{\mathcal{C}^{n,n} : n \geq 1\}$ satisfies the rate condition. On the other hand, based on the assumption that Π is zero-rate uniformly consistent, it follows that

$$\lim_{n\to\infty} \sup_{\mu\in\mathcal{F}} \mathbb{E}(V(\mu_{\hat{\theta}_n(Z^n)}, \mu)) = 0, \tag{30}$$

where $\hat{\theta}_n(Z^n) = \phi_n(f_n(Z^n))$. Then $\{\mathcal{C}^{n,n} : n \geq 1\}$ achieves the modeling objective. Concerning the coding objective, we use the following key result:

Lemma 2 ([18] (Lemma C.1)). *Let P and Q be two probability measures in $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$. Let $\mathcal{C}^n = (f, \phi)$ be a zero-memory n-block coder with the nearest neighbor property (i.e., $\mathcal{C}^n$ is nearest neighbor if, $\forall x_1^n \in \mathbb{X}^n$, $\phi(f(x_1^n)) = \arg\min_{\hat{x}_1^n \in \Gamma_{\mathcal{C}^n}} \rho(x_1^n, \hat{x}_1^n)$ with $\Gamma_{\mathcal{C}^n}$ the reproduction codebook of $\mathcal{C}^n$.). If we denote the performance of $\mathcal{C}^n$ ($\mathcal{C}^n = \phi \circ f$) with respect to P by*

$$D_P(\mathcal{C}^n) \equiv \frac{1}{n}\mathbb{E}_{X^n \sim P^n}\left(\rho(\mathcal{C}^n(X^n), X^n)\right), \tag{31}$$

where P^n denotes the product measure with marginal P in $(\mathbb{X}^n, \mathcal{B}(\mathbb{X}^n))$, and ρ satisfies the condition i) of Theorem 1 and is bounded by d_{max}, then

$$\left| D_P(\mathcal{C}^n)^{1/p} - D_Q(\mathcal{C}^n)^{1/p} \right| \leq 2^{1/p} d_{max} \cdot V(P, Q). \tag{32}$$

Furthermore, the inequality can be extended for the n-order operational distortions in (7), i.e.,

$$\left| D_P^n(R)^{1/p} - D_Q^n(R)^{1/p} \right| \leq 2^{1/p} d_{max} \cdot V(P, Q), \tag{33}$$

$\forall R > 0$.

Let us work with the following distortion redundancy,

$$\begin{aligned} D_\mu(\mathcal{C}^{n,n}|Z^n) - D_\mu^n(R) = & \left[\frac{1}{n} \mathbb{E}_{X^n \sim \mathbb{P}_\mu^n} (\rho^n(X^n, \mathcal{C}^{n,n}(X^n))) - D^n_{\mu_{\hat{\theta}_n(Z^n)}}(R) \right] + \\ & \left[D^n_{\mu_{\hat{\theta}_n(Z^n)}}(R) - D_\mu^n(R) \right] & (34) \\ & \leq D_\mu(\mathcal{C}^{*n}_{\mu_{\hat{\theta}_n(Z^n)}}) - D^n_{\mu_{\hat{\theta}_n(Z^n)}}(R) + 2^{1/p} d_{max} \cdot V(\mu_{\hat{\theta}_n(Z^n)}, \mu) & (35) \\ & = D_\mu(\mathcal{C}^{*n}_{\mu_{\hat{\theta}_n(Z^n)}}) - D_{\mu_{\hat{\theta}_n(Z^n)}}(\mathcal{C}^{*n}_{\mu_{\hat{\theta}^n(Z^n)}}) \\ & + 2^{1/p} d_{max} \cdot V(\mu_{\hat{\theta}_n(Z^n)}, \mu) & (36) \\ & \leq 2^{1/p+1} d_{max} \cdot V(\mu_{\hat{\theta}_n(Z^n)}, \mu). & (37) \end{aligned}$$

For the first equality we use (5). The inequality in (35) is from the definition in (31) and (33), and the equality in (36) is from the construction of $\mathcal{C}^{*n}_{\mu_{\hat{\theta}_n(Z^n)}}$ which is n-operational optimal for the distribution $\mu_{\hat{\theta}_n(Z^n)}$ at rate R. Finally, (37) is from (32).

Concluding, $D_\mu(\mathcal{C}^{n,n}|Z^n) - D_\mu^n(R)$ is random (a measurable function of Z^n) and dominated by $V(\mu_{\hat{\theta}_n(Z^n)}\mu)$. Hence taking the expected value (with respect to Z^n) on both sides of this inequality (see (6)), we have the uniform convergence in (30) implying that

$$\lim_{n \to \infty} \sup_{\mu \in \mathcal{F}} \left[D_\mu(\mathcal{C}^{n,n}) - D_\mu^n(R) \right] = 0, \tag{38}$$

and then the coding objective is achieved. □

8.3. Theorem 2

Proof. Let us first assume that $\mathcal{F}$ is L_1-totally bounded and prove the direct part of the statement. We adopt the skeleton estimate proposed by Yatracos [29] and extended by Devroye et al. [42,43] (a complete presentation can be found in [30] (Chapter 7)). For any arbitrary $\epsilon > 0$, let us consider the ϵ-skeleton $\mathcal{G}_\epsilon = \left\{ \mu_{\theta_i^\epsilon} : i = 1, ..., N_\epsilon \right\}$ of $\mathcal{F}$. We use $g_{\theta_i^\epsilon}(x) \equiv \frac{d\mu_{\theta_i^\epsilon}}{d\lambda}(x)$ as short-hand for the i-th pdf in $\mathcal{G}_\epsilon$, and we define

$$\Theta_\epsilon \equiv \{\theta_i^\epsilon : i = 1, ..., N_\epsilon\} \subset \Theta$$

to represent the index set of $\mathcal{G}_\epsilon$. Let us consider the Yatracos class of $\mathcal{G}_\epsilon$ given by [30]

$$\mathcal{A}_\epsilon \equiv \left\{ A_{i,j}^\epsilon, A_{j,i}^\epsilon : 1 \leq i < j \leq N_\epsilon \right\}, \tag{39}$$

where $A^{\epsilon}_{i,j} = \left\{ x \in \mathbb{X} : g_{\theta^{\epsilon}_i}(x) > g_{\theta^{\epsilon}_j}(x) \right\} \in \mathcal{B}(\mathbb{X})$ is the Scheffé set of $\mu_{\theta^{\epsilon}_i}$ with respect to $\mu_{\theta^{\epsilon}_j}$ in (2) [30,33]. Hence, given i.i.d. realizations $X_1, ..., X_n$ with $X_i \sim \mu_\theta$ ($\mu_\theta \in \mathcal{F}$), let us propose the encoder-decoder pair $(f_{n,\epsilon}, \phi_{n,\epsilon})$ associated with $\mathcal{A}_\epsilon$ by,

$$f_{n,\epsilon}(X^n) \equiv \arg \min_{i \in \{1,...,N_\epsilon\}} \sup_{B \in \mathcal{A}_\epsilon} \left| \mu_{\theta^{\epsilon}_i}(B) - \hat{\mu}_n(B) \right| \in [N_\epsilon], \tag{40}$$

$$\phi_{n,\epsilon}(i) \equiv \theta^{\epsilon}_i \in \Theta_\epsilon \subset \Theta, \tag{41}$$

where $\hat{\mu}_n(B) = \sum_{j=1}^{n} \mathbf{1}_B(X_j)$ is the standard empirical distribution. In this context,

$$\hat{\theta}_\epsilon(X^n) = \phi_{n,\epsilon}((f_{n,\epsilon}(X^n))) = \arg \min_{\theta^{\epsilon}_i \in \Theta_\epsilon} \sup_{B \in \mathcal{A}_\epsilon} \left| \mu_{\theta^{\epsilon}_i}(B) - \hat{\mu}_n(B) \right|, \tag{42}$$

is the well-known skeleton estimate [29]. $\hat{\theta}_\epsilon(X^n_1)$ is the minimum-distance approximation of $\hat{\mu}_n$ with elements of $\mathcal{G}_\epsilon$ [29,30], adopting the measure in the right-hand-side of (42) that is reminiscent of the total variational distance in (1). In order to choose a sequence $(\epsilon_n)_{n \geq 1}$, we consider the following performance bound.

Lemma 3 ([30] (Theorem 6.3))**.** *For any $\mu \in \mathcal{F}$,*

$$V(\mu_{\hat{\theta}_\epsilon(X^n)}, \mu) \leq 3 \min_{v \in \mathcal{G}_\epsilon} V(v, \mu) + 4 \sup_{B \in \mathcal{A}_\epsilon} |\hat{\mu}_n(B) - \mu(B)| \,. \tag{43}$$

Equation (43) is valid for any $\epsilon > 0$ and, consequently, it provides a trade-off between an approximation error term and an estimation error term. The approximation error is $\min_{v \in \mathcal{G}_\epsilon} V(v, \mu)$, which is bounded by the definition of $\mathcal{G}_\epsilon$. For the estimation error, on the other hand, Yatracos proposed the use of Hoeffding's inequality [44] to obtain that $\forall \mu \in \mathcal{P}(\mathbb{X})$ [30] (Theorem 7.1),

$$\mathbb{E}_{X^n \sim \mu^n} \left(\sup_{B \in \mathcal{A}_\epsilon} |\hat{\mu}_n(B) - \mu(B)| \right) \leq \sqrt{\frac{\log(2N^2_\epsilon)}{2n}}. \tag{44}$$

Using (44) in (43), it follows that, $\sup_{\mu_\theta \in \mathcal{F}} \mathbb{E}\left\{ V(\mu_{\hat{\theta}_\epsilon(X^n)}, \mu_\theta) \right\} \leq 3\epsilon + \sqrt{\frac{8 \log(2N^2_\epsilon)}{n}}$. This last expression is distribution-free and it is valid if the approximation fidelity ϵ is a chosen function of n [30]. Consequently, for any sequence $(\epsilon_n)_{n \geq 1}$,

$$\sup_{\mu_\theta \in \mathcal{F}} \mathbb{E}\left\{ V(\mu_{\hat{\theta}_{\epsilon_n}(X^n)}, \mu_\theta) \right\} \leq 3\epsilon_n + \sqrt{\frac{8 \log(2N^2_{\epsilon_n})}{n}}, \tag{45}$$

for all $n \geq 1$. Hence, we consider $\epsilon^*_n \equiv \inf\left\{ \epsilon > 0 : \log(2N^2_\epsilon) \leq \sqrt{n} \right\}$ proposed in [30] (Chapter 7.2), which is well-defined and converges to zero as n tends to infinity. Consequently from (45), $\lim_{n \to \infty} \sup_{\mu_\theta \in \mathcal{F}} \mathbb{E}\left\{ V(\mu_{\hat{\theta}_{\epsilon^*_n}(X^n)}, \mu_\theta) \right\} = 0$. Then the learning scheme $\Pi((\epsilon^*_n)_{n \geq 1}) \equiv \left\{ (f_{n,\epsilon^*_n}, \phi_{n,\epsilon^*_n}) : n \geq 1 \right\}$ satisfies the learning requirement in Definition 6, where in particular $R(\phi_{n,\epsilon^*_n} \circ f_{n,\epsilon^*_n}) = \frac{\log_2(N_{\epsilon^*_n})}{n}$ is $O(1/\sqrt{n})$ by construction. To conclude the argument of this part (i.e., presenting the construction of the second stage of a joint coding & modeling scheme), we adopt the result and the construction presented in the proof of Theorem 1 (see Remark 1 for details). This result implies that $\forall R > 0$ there is a strongly minimax universal joint coding and modeling scheme for $\mathcal{F}$ at rate R.

For the other implication (the converse part of the statement), let us fix $R > 0$ and assume that we have a joint coding & modeling scheme that is strongly minimax universal (Definition 4) for $\mathcal{F}$

at rate R. Then from Proposition 1, we have a learning scheme $\Pi = \{(f_n, \phi_n) : n \geq 1\}$ such that $\lim_{n\to\infty} R(\pi_n = \phi_n \circ f_n) = 0$ and

$$\lim_{n\to\infty} \sup_{\mu\in\mathcal{F}} \mathbb{E}_{\mathbb{P}^n_\mu} \left\{ V(\mu_{\pi_n(X^n)}, \mu) \right\} = 0. \tag{46}$$

For the learning rule of length n, we have its reproduction codebook that we denote by $\Theta^n \equiv \left\{ \theta^n_j : j = 1, ..., 2^{nR(\pi_n)} \right\} \subset \Theta$. Let us define the minimum-distance oracle solution in Θ^n by

$$\tilde{\theta}_n(\mu) = \arg \inf_{\theta\in\Theta^n} V(\mu_\theta, \mu). \tag{47}$$

From (46), we have that $\lim_{n\to\infty} \sup_{\mu\in\mathcal{F}} V(\mu_{\tilde{\theta}_n(\mu)}, \mu) = 0$. In other words, $\forall\epsilon > 0$, there exists $N(\epsilon) < \infty$, such that for all $n \geq N(\epsilon)$, $V(\mu_{\tilde{\theta}_n(\mu)}, \mu) < \epsilon$ uniformly for every element $\mu \in \mathcal{F}$. This means that $\forall\epsilon > 0$ there exists $N(\epsilon) < \infty$, such that for any arbitrary $\bar{n} > N(\epsilon)$, $\mathcal{F} \subset \bigcup_{\theta\in\Theta^{\bar{n}}} B_\epsilon(\mu_\theta)$, where by construction $|\Theta^{\bar{n}}| < \infty$. Then $\mathcal{F}$ is totally bounded, which concludes the proof. □

8.4. Theorem 3

Proof. From Lemma 3, for any arbitrary sequence $(\epsilon_n)_{n\geq 1}$

$$V(\mu_{\hat{\theta}_{\epsilon_n}(X^n)}, \mu_\theta) \leq 3\epsilon_n + 4 \sup_{B\in\mathcal{A}_{\epsilon_n}} |\hat{\mu}_n(B) - \mu_\theta(B)|\,. \tag{48}$$

with $\mathcal{A}_{\epsilon_n}$ the Yatracos class of the skeleton $\mathcal{G}_{\epsilon_n}$. It is clear that $\forall\epsilon > 0$, $\mathcal{A}_\epsilon \subset \mathcal{A}_\Theta$. Then by monotonicity $\mathbb{E}\left(\sup_{B\in\mathcal{A}_\epsilon} |\hat{\mu}_n(B) - \mu(B)|\right) \leq \mathbb{E}\left(\sup_{B\in\mathcal{A}_\Theta} |\hat{\mu}_n(B) - \mu(B)|\right)$, for all $\epsilon > 0$ and for any distribution $\mu \in \mathcal{P}(\mathbb{X})$. Here is where we use the assumption that $\mathcal{A}_\Theta$ has finite VC dimension J, which implies from [30] (Theorem 3.1) that

$$\sup_{\mu\in\mathcal{F}} \mathbb{E}\left(\sup_{B\in\mathcal{A}_\Theta} |\hat{\mu}_n(B) - \mu(B)| \right) \leq c\sqrt{\frac{J}{n}} \tag{49}$$

for some constant $c > 0$. Substituting this result in (48), the argument concludes by replacing $(\epsilon_n) = (1/\sqrt{n})$, a solution which achieves the intended rate of convergence for $\sup_{\mu_\theta\in\mathcal{F}} \mathbb{E}\left\{ V(\mu_{\hat{\theta}_{1/\sqrt{n}}(X^n)}, \mu_\theta) \right\}$. Finally, the rate of the learning rule is $\frac{\lceil \log_2(N_{1/\sqrt{n}}) \rceil}{n}$, which tends to zero by the last hypothesis.

For the almost-sure convergence part if $\epsilon^*_n = \frac{1}{\sqrt{n}}$, it is sufficient to show that the second term in the right hand side (RHS) of (48) is $O(\sqrt{\log n/n})$ $\mathbb{P}_\mu$-almost surely. From the fact that $\mathcal{A}_\Theta$ has finite VC dimension (Definition A1), and from the classical VC inequality [30] (Corollary 4.1 and Theorem 3.1) and [45] (Chapter 12.4), it follows that

$$\mathbb{P}\left(\sup_{B\in\mathcal{A}_{\epsilon^*_n}} |\hat{\mu}_n(B) - \mu_\theta(B)| > \delta \right) \leq 8(n+1)^J \cdot e^{-\frac{n\delta^2}{32}},$$

$\forall n \geq 0$ and $\forall\epsilon > 0$. Then considering $a_n = \sqrt{\log n/n}$ and $M^2/32 > J + 2$,

$$\mathbb{P}\left(\sup_{B\in\mathcal{A}_{\epsilon^*_n}} |\hat{\mu}_n(B) - \mu_\theta(B)| > M \cdot a_n \right) \leq 8\frac{(n+1)^J}{n^{M^2/32}} \leq \frac{K}{n^2}$$

for some $K > 0$, hence $\sum_{n\geq 0} \mathbb{P}\left(\frac{1}{a_n} \cdot \sup_{B\in\mathcal{A}_{\epsilon^*_n}} |\hat{\mu}_n(B) - \mu_\theta(B)| > M \right) < \infty$. Then from the Borel Cantelli Lemma, $\limsup_{n\to\infty} \frac{1}{a_n} \cdot \sup_{B\in\mathcal{A}_{\epsilon^*_n}} |\hat{\mu}_n(B) - \mu_\theta(B)| \leq M$ $\mathbb{P}_\mu$-almost surely, which concludes the proof. As (a_n) is $o(1)$, this result implies the almost-sure convergences to zero of $V(\mu_{\hat{\theta}_{\epsilon_n^*}(X^n)}, \mu_\theta)$ as n goes to infinity.

Finally, using similar arguments, it is possible to show that $V(\mu_{\hat{\theta}_{\epsilon_n^*}(X^n)}, \mu_\theta)$ is $o(1/n^\tau)$ $\mathbb{P}_\mu$-almost surely for any $\tau \in (0, 1/2)$. □

8.5. Lemma 1

Proof. First note that Θ is contained in a compact set $\bigotimes_{i=1}^k [-L, L] \subset \mathbb{R}^k$, consequently, Θ inherits the finite covering property of a compact set, i.e., $\forall \epsilon > 0$, there exists a finite covering $\Theta^\epsilon = \left\{\theta_1^\epsilon, ., \theta_{K(\epsilon)}^\epsilon\right\} \subset \Theta$ such that,

$$\Theta \subset \bigcup_{\theta \in \Theta} B_\epsilon(\theta) = \bigcup_{i=1}^{K(\epsilon)} B_\epsilon(\theta_i^\epsilon). \tag{50}$$

On the other hand, from the locally uniformly Lipschitz assumption on $I_{\mathcal{F}} : \Theta \to \mathcal{F}$, there exists $r > 0$ and $m > 0$ such that $V(\mu_\theta, \mu_\phi) \leq m\,||\theta - \phi||$, $\forall \theta \in \Theta$, $\forall \phi \in B_r(\theta)$. Then, by considering $\epsilon_o < r$, it follows by construction of Θ^{ϵ_o} that

$$\mathcal{F} \subset \bigcup_{i=1}^{K(\epsilon_o)} \mathcal{I}_{\mathcal{F}}(B_{\epsilon_o}(\theta_i^{\epsilon_o})) \subset \bigcup_{i=1}^{K(\epsilon_o)} B^V_{m\cdot r}(\mu_{\theta_i^{\epsilon_o}}), \tag{51}$$

where $B^V_\delta(\mu) = \{v \in \mathcal{P}(\mathbb{X}) : V(v, \mu) < \delta\}$ is the ball centered at $\mu \in \mathcal{P}(\mathbb{X})$, induced from the total variational distance, and the last inequality stems from the Lipschitz condition. Hence, from (51), $\forall \epsilon > 0$ there exists $M(\epsilon) = K(\min\{\epsilon/m, r\}) < \infty$ and $\left\{\mu_1^\epsilon, ..., \mu_{M(\epsilon)}^\epsilon\right\} \subset \mathcal{P}(\mathbb{X})$, such that $\mathcal{F} \subset \bigcup_{i=1}^{M(\epsilon)} B(\mu_{\theta_i^\epsilon}, \epsilon)$, which proves the result.

For the final part, let (m, r) be the uniform parameters that characterize the Lipschitz condition of $I_{\mathcal{F}}(\cdot)$ (Definition 11). Without loss of generality, let us assume the critical regime where $\frac{\epsilon}{m} < r$, hence from (51) N_ϵ is upper bounded by $K(\epsilon/m)$, which is the covering number of Θ. As $\Theta \subset \bigotimes_{i=1}^k [-L, L] \subset \mathbb{R}^k$, we will work with a uniform partition of $\bigotimes_{i=1}^k [-L, L]$ to find a bound for $K(\epsilon/m)$. Let $\bar{\epsilon} = \frac{\epsilon}{m}$, then inducing a product-type partition, where in each coordinate we have $\lceil \frac{L\sqrt{k}}{\bar{\epsilon}} \rceil$ uniform length cells, we have the required $\bar{\epsilon}$-covering. The number of prototypes is $O(\frac{(L\sqrt{k})^k}{\bar{\epsilon}^k})$, which is $O(1/\epsilon^k)$ as a function of ϵ ($\epsilon = \bar{\epsilon} \cdot m$).

To clarify the constructive nature of the ϵ-covering used to prove this result, an algorithm with the basic steps of the construction of this practical covering is sketched in Appendix C. □

8.6. Theorem 4

Proof. Let $\tilde{\mathcal{G}}_\epsilon \subset \mathcal{F}$ be the ϵ-covering induced from the uniform partition of Θ presented in Lemma 1. From this we can construct the minimum-distance estimate in (42) adopting the Yatracos class of $\tilde{\mathcal{G}}_\epsilon$ (with index set $\tilde{\Theta}_\epsilon$), i.e., $\tilde{\mathcal{A}}_\epsilon$, which, from (39), yields

$$\tilde{\theta}_\epsilon(X^n) \equiv \arg\min_{\theta_i^\epsilon \in \tilde{\Theta}_\epsilon} \sup_{B \in \tilde{\mathcal{A}}_\epsilon} \left| \mu_{\theta_i^\epsilon}(B) - \hat{\mu}_n(B) \right|. \tag{52}$$

Considering $\epsilon_n = 1/\sqrt{n}$, from (45) it follows that

$$\sup_{\mu_\theta \in \mathcal{F}} \mathbb{E}\left\{V(\mu_{\tilde{\theta}_{\epsilon_n}(X^n)}, \mu_\theta)\right\} \leq \frac{3}{\sqrt{n}} + \sqrt{\frac{8\log(2\log\left|\tilde{\mathcal{G}}_{1/\sqrt{n}}\right|^2)}{n}}.$$

The latter upper bound is asymptotically dominated by $(\sqrt{\log n/n})$ from the fact that $\log\left|\tilde{\mathcal{G}}_{1/\sqrt{n}}\right|$ is $O(k\log(n))$ (Lemma 1), which proves the assertions made in (26).

Concerning part (ii), using the arguments presented in the proof of Theorem 3, we can obtain that $\forall \epsilon > 0$,

$$\sup_{\mu_\theta \in \mathcal{F}} \mathbb{E}\left\{ V(\mu_{\tilde{\theta}_\epsilon(X^n)}, \mu_\theta) \right\} \leq 3\epsilon + 4 \cdot c\sqrt{\frac{J}{n}}. \tag{53}$$

From this point, the proof follows from the arguments of Theorem 3 and the fact that $\log_2 \left| \tilde{\mathcal{G}}_{1/\sqrt{n}} \right|$ is $O(k/2 \cdot \log_2 n)$. $\square$

Author Contributions: Conceptualization, J.F. Silva and M.S. Derpich; Methodology, J.F. Silva and M.S. Derpich; Formal Analysis, J.F. Silva and M.S. Derpich; Investigation and Results, J.F. Silva and M.S. Derpich; Writing—Original Draft Preparation, J.F. Silva and M.S. Derpich; Writing—J.F. Silva & Editing, M.S. Derpich; Project Administration, J.F. Silva; Funding Acquisition, J.F. Silva.

Funding: The work is supported by funding from FONDECYT Grants 1170854 and 1171059, CONICYT-Chile and the Advanced Center for Electrical and Electronic Engineering (AC3E), Basal Project FB0008. In addition, J.F. Silva acknowledges support from Project Anillos ACTI 1405, CONICYT-Chile.

Acknowledgments: We want to thank the anonymous reviewers for their constructive comments that were instrumental to improve the technical content and organization of this work. We thank Diane Greenstein for editing and proofreading all this material and Sebastian Espinosa for preparing Figure 1.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of (22) and (23)

First, we show that the zero-rate skeleton estimate $\Pi((\epsilon_n)) = \{(f_{n,\epsilon_n}, \phi_{n,\epsilon_n}) : n \geq 1\}$ proposed in (40) and (41) is also strongly consistent.

Proposition A1. $\Pi((\epsilon_n^*)) = \left\{(f_{n,\epsilon_n^*}, \phi_{n,\epsilon_n^*}) : n \geq 1\right\}$ *is strongly consistent, i.e., for any* $\mu \in \mathcal{F}$,

$$\lim_{n\to\infty} V(\mu_{\hat{\theta}_{\epsilon_n^*}(X^n)}, \mu) = 0,\ \mathbb{P}_\mu\text{-almost surely.}$$

Proof. Let us consider the skeleton estimate $\mu_{\hat{\theta}_{\epsilon_n^*}(X^n)}$, where the sequence was chosen by the rule $\epsilon_n^* = \inf\left\{\epsilon > 0 : \log(2N_\epsilon^2) \leq \sqrt{n}\right\}$. Then $\log N_{\epsilon_n^*}^2 \leq (\sqrt{n} - \log 2) \leq \sqrt{n}$ for all n. From Lemma 3, $V(\mu_{\hat{\theta}_{\epsilon_n^*}(X^n)}, \mu) \leq 3\epsilon_n^* + 4 \sup_{B \in \mathcal{A}_{\epsilon_n^*}} |\hat{\mu}_n(B) - \mu(B)|$. As by construction (ϵ_n^*) is $o(1)$, we just need to concentrate on the estimation error term. Applying Hoeffding's inequality [44] $\forall \delta > 0$,

$$\mathbb{P}\left(\sup_{B \in \mathcal{A}_{\epsilon_n^*}} |\hat{\mu}_n(B) - \mu(B)| > \delta \right) \leq 2 \cdot N_{\epsilon_n^*}^2 \cdot e^{-2n\delta^2} \leq 2e^{(\sqrt{n}/\log e - 2n\delta^2)}, \tag{A1}$$

where from the Borel-Cantelli lemma [46,47], the estimation error convergences to zero almost-surely. $\square$

Finally considering the inequality in (37), we have that $D_\mu(\mathcal{C}^{n,n}|Z^n) - D_\mu^n(R) \leq 2^{1/p+1} d_{max} \cdot V(\mu_{\hat{\theta}_n(Z^n)}, \mu)$, $\forall \mu \in \mathcal{F}$, which concludes the argument.

Appendix B. Basic Definitions of Vapnik and Chervonenkis Theory

Let $\mathcal{C} \subset \mathcal{B}(\mathbb{X})$ be a collection of measurable events, and $x^n = (x_1, ..., x_n)$ be a sequence of n points in $\mathbb{X}^n$. Then we define by $\mathcal{S}(\mathcal{C}, x^n)$ the number of different sets in

$$\{\{x_1, x_2, ..., x_n\} \cap B : B \in \mathcal{C}\},$$

and the shatter coefficient of $\mathcal{C}$ by [40,45]

$$S_n(\mathcal{C}) = \sup_{x^n \in \mathbb{X}^n} \mathcal{S}(\mathcal{C}, x^n). \tag{A2}$$

The shatter coefficient is an indicator of the richness of $\mathcal{C}$ to dichotomize a finite sequence of points in the space, where by definition $S_n(\mathcal{C}) \leq 2^n$.

Definition A1. *The first time (in the index n) where $S_n(\mathcal{C})$ is strictly less than 2^n is called the Vapnik and Chervonenkis (VC) dimension of $\mathcal{C}$ [45]. If $\mathcal{C}$ has a finite VC dimension then it is called a* VC class; *otherwise if $S_n(\mathcal{C}) = 2^n$ $\forall n \geq 1$, then the class is said to have an infinite VC-dimension.*

Appendix C. Pseudo Algorithm to Implement the Practical ϵ-Covering Presented in Lemma 1

Under the parametric assumptions of Lemma 1, we recognize four structural parameters that characterize $\mathcal{F}$: k the dimension of the Euclidean space that contains Θ, $L > 0$ associated with the assumption that $\Theta \subset \bigotimes_{i=1}^{k}[-L, L]$, and (r, m) the parameters associated with the locally Lipschitz assumption of $I_{\mathcal{F}}$. Given these four parameters (k, L, m, r) and $\epsilon > 0$, there is a constructive ϵ-covering presented in the proof of Lemma 1 that can be implemented in the following steps:

1. In each of the k dimensions of Θ, the interval $[-L, L]$ is partitioned uniformly with sub-intervals of length $2\epsilon/(m\sqrt{k})$. This produces a scalar quantization of $[-L, L]$ with $\lceil m\sqrt{k}L/\epsilon \rceil$ prototypes per coordinate.
2. A product partition of $\bigotimes_{i=1}^{k}[-L, L]$ is made with the scalar quantizations of the previous step. From the proof of Lemma 1, this is a ϵ/m-covering of Θ with $K = \lceil m\sqrt{k}L/\epsilon \rceil^k$ prototypes. Let us denote this set by $\{\theta_i, i = 1, ..., K\} \subset \Theta$.
3. From the proof of Lemma 1, the covering of Θ constructed in the previous step induces an ϵ-covering of $\mathcal{F}$ by applying the indexing function $I_{\mathcal{F}}$, i.e., by $\{I_{\mathcal{F}}(\theta_i) : i = 1, ..., K\}$.

References

1. Csiszár, I.; Shields, P.C. *Information Theory and Statistics: A Tutorial*; Now Inc.: Houston, TX, USA, 2004.
2. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley Interscience: New York, NY, USA, 2006.
3. Gyorfi, L.; Pali, I.; van der Meulen, E. There is no unieversal soruce code for an infinite source alphabet. *IEEE Trans. Inf. Theory* **1994**, *40*, 267–271. [CrossRef]
4. Davisson, L.D. Universal noiseless coding. *IEEE Trans. Inf. Theory* **1973**, *19*, 783–785. [CrossRef]
5. Kieffer, J.C. A unified approach to weak universal source coding. *IEEE Trans. Inf. Theory* **1978**, *24*, 674–682. [CrossRef]
6. Rissanen, J. Universal coding, information, prediction, and estimation. *IEEE Trans. Inf. Theory* **1984**, *30*, 629–636. [CrossRef]
7. Boucheron, S.; Garivier, A.; Gassiat, E. Codign on countable infininite alphabets. *IEEE Trans. Inf. Theory* **2009**, *55*, 358–373. [CrossRef]
8. Bontemps, D.; Boucheron, S.; Gassiat, E. About adaptive coding on countable alphabets. *IEEE Trans. Inf. Theory* **2014**, *60*, 808–821. [CrossRef]
9. Bontemps, D. Universal coding on infinite alphabets: exponentially decreasing envelops. *IEEE Trans. Inf. Theory* **2011**, *57*, 1466–1478. [CrossRef]
10. Silva, J.F.; Piantanida, P. Almost lossless variable-length source coding on countably infinite alphabets. In Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016; pp. 1–5. [CrossRef]
11. Silva, J.F.; Piantanida, P. The redundancy gains of almost lossless universal source coding over envelope families. In Proceedings of the IEEE International Symposium on Information Theory, Aachen, Germany, 25–30 June 2017; pp. 1–5.
12. Silva, J.F.; Piantanida, P. Universal weak variable-length source coding on countable infinite alphabets. *arXiv* **2017**, arXiv:1708.08103.
13. Berger, T.; Gibson, J.D. Lossy source coding. *IEEE Trans. Inf. Theory* **1998**, *44*, 2693–2723. [CrossRef]
14. Linder, T.; Lugosi, G.; Zeger, K. Rates of convergence in the source codign theorem, in empirical quantization design, and in univesal lossy source codign. *IEEE Trans. Inf. Theory* **1994**, *40*, 1728–1740. [CrossRef]
15. Linder, T.; Lugosi, G.; Zeger, K. Fixed-rate universal lossy soruce coding and rate of convergence for memoryless sources. *IEEE Trans. Inf. Theory* **1995**, *41*, 665–676. [CrossRef]

16. Neuhoff, D.L.; Gray, R.M.; Davisson, L.D. Fixed rate universal block source coding with a fidelity criterion. *IEEE Trans. Inf. Theory* **1975**, *21*, 511–523. [CrossRef]
17. Ziv, J. Coding of sources with unkown statistics-Part II: Distortion relative to a fidelity criterion. *IEEE Trans. Inf. Theory* **1972**, *18*, 389–394. [CrossRef]
18. Raginsky, M. Joint fixed-rate univesal lossy coding and identification of continuous-alphabet memoryless sources. *IEEE Trans. Inf. Theory* **2008**, *54*, 3059–3077. [CrossRef]
19. Chou, P.; Effros, M.; Gray, R.M. A vector quantization approach to universal noiseless coding and quantization. *IEEE Trans. Inf. Theory* **1996**, *42*, 1109–1138. [CrossRef]
20. Rissanen, J. Stochastic complexity and modeling. *Ann. Stat.* **1986**, *14*, 1080–1100. [CrossRef]
21. Barron, A.; Rissanen, J.; Yu, B. The minimun description lenght principle in coding and modeling. *IEEE Trans. Inf. Theory* **1998**, *44*, 2743–2760. [CrossRef]
22. Barron, A.; Györfi, L.; van der Meulen, E.C. Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Trans. Inf. Theory* **1992**, *38*, 1437–1454. [CrossRef]
23. Tao, G. *Adaptive Control Design and Analysis*; Wiley-IEEE Press: Hoboken, NJ, USA, 2003.
24. Berger, T. *Rate Distortion Theory*; Prentice Hall: Upper Saddle River, NJ, USA, 1971.
25. Devroye, L.; Györfi, L. *Nonparametric Density Estimation: The L_1 View*; Wiley Interscience: New York, NY, USA, 1985.
26. Devroye, L.; Györfi, L. *Principles of Nonparametric Learning*; Chapter Distribution and Density Estimation; Springer: New York, NY, USA, 2001.
27. Shannon, C.E. Coding theorems for a discrete source with fidelity criterion. *IRE Int. Conv. Rec.* **1959**, *4*, 325–350.
28. Gallager, R.G. *Information Theory and Realiable Communication*; John Wiley & Songs: Hoboken, NJ, USA, 1968.
29. Yatracos, Y.G. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Ann. Stat.* **1985**, *13*, 768–774. [CrossRef]
30. Devroye, L.; Lugosi, G. *Combinatorial Methods in Density Estimation*; Springer: New York, NY, USA, 2001.
31. Silva, J.F.; Derpich, M.S. Necessary and sufficient conditions for zero-rate density estimation. In Proceedings of the Information Theory Workshop (ITW), Paraty, Brazil, 16–20 October 2011.
32. Halmos, P.R. *Measure Theory*; Van Nostrand: New York, NY, USA, 1950.
33. Scheffé, H. A useful convergence theorem for probability distribution. *Ann. Math. Stat.* **1947**, *18*, 434–458.
34. Gersho, A.; Gray, R. *Vector Quantization and Signal Compression*; Kluwer Academic: Norwell, MA, USA, 1992.
35. Gray, R.; Neuhoff, D. Quantization. *IEEE Trans. Inf. Theory* **1998**, *44*, 2325–2384. [CrossRef]
36. Gray, R.M. *Entropy and Information Theory*; Springer: New York, NY, USA, 1990.
37. Kolmogorov, A.N.; Tikhomirov, V.M. ϵ-emtropy and ϵ-capacity of sets in function spaces. *Transl. Am. Math. Soc.* **1961**, *17*, 277–364.
38. Yatracos, Y.G. A note on L_1 consistent estimation. *Can. J. Stat.* **1988**, *16*, 283–292.
39. Vapnik, V.; Chervonenkis, A.J. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **1971**, *16*, 264–280. [CrossRef]
40. Vapnik, V. *Statistical Learning Theory*; John Wiley: Hoboken, NJ, USA, 1998.
41. Dudley, R.M. Central limits theorems for empirical measures. *Ann. Probab.* **1978**, *6*, 899–929. [CrossRef]
42. Devroye, L.; Lugosi, G. A universally acceptable smoothing factor for kernel density estimation. *Ann. Stat.* **1996**, *24*, 2499–2512.
43. Devroye, L.; Lugosi, G. Nonasymtotic universal smothing factors, kernel complexity and Yatracos classes. *Ann. Stat.* **1997**, *25*, 2626–2637.
44. Hoeffding, W. Probability inequalities for sums of bounded random variables. *J. Am. Stat. Assoc.* **1963**, *58*, 13–30. [CrossRef]
45. Devroye, L.; Györfi, L.; Lugosi, G. *A Probabilistic Theory of Pattern Recognition*; Springer: New York, NY, USA, 1996.
46. Breiman, L. *Probability*; Addison-Wesley: Boston, MA, USA, 1968.
47. Varadhan, S. *Probability Theory*; American Mathematical Society: Providence, RI, USA, 2001.