Joint Fixed-Rate Universal Lossy Coding and Identification of Continuous-Alphabet Memoryless Sources

Maxim Raginsky, *Member, IEEE*

Abstract—The problem of joint universal source coding and density estimation is considered in the setting of fixed-rate lossy coding of continuous-alphabet memoryless sources. For a wide class of bounded distortion measures, it is shown that any compactly parametrized family of $\mathbb{R}^d$-valued independent and identically distributed (i.i.d.) sources with absolutely continuous distributions satisfying appropriate smoothness and Vapnik–Chervonenkis (VC) learnability conditions, admits a joint scheme for universal lossy block coding and parameter estimation, such that when the block length n tends to infinity, the overhead per-letter rate and the distortion redundancies converge to zero as $O(n^{-1} \log n)$ and $O(\sqrt{n^{-1} \log n})$, respectively. Moreover, the active source can be determined at the decoder up to a ball of radius $O(\sqrt{n^{-1} \log n})$ in variational distance, asymptotically almost surely. The system has finite memory length equal to the block length, and can be thought of as blockwise application of a time-invariant nonlinear filter with initial conditions determined from the previous block. Comparisons are presented with several existing schemes for universal vector quantization, which do not include parameter estimation explicitly, and an extension to unbounded distortion measures is outlined. Finally, finite mixture classes and exponential families are given as explicit examples of parametric sources admitting joint universal compression and modeling schemes of the kind studied here.

Index Terms—Learning, minimum-distance density estimation, two-stage codes, universal vector quantization, Vapnik–Chervonenkis (VC) dimension.

I. INTRODUCTION

IN a series of influential papers [1]–[3], Rissanen has elucidated and analyzed deep connections between universal lossless coding and statistical modeling. His approach hinges on the following two key insights.

1) A given parametric class of information sources admits universal lossless codes if a) the statistics of each source in the class (or, equivalently, the parameters of the source) can be determined with arbitrary precision from a sufficiently long data sequence and if b) the parameter space can be partitioned into a finite number of subsets, such that the

sources whose parameters lie in the same subset are "equivalent" in the sense of requiring "similar" optimal coding schemes. This idea extends naturally to hierarchical model classes (e.g., when the dimension of the parameter vector is unknown), provided that the parametric family of sources governed by each model satisfies the above regularity conditions individually.

2) Given a sequence of symbols emitted by an information source from a hierarchical model class, an asymptotically correct model of the source is obtained by finding the best tradeoff between the number of bits needed to describe it and the number of bits needed to losslessly encode the data assuming that the data are drawn from the maximum-likelihood distribution relative to this model. This is the basis of the so-called minimum description length (MDL) principle for model selection and, more generally, statistical inference (see, e.g., the survey article of Barron, Rissanen, and Yu [4] or the recent book by Grünwald [5]).

There is, in fact, a natural symmetry between these two insights, owing to the well-known one-to-one correspondence between (almost) optimal lossless codes and probability distributions on the space of all input sequences [6]. For this reason, when considering universal lossless coding, we can use the term "model" to refer either to the probability distribution of the source or to an optimal lossless code for the source, where we allow codes with ideal (noninteger) codeword lenghts. The main point of Rissanen's approach is precisely that the objectives of source coding and modeling can be accomplished jointly and in an asymptotically optimal manner.

Consider the case of a parametric class of sources where the parameter space has finite dimension k. Then the redundancy of the corresponding universal lossless code (i.e., the excess average code length relative to the optimal code for the actual source at a given block length) is controlled essentially by the number of bits required to describe the source parameters to the decoder. In particular, the achievability theorem of Rissanen [1, Theorem 1b] states that one can use a scheme of this kind to achieve the redundancy of about $(k/2) \log n / n$ bits per symbol, where n is the block length. The universal lossless coder used by Rissanen in [1] operates as follows: first, the input data sequence is used to compute the maximum-likelihood estimate of the parameters of the source, then the estimate is quantized to a suitable resolution, and finally, the data are encoded with the corresponding optimum lossless code. Structurally, this is an example of a *two-stage code*, in which the binary description of the input data sequence produced by the encoder consists of two parts:

the first part describes the (quantized) maximum-likelihood estimate of the source parameters, while the second part describes the data using the code matched to the estimated source.

In this paper, we investigate the achievable redundancies of schemes for joint source coding and identification (modeling) in the setting of fixed-rate universal *lossy* block coding (vector quantization) of continuous-alphabet memoryless sources. Once we pass from lossless codes to lossy ones, the term "model" can refer either to a probabilistic description of the source or to a probability distribution over codebooks in the *reproduction* space. In particular, whereas a lossless code for an information source is equivalent to a probabilistic model of the source, a *lossy* code corresponds to a discrete probability distribution over sequences in the *reproduction* alphabet, and is thus related to the *source* distribution only indirectly. However, there are situations in which one would like to compress the source and identify its statistics at the same time. For instance, in *indirect adaptive control* (see, e.g., Tao [7, Ch. 7]) the parameters of the plant (the controlled system) are estimated on the basis of observation, and the controller is modified accordingly. Consider the discrete-time stochastic setting, in which the plant state sequence is a random process whose statistics are governed by a finite set of parameters. Suppose that the controller is geographically separated from the plant and connected to it via a noiseless digital channel whose capacity is R bits per use. Then, given the time horizon T, the objective is to design an encoder and a decoder for the controller to obtain reliable estimates of both the plant parameters and the plant state sequence from the 2^{TR} possible outputs of the decoder. In this paper, we are concerned with modeling the actual source directly, and not through a codebook distribution in the reproduction space. To place the present work in a wider context, in Section VI we shall briefly comment on the line of research concerned with relating lossy codes to codebook models, which can be thought of as a lossy variant of the MDL principle.

The objective of universal lossy coding (see, e.g., [8]–[13]) is to construct lossy block source codes (vector quantizers) that perform well in incompletely or inaccurately specified statistical environments. Roughly speaking, a sequence of vector quantizers is universal for a given class of information sources if it has asymptotically optimal performance, in the sense of minimizing the average distortion under the rate constraint, on any source in the class. Two-stage codes have also proved quite useful in universal lossy coding [10], [11], [13]. For instance, the two-stage universal quantizer introduced by Chou, Effros, and Gray [13] is similar in spirit to the adaptive lossless coder of Rice and Plaunt [14], [15], known as the "Rice machine": each input data sequence is encoded in parallel with a number of codes, where each code is matched to one of the finitely many "representative" sources, and the code that performs the best on the given sequence (in the case of lossy codes, compresses it with the smallest amount of distortion) wins. Similar to the setting of Rissanen's achievability theorem, the approach of [13] assumes a sufficiently smooth dependence of optimum coding schemes on the parameters of the source. However, the decision rule used in selection of the second-stage code does not rely on explicit modeling of the source statistics as the second-stage code is chosen on the basis of local (pointwise), rather than

average, behavior of the data sequence with respect to a fixed collection of quantizers. This approach emphasizes the coding objective at the expense of the modeling objective, thus falling short of exhibiting a relation between the latter and the former.

In the present work, we consider parametric spaces $\{P_\theta\}$ of independent and identically distributed (i.i.d.) sources with values in $\mathbb{R}^d$, such that the P_θ's are absolutely continuous and the parameter θ belongs to a bounded subset of $\mathbb{R}^k$. We show in a constructive manner that, for a wide class of bounded distortion functions and under certain regularity conditions, such parametric families admit universal sequences of quantizers with distortion redundancies[1] converging to zero as $O(\sqrt{n^{-1}\log n})$ and with an overhead per-letter rate converging to zero as $O(n^{-1}\log n)$, as the block length $n \to \infty$. These convergence rates are, more or less, typical for universal coding schemes relying on explicit or implicit acquisition of the statistical model of the source (cf. the discussion in Section IV of this paper). For unbounded distortion functions satisfying a certain moment condition with respect to a fixed reference letter, the distortion redundancies are shown to converge to zero as $O(\sqrt[4]{n^{-1}\log n})$. The novel feature of our method, however, is that the decoder can use the two-stage binary description of the data not only to reconstruct the data with asymptotically optimal fidelity, but also to identify the active source up to a variational ball of radius $O(\sqrt{n^{-1}\log n})$ with probability approaching unity. In fact, the universality and the rate of convergence of the compression scheme are directly tied to the performance of the source identification procedure.

While our approach parallels Rissanen's method for proving his achievability result from [1], there are two important differences with regard to both his work on lossless codes and subsequent work by others on universal lossy codes. The first difference is that the maximum-likelihood estimate, which fits naturally into the lossless framework, is no longer appropriate in the lossy case. In order to relate coding to modeling, we require that the probability distributions of the sources under consideration behave smoothly as functions of the parameter vectors; for compactly parametrized sources with absolutely continuous probability distributions, this smoothness condition is stated as a local Lipschitz property in terms of the L_1 distance between the probability densities of the sources and the Euclidean distance in the parameter space. For bounded distortion measures, this implies that the expected performance of the corresponding optimum coding schemes also exhibits smooth dependence on the parameters. (By contrast, Chou, Effros, and Gray [13] impose the smoothness condition directly on the optimum codes. This point will be elaborated upon in Section III.) Now, one can construct examples of sources with absolutely continuous probability distributions for which the maximum-likelihood estimate behaves rather poorly in terms of the L_1 distance between the true and the estimated probability densities [16]. Instead, we propose the use of the so-called *minimum-distance estimate*, introduced by Devroye and Lugosi [17], [18] in the context of kernel density estimation. The introduction of the minimum-distance estimate allows us to draw upon the powerful machinery of Vapnik–Chervonenkis (VC) theory (see, e.g., [19] and Appendix A in this paper) both for estimating the convergence rates of density es-

---

[1]The distortion redundancy of a lossy block code relative to a source is the excess distortion of the code compared to the optimum code for that source.

timates and distortion redundancies, as well as for characterizing the classes of sources that admit joint universal coding and identification schemes. The merging of VC techniques with two-stage coding further underscores the forward relation between statistical learning/modeling and universal lossy coding.

The second difference is that, unlike previously proposed schemes, our two-stage code has nonzero memory length. The use of memory is dictated by the need to force the code selection procedure to be blockwise causal and robust to local variations in the behavior of data sequences produced by "similar" sources. For a given block length n, the stream of input symbols is parsed into contiguous blocks of length n, and each block is quantized with a quantizer matched to the source with the parameters estimated from the preceding block. In other words, the coding process can be thought of as blockwise application of a nonlinear time-invariant filter with initial conditions determined by the preceding block. In the terminology of Neuhoff and Gilbert [20], this is an instance of a *block-stationary causal source code*.

The remainder of the paper is organized as follows. In Section II, we state the basic notions of universal lossy coding specialized to block codes with finite memory. Two-stage codes with memory are introduced in Section III and placed in the context of statistical modeling and parameter estimation. The main result of this paper, Theorem 3.2, is also stated and proved in Section III. Next, in Section IV, we present comparisons of our two-stage coding technique with several existing techniques, as well as discuss some generalizations and extensions. In Section V, we show that two well-known types of parametric sources—namely, mixture classes and exponential families—satisfy, under mild regularity requirements, the conditions of our main theorem and thus admit joint universal quantization and identification schemes. Section VI offers a quick summary of the paper, together with a list of potential topics for future research. Appendix A contains a telegraphic summary of notions and results from VC theory. Appendices B, C, and D are devoted to proofs of certain technical results used throughout the paper.

II. PRELIMINARIES

Let $\{X_i\}_{i=-\infty}^{\infty}$ be a memoryless stationary source with alphabet $\mathcal{X}$ (the *source alphabet*), i.e., the X_i's are i.i.d. random variables with values in $\mathcal{X}$. Suppose that the common distribution of the X_i's belongs to a given indexed class $\{P_\theta : \theta \in \Theta\}$ of probability measures on $\mathcal{X}$ (with respect to an appropriate σ-field). The distributions on the n-blocks $X^n = (X_1, \ldots, X_n)$ will be denoted by P_θ^n. The superscript n will be dropped whenever it is clear from the argument, such as in $P_\theta(x^n)$. Expectations with respect to the corresponding process distributions will be denoted by $\mathbb{E}_\theta[\cdot]$, e.g., $\mathbb{E}_\theta[X^n]$. In this paper, we assume that $\mathcal{X}$ is a Borel subset of $\mathbb{R}^d$, although this qualification is not required in the rest of the present section.

Consider coding $\{X_i\}$ into another process $\{\hat{X}_i\}$ with alphabet $\hat{\mathcal{X}}$ (the *reproduction alphabet*) by means of a finite-memory stationary block code. Given any $m, n, t \in \mathbb{Z}$ with $m, n \geq 1$, let $X_m^n(t)$ denote the segment

$$(X_{tn-m+1}, X_{tn-m+2}, \ldots, X_{tn})$$

of $\{X_i\}$. When $n = m$, we shall abbreviate this notation to $X^n(t)$; when $m = 1$, we shall write $X_n(t)$; finally, when $t = 1$, we shall write X_m^n, X^n, X_m. A code with block length n and memory length m (or an (n, m)-block code, for short) is then described as follows. Each reproduction n-block $\hat{X}^n(t), t \in \mathbb{Z}$, is a function of the corresponding source n-block $X^n(t)$, as well as of $X_m^n(t-1)$, the m source symbols immediately preceding $X^n(t)$, and this function is independent of t:

$$\hat{X}^n(t) = C^{n,m}\left(X^n(t), X_m^n(t-1)\right), \quad \forall t \in \mathbb{Z}.$$

When the code has zero memory, i.e., $m = 0$, we shall denote it more compactly by C^n. The performance of the code is measured in terms of a *single-letter distortion* (or *fidelity criterion*), i.e., a measurable map $\rho : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$. The loss incurred in reproducing a string $x^n \in \mathcal{X}^n$ by $\hat{x}^n \in \hat{\mathcal{X}}^n$ is given by

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

When the statistics of the source are described by P_θ, the average per-letter distortion of $C^{n,m}$ is defined as

$$D_\theta(C^{n,m}) \triangleq \limsup_{k \to \infty} \frac{1}{k} \mathbb{E}_\theta[\rho(X^k, \hat{X}^k)]$$

$$= \limsup_{k \to \infty} \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}_\theta[\rho(X_i, \hat{X}_i)]$$

where the $\hat{X}_i$'s are determined from the rule $\hat{X}^n(t) = C^{n,m}(X^n(t), X_m^n(t-1))$ for all $t \in \mathbb{Z}$. Since the source is i.i.d., hence stationary, for each $\theta \in \Theta$, both the reproduction process $\{\hat{X}_i\}$ and the pair process $\{(X_i, \hat{X}_i)\}$ are n-stationary, i.e., the vector processes $\{X^n(t)\}_{t=-\infty}^{\infty}$ and $\{(X^n(t), \hat{X}^n(t))\}_{t=-\infty}^{\infty}$ are stationary [20]. This implies [21] that

$$D_\theta(C^{n,m}) = \frac{1}{n} \mathbb{E}_\theta[\rho(X^n, \hat{X}^n)]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta[\rho(X_i, \hat{X}_i)]$$

where $\hat{X}^n = C^{n,m}(X^n, X_m^n(0))$.

More specifically, we shall consider fixed-rate lossy block codes (also referred to as vector quantizers). A fixed-rate lossy (n, m)-block code is a pair (f, ϕ) consisting of an encoder $f : \mathcal{X}^n \times \mathcal{X}^m \to \mathcal{S}$ and a decoder $\phi : \mathcal{S} \to \hat{\mathcal{X}}^n$, where $\mathcal{S} \subset \{0, 1\}^*$ is a collection of fixed-length binary strings. The quantizer function $C^{n,m} : \mathcal{X}^n \times \mathcal{X}^m \to \hat{\mathcal{X}}^n$ is the composite map $\phi \circ f$; we shall often abuse notation, denoting by $C^{n,m}$ also the pair (f, ϕ). The number $R(C^{n,m}) = n^{-1} \log |\mathcal{S}|$ is called the *rate* of $C^{n,m}$, in bits per letter (unless specified otherwise, all logarithms in this paper will be taken to base 2). The set $\Gamma = \{\phi(s) : s \in \mathcal{S}\}$ is the *reproduction codebook* of $C^{n,m}$.

The optimum performance achievable on the source P_θ by *any* finite-memory code with block length n is given by the *nth-order operational distortion-rate function* (DRF)

$$\hat{D}_\theta^{n,*}(R) \triangleq \inf_m \inf_{C^{n,m}} \{D_\theta(C^{n,m}) : R(C^{n,m}) \leq R\}$$

where the infimum is over all finite-memory block codes with block length n and with rate at most R bits per letter. If we restrict the codes to have zero memory, then the corresponding nth-order performance is given by

$$\hat{D}_\theta^n(R) \triangleq \inf_{C^n}\{D_\theta(C^n) : R(C^n) \leq R\}.$$

Clearly, $\hat{D}_\theta^{n,*}(R) \leq \hat{D}_\theta^n(R)$. However, as far as optimal performance goes, allowing nonzero memory length does not help, as the following elementary lemma shows.

Lemma 2.1: $\hat{D}_\theta^{n,*}(R) = \hat{D}_\theta^n(R)$.

Proof: It suffices to show that $\hat{D}_\theta^n(R) \leq \hat{D}_\theta^{n,*}(R)$. Consider an arbitrary (n, m)-block code

$$C^{n,m} = \phi \circ f, \quad f : \mathcal{X}^n \times \mathcal{X}^m \to \mathcal{S}, \quad \phi : \mathcal{S} \to \hat{\mathcal{X}}^n.$$

We claim that there exists a zero-memory code $C_*^n = \phi_* \circ f_*$, $f_* : \mathcal{X}^n \to \mathcal{S}, \phi_* : \mathcal{S} \to \hat{\mathcal{X}}^n$, such that

$$\rho(x^n, C_*^n(x^n)) \leq \rho(x^n, C^{n,m}(x^n, z^m))$$

for all $x^n \in \mathcal{X}^n, z^m \in \mathcal{X}^m$. Indeed, define f_* as the minimum-distortion encoder

$$x^n \mapsto \arg\min_{s \in \mathcal{S}} \rho(x^n, \phi(s))$$

for the reproduction codebook of $C^{n,m}$, and let $\phi_*(s) = \phi(s)$. Then it is easy to see that $R(C_*^n) \leq R(C^{n,m})$ and $D_\theta(C_*^n) \leq D_\theta(C^{n,m})$ for all $\theta \in \Theta$, and the lemma is proved. $\square$

Armed with this lemma, we can compare the performance of all fixed-rate quantizers with block length n, with or without memory, to the nth-order operational DRF $\hat{D}_\theta^n(R)$. If we allow the block length to grow, then the best performance that can be achieved by a fixed-rate quantizer with or without memory on the source P_θ is given by the *operational distortion-rate function*

$$\hat{D}_\theta(R) \triangleq \inf_n \hat{D}_\theta^n(R) = \lim_{n \to \infty} \hat{D}_\theta^n(R).$$

Since an i.i.d. source is stationary and ergodic, the source coding theorem and its converse [22, Ch. 9] guarantee that the operational DRF $\hat{D}_\theta(R)$ is equal to the *Shannon DRF* $D_\theta(R)$, which in the i.i.d. case admits the following single-letter characterization:

$$D_\theta(R) \triangleq \inf_Q \{\mathbb{E}_{P_\theta Q}[\rho(X, \hat{X})] : I_{P_\theta Q}(X, \hat{X}) \leq R\}.$$

Here, the infimum is taken over all conditional probabilities (or test channels) Q from $\mathcal{X}$ to $\hat{\mathcal{X}}$, so that $P_\theta Q$ is the corresponding joint probability on $\mathcal{X} \times \hat{\mathcal{X}}$, and I is the mutual information.

A universal lossy coding scheme at rate R for the class $\{P_\theta : \theta \in \Theta\}$ is a sequence of codes $\{C^{n,m}\}$, where $n = 1, 2, \ldots$, and m is either a constant or a function of n, such that for each $\theta \in \Theta$, $R(C^{n,m})$ and $D_\theta(C^{n,m})$ converge to R and $D_\theta(R)$, respectively, as $n \to \infty$. Depending on the mode of convergence with respect to θ, one gets different types of universal codes.

Specifically, let $\{C^{n,m}\}_{n=1}^\infty$ be a sequence of lossy codes satisfying $R(C^{n,m}) \to R$ as $n \to \infty$. Then, following [9], we can distinguish between the following three types of universality.

Definition 2.1 (Weighted Universal): $\{C^{n,m}\}$ is *weighted universal* for $\{P_\theta : \theta \in \Theta\}$ with respect to a probability distribution W on Θ (on an appropriate σ-field) if the *distortion redundancy*

$$\delta_\theta(C^{n,m}) \triangleq D_\theta(C^{n,m}) - D_\theta(R)$$

converges to zero in the mean, i.e.,

$$\lim_{n \to \infty} \int_\Theta \delta_\theta(C^{n,m}) dW(\theta) = 0.$$

Definition 2.2 (Weakly Minimax Universal): $\{C^{n,m}\}$ is *weakly minimax universal* for $\{P_\theta : \theta \in \Theta\}$ if

$$\lim_{n \to \infty} \delta_\theta(C^{n,m}) = 0, \quad \forall \theta \in \Theta$$

i.e., $\delta_\theta(C^{n,m})$ converges to zero pointwise in θ.

Definition 2.3 (Strongly Minimax Universal): $\{C^{n,m}\}$ is *strongly minimax universal* for $\{P_\theta : \theta \in \Theta\}$ if $\delta_\theta(C^{n,m}) \to 0$ as $n \to \infty$ uniformly in θ.

The various relationships between the three types of universality have been explored in detail, e.g., in [9]. From the practical viewpoint, the differences between them are rather insubstantial. For instance, the existence of a weighted universal sequence of codes for $\{P_\theta : \theta \in \Theta\}$ with respect to W implies, for any $\epsilon > 0$, the existence of a strongly minimax universal sequence for $\{P_\theta : \theta \in \Theta_\epsilon\}$ for some $\Theta_\epsilon \subseteq \Theta$ satisfying $W(\Theta_\epsilon) \geq 1 - \epsilon$. In this paper, we shall concentrate exclusively on weakly minimax universal codes.

Once the existence of a universal sequence of codes is established in an appropriate sense, we can proceed to determine the rate of convergence. To facilitate this, we shall follow Chou, Effros, and Gray [13] and split the redundancy $\delta_\theta(C^{n,m})$ into two nonnegative terms

$$\delta_\theta(C^{n,m}) = \left(D_\theta(C^{n,m}) - \hat{D}_\theta^n(R)\right) + \left(\hat{D}_\theta^n(R) - D_\theta(R)\right). \tag{1}$$

The first term, which we shall call the *nth-order redundancy* and denote by $\delta_\theta^n(C^{n,m})$, quantifies the difference between the performance of $C^{n,m}$ and the nth-order operational DRF, while the second term tells us by how much the nth-order operational DRF exceeds the Shannon DRF, with respect to the source P_θ. Note that $\delta_\theta(C^{n,m})$ converges to zero if and only if $\delta_\theta^n(C^{n,m})$ does, because $\hat{D}_\theta^n(R) \to D_\theta(R)$ as $n \to \infty$ by the source coding theorem. Thus, in proving the existence of universal codes, we shall determine the rates at which the two terms on the right-hand side of (1) converge to zero as $n \to \infty$.

III. TWO-STAGE JOINT UNIVERSAL CODING AND MODELING

As discussed in the Introduction, two-stage codes are both practically and conceptually appealing for analysis and design of universal codes. A two-stage lossy block code (vector quantizer) with block length n is a code that describes each source

sequence x^n in two stages: in the first stage, a quantizer of block length n is chosen as a function of x^n from some collection of available quantizers; this is followed by the second stage, in which x^n is encoded with the chosen code.

In precise terms, a two-stage fixed-rate lossy code is defined as follows [13]. Let $\tilde{f} : \mathcal{X}^n \to \tilde{\mathcal{S}}$ be a mapping of $\mathcal{X}^n$ into a collection $\tilde{\mathcal{S}}$ of fixed-length binary strings, and assume that to each $\tilde{s} \in \tilde{\mathcal{S}}$ there corresponds an n-block code $C_{\tilde{s}}^n = (f_{\tilde{s}}, \phi_{\tilde{s}})$ at rate of R bits per letter. A two-stage code C^n is defined by the encoder

$$f(x^n) = \tilde{f}(x^n) f_{\tilde{f}(x^n)}(x^n)$$

and the decoder

$$\phi(\tilde{f}(x^n) f_{\tilde{f}(x^n)}(x^n)) = \phi_{\tilde{f}(x^n)}(x^n).$$

Here, the juxtaposition of two binary strings stands for their concatenation. The map $\tilde{f}$ is called the first-stage encoder. The rate of this code is $R + n^{-1} \log |\tilde{\mathcal{S}}|$ bits per letter, while the instantaneous distortion is

$$\rho(x^n, C^n(x^n)) = \rho\left(x^n, C_{\tilde{f}(x^n)}^n(x^n)\right).$$

Now consider using C^n to code an i.i.d. process $\{X_i\}$ with all the X_i's distributed according to P_θ for some $\theta \in \Theta$. This will result in the average per-letter distortion

$$\begin{aligned}
D_\theta(C^n) &= \frac{1}{n} \mathbb{E}_\theta[\rho(X^n, C^n(X^n))] \\
&= \frac{1}{n} \mathbb{E}_\theta\left[\rho\left(X^n, C_{\tilde{f}(X^n)}^n(X^n)\right)\right] \\
&= \frac{1}{n} \int \rho\left(x^n, C_{\tilde{f}(x^n)}^n(x^n)\right) dP_\theta(x^n).
\end{aligned}$$

Note that it is not possible to express $D_\theta(C^n)$ in terms of expected distortion of any single code because the identity of the code used to encode each $x^n \in \mathcal{X}^n$ itself varies with x^n.

Let us consider the following modification of two-stage coding. As before, we wish to code an i.i.d. source $\{X_i\}$ with an n-block lossy code, but this time we allow the code to have finite memory m. Assume once again that we have an indexed collection $\{C_{\tilde{s}}^n : \tilde{s} \in \tilde{\mathcal{S}}\}$ of n-block codes, but this time the first-stage encoder is a map $\tilde{f} : \mathcal{X}^m \to \tilde{\mathcal{S}}$ from the space $\mathcal{X}^m$ of m-blocks over $\mathcal{X}$ into $\tilde{\mathcal{S}}$. In order to encode the current n-block $X^n(t), t \in \mathbb{Z}$, the encoder first looks at $X_m^n(t-1)$, the m-block immediately preceding $X^n(t)$, selects a code $C_{\tilde{s}}^n$ according to the rule $\tilde{s} = \tilde{f}(X_m^n(t-1))$, and then codes $X^n(t)$ with that code. In this way, we have a two-stage (n, m)-block code $C^{n,m}$ with the encoder

$$f(x^n, z^m) = \tilde{f}(z^m) f_{\tilde{f}(z^m)}(x^n)$$

and the decoder

$$\phi\left(\tilde{f}(z^m) f_{\tilde{f}(z^m)}(x^n)\right) = \phi_{\tilde{f}(z^m)}(x^n).$$

The operation of this code can be pictured as a blockwise application of a nonlinear time-invariant filter with the initial conditions determined by a fixed finite amount of past data. Just as in

the memoryless case, the rate of $C^{n,m}$ is $R + n^{-1} \log |\tilde{\mathcal{S}}|$ bits per letter, but the instantaneous distortion is now given by

$$\rho(x^n, C^{n,m}(x^n, z^m)) = \rho\left(x^n, C_{\tilde{f}(z^m)}^n(x^n)\right).$$

When the common distribution of the X_i's is P_θ, the average per-letter distortion is given by

$$\begin{aligned}
D_\theta(C^{n,m}) &= \frac{1}{n} \int_{\mathcal{X}^n \times \mathcal{X}^m} \rho\left(x^n, C_{\tilde{f}(z^m)}^n(x^n)\right) dP_\theta(x^n, z^m) \\
&= \frac{1}{n} \mathbb{E}_\theta\left\{\mathbb{E}_\theta\left[\rho\left(X^n, C_{\tilde{f}(Z^m)}^n(X^n)\right) \Big| Z^m\right]\right\} \\
&= \mathbb{E}_\theta\left[D_\theta\left(C_{\tilde{f}(X^m)}^n\right)\right]. \tag{2}
\end{aligned}$$

Observe that the use of memory allows us to decouple the choice of the code from the actual encoding operation, which in turn leads to an expression for the average distortion of $C^{n,m}$ that involves iterated expectations.

Intuitively, this scheme will yield a universal code if

$$\mathbb{E}_\theta\left[D_\theta\left(C_{\tilde{f}(X^m)}^n\right)\right] \approx \hat{D}_\theta^n(R) \tag{3}$$

for each $\theta \in \Theta$. Keeping in mind that $\tilde{f}(X^m)$ is allowed to take only a finite number $|\tilde{\mathcal{S}}|$ of values, we see that condition (3) must be achieved through some combination of parameter estimation and quantization. To this end, we impose additional structure on the map $\tilde{f}$. Namely, we assume that it is composed of a *parameter estimator* $\tilde{\theta} : \mathcal{X}^m \to \Theta$ that uses the past data $X_m^n(t-1)$ to estimate the parameter label $\theta \in \Theta$ of the source in effect, and a lossy *parameter encoder* $\tilde{g} : \Theta \to \tilde{\mathcal{S}}$, whereby the estimate $\tilde{\theta}$ is quantized to $\tilde{R} \equiv \log |\tilde{\mathcal{S}}|$ bits, with respect to a suitable distortion measure on Θ. A binary description of the quantized version $\hat{\theta}$ of $\tilde{\theta}(X_m^n(t-1))$ is then passed on to the second-stage encoder which will quantize the current n-block $X^n(t)$ with an n-block code matched to $P_{\hat{\theta}}$. Provided that P_θ and $P_{\hat{\theta}}$ are "close" to each other in an appropriate sense, the resulting performance will be almost as good as if the actual parameter θ were known all along. As a bonus, the decoder will also receive a good $\tilde{R}$-bit binary representation (model) of the source in effect. Therefore, we shall also define a *parameter decoder* $\tilde{\psi} : \tilde{\mathcal{S}} \to \Theta$, so that $\hat{\theta} = \tilde{\psi}(\tilde{f}(X_m^n(t - 1))$ can be taken as an estimate of the parameter $\theta \in \Theta$ of the active source. The structure of the encoder and the decoder in this two-stage scheme for joint modeling and lossy coding is displayed in Fig. 1.

These ideas are formalized in Theorem 3.2 below for i.i.d. vector sources $\{X_i\}, X_i \in \mathbb{R}^d$, where the common distribution of the X_i's is a member of a given indexed class $\{P_\theta : \theta \in \Theta\}$ of absolutely continuous distributions, and the parameter space Θ is a bounded subset of $\mathbb{R}^k$. For simplicity, we have set $m = n$, although other choices for the memory length are also possible. Before we state and prove the theorem, let us fix some useful results and notation. The following proposition generalizes Theorem 2 of Linder, Lugosi, and Zeger [11] to i.i.d. vector sources and characterizes the rate at which the nth-order operational DRF converges to the Shannon DRF (the proof, which uses Csiszár's generalized parametric representation of the DRF [23], as well as a combination of standard random coding arguments and large-deviation estimates, is an almost verbatim

Second-stage encoder

$X^n(t)$

Parameter estimator

$\widetilde{\theta}(t+1)$

Parameter encoder

$\widetilde{s}(t+1)$

Unit delay

First-stage encoder

$C^n_{11\ldots1}$ → $s(t)$

$C^n_{00\ldots1}$

$C^n_{00\ldots0}$

→ $\widetilde{s}(t)$

$s(t)$

$\widetilde{s}(t)$

Parameter decoder

$C^n_{11\ldots1}$ → $\hat{X}^n(t)$

$C^n_{00\ldots1}$

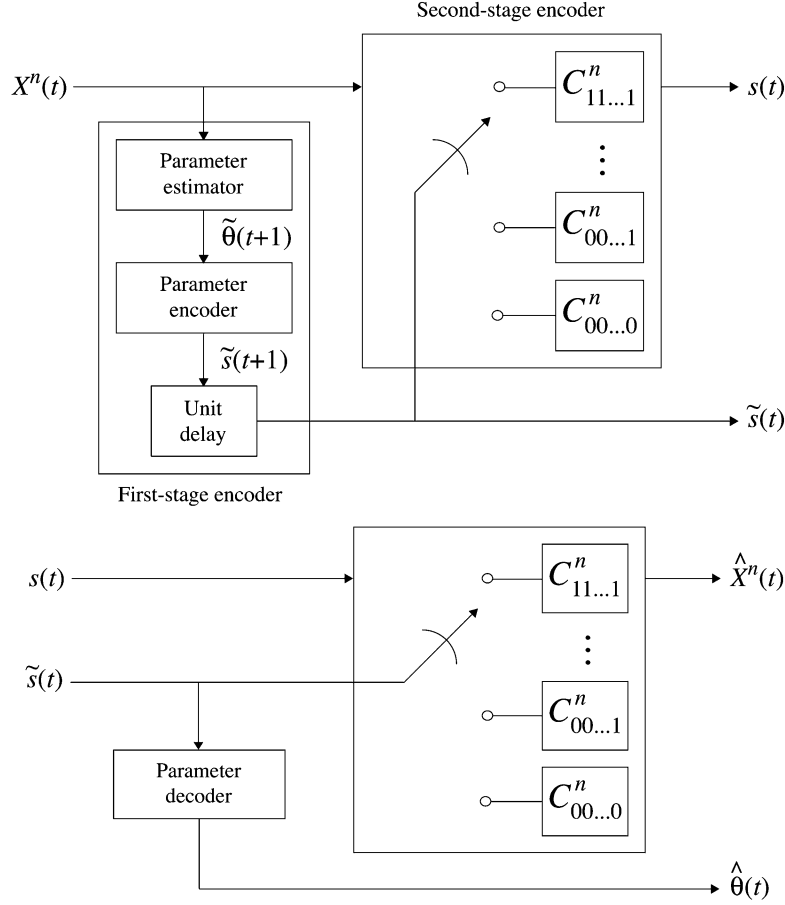$C^n_{00\ldots0}$

→ $\hat{\theta}(t)$

Fig. 1.   Two-stage scheme for joint modeling and lossy coding (top: encoder; bottom: decoder).

adaptation of the proof of Linder *et al.* to vector sources, and is presented for completeness in Appendix B).

Proposition 3.1: Let $\{X_i\}$ be an i.i.d. source with alphabet $\mathcal{X} \subseteq \mathbb{R}^d$, where the common distribution of the X_i's comes from an indexed class $\{P_\theta : \theta \in \Theta\}$. Let $\rho : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ be a distortion function satisfying the following two conditions:
1) $\inf_{\hat{x} \in \hat{\mathcal{X}}} \rho(x, \hat{x}) = 0$ for all $x \in \mathcal{X}$;
2) $\sup_{x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}} \rho(x, \hat{x}) = \rho_{\max} < \infty$.
Then for every $\theta \in \Theta$ and every $R > 0$ such that $D_\theta(R) > 0$ there exists a constant $c_\theta(R)$ such that

$$\hat{D}^n_\theta(R) - D_\theta(R) \leq (c_\theta(R) + o(1))\sqrt{\frac{\log n}{n}}.$$

The function $c_\theta(R)$ is continuous in R, and the $o(1)$ term converges to zero uniformly in a sufficiently small neighborhood of R.

Remark 3.1: The condition $D_\theta(R) > 0$ is essential to the proof and holds for all $R > 0$ whenever P_θ has a continuous component, which is assumed in the following.

Remark 3.2: The constant $c_\theta(R)$ depends on the derivative of the DRF $D_\theta(R)$ at R and on the maximum value $\rho_{\max}$ of the distortion function.

The distance between two i.i.d. sources will be measured by the *variational distance* between their respective single-letter distributions [19, Ch. 5]

$$d_V(P_\theta, P_\eta) \overset{\triangle}{=} \sup_B |P_\theta(B) - P_\eta(B)|, \quad \forall \theta, \eta \in \Theta$$

where the supremum is taken over all Borel subsets of $\mathbb{R}^d$. Also, given a sequence of real-valued random variables $V_1, V_2, \ldots,$ and a sequence of nonnegative numbers $a_1, a_2, \ldots,$ the notation $V_n = O(a_n)$ almost surely (a.s.) means that there exists a constant $c > 0$ and a nonnegative random variable $N \in \mathbb{Z}$ such that $V_n \leq c a_n$ for all $n \geq N$. Finally, both the statement and the proof of the theorem rely on certain notions from VC theory; for the reader's convenience, Appendix A contains a summary of the necessary definitions and results.

Theorem 3.2: Let $\{X_i\}_{-\infty}^{\infty}$ be an i.i.d. source with alphabet $\mathcal{X} \subseteq \mathbb{R}^d$, where the common distribution of the X_i's is a member of a class $\{P_\theta : \theta \in \Theta\}$ of absolutely continuous distributions with the corresponding densities p_θ. Assume the following conditions are satisfied.
1) Θ is a bounded subset of $\mathbb{R}^k$.
2) The map $\theta \mapsto P_\theta$ is uniformly locally Lipschitz: there exist constants $r > 0$ and $m > 0$ such that, for each $\theta \in \Theta$

$$d_V(P_\theta, P_\eta) \leq m\|\theta - \eta\|$$

for all $\eta \in B_r(\theta)$, where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^k$ and $B_r(\theta)$ is an open ball of radius r centered at θ.
3) The *Yatracos class* [17], [18], [24] associated with Θ, defined as

$$\mathcal{A}_\Theta \overset{\triangle}{=} \{A_{\theta,\eta} = \{x \in \mathcal{X} : p_\theta(x)$$
$$> p_\eta(x)\} : \theta, \eta \in \Theta; \theta \neq \eta\},$$

is a VC class, $\mathsf{V}(\mathcal{A}_\Theta) = V < \infty$.

Let $\rho : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ be a single-letter distortion function of the form $\rho(x, \hat{x}) = [d(x, \hat{x})]^p$ for some $p > 0$, where $d(\cdot, \cdot)$ is a bounded metric on $\mathcal{X} \cup \hat{\mathcal{X}}$ with the bound $d_{\max}$. Suppose that for each n and each $\theta \in \Theta$ there exists an n-block code $C_\theta^n = (f_\theta, \phi_\theta)$ at rate of $R > 0$ bits per letter that achieves the nth-order operational DRF for P_θ: $D_\theta(C_\theta^n) = \hat{D}_\theta^n(R)$. Then there exists an (n, n)-block code $C^{n,n}$ with

$$R(C^{n,n}) = R + O\left(\frac{\log n}{n}\right) \tag{4}$$

such that for every $\theta \in \Theta$

$$\delta_\theta(C^{n,n}) = O\left(\sqrt{\frac{\log n}{n}}\right). \tag{5}$$

The resulting sequence of codes $\{C^{n,n}\}_{n=1}^\infty$ is therefore weakly minimax universal for $\{P_\theta : \theta \in \Theta\}$ at rate R. Furthermore, for each n the first-stage encoder $\tilde{f}$ and the corresponding parameter decoder $\tilde{\psi}$ are such that

$$d_V\left(P_\theta, P_{\tilde{\psi}(\tilde{f}(X^n))}\right) = O\left(\sqrt{\frac{\log n}{n}}\right) \quad \text{a.s.} \tag{6}$$

where the probability is with respect to P_θ. The constants implicit in the $O(\cdot)$ notation in (4) and (6) are independent of θ.

Proof: The theorem will be proved by construction of a two-stage code, where the first-stage encoder $\tilde{f} : \mathcal{X}^n \to \tilde{\mathcal{S}}$ is a cascade of the parameter estimator $\tilde{\theta} : \mathcal{X}^n \to \Theta$ and the lossy parameter encoder $\tilde{g} : \Theta \to \tilde{\mathcal{S}}$. Estimation of the parameter vector θ at the decoder will be facilitated by the corresponding decoder $\tilde{\psi} : \tilde{\mathcal{S}} \to \Theta$.

Our parameter estimator will be based on the so-called *minimum-distance density estimator* [16, Sec. 5.5], originally developed by Devroye and Lugosi [17], [18] in the context of kernel density estimation. It is constructed as follows. Let $Z^n = (Z_1, \ldots, Z_n)$ be i.i.d. according to P_θ for some $\theta \in \Theta$. Given any $\eta \in \Theta$ let

$$\Delta_\eta(Z^n) \triangleq \sup_{A \in \mathcal{A}_\Theta} |P_\eta(A) - P_{Z^n}(A)|$$

where P_{Z^n} is the empirical distribution of Z^n:

$$P_{Z^n}(B) \triangleq \frac{1}{n} \sum_{i=1}^n 1_{\{Z_i \in B\}}$$

for any Borel set B. Define $\tilde{\theta}(Z^n)$ as any $\theta^* \in \Theta$ satisfying

$$\Delta_{\theta^*}(Z^n) < \inf_{\eta \in \Theta} \Delta_\eta(Z^n) + \frac{1}{n}$$

where the extra $1/n$ term has been added to ensure that at least one such θ^* exists. Then $P_{\tilde{\theta}(Z^n)}$ is called the minimum-distance estimate of P_θ. Through an abuse of terminology, we shall also say that $\tilde{\theta}$ is the minimum-distance estimate of θ. The key property of the minimum-distance estimate [16, Theorem 5.13] is that

$$\int_{\mathcal{X}} \left| p_{\tilde{\theta}(Z^n)}(x) - p_\theta(x) \right| dx \le 4\Delta_\theta(Z^n) + \frac{3}{n}. \tag{7}$$

Since the variational distance between any two absolutely continuous distributions P, Q on $\mathbb{R}^d$ is equal to one half of the L_1 distance between their respective densities p, q [19, Theorem 5.1], i.e.,

$$d_V(P, Q) = \frac{1}{2} \int_{\mathbb{R}^d} |p(x) - q(x)| dx$$

we can rewrite (7) as

$$d_V\left(P_\theta, P_{\tilde{\theta}(Z^n)}\right) \le 2\Delta_\theta(Z^n) + \frac{3}{2n}. \tag{8}$$

Since $\mathcal{A}_\Theta$ is a VC class, Lemma A.2 in the Appendix asserts that

$$\mathbb{E}_\theta[\Delta_\theta(Z^n)] \le c_1 \sqrt{\frac{\log n}{n}} \tag{9}$$

where c_1 is a constant that depends only on the VC dimension of $\mathcal{A}_\Theta$. Taking expectations of both sides of (8) and applying (9), we get

$$\mathbb{E}_\theta\left[d_V(P_\theta, P_{\tilde{\theta}(Z^n)})\right] \le 2c_1 \sqrt{\frac{\log n}{n}} + \frac{3}{2n}. \tag{10}$$

Next, we construct the lossy encoder $\tilde{g}$. Since Θ is bounded, it is contained in some hypercube M of side J, where J is some positive integer. Let $\mathcal{M}^{(n)} = \{M_1^{(n)}, M_2^{(n)}, \ldots, M_K^{(n)}\}$ be a partitioning of M into nonoverlapping hypercubes of side $1/\lceil n^{1/2} \rceil$, so that $K \le (Jn^{1/2})^k$. Represent each $M_j^{(n)}$ that intersects Θ by a unique fixed-length binary string $\tilde{s}_j$, and let $\tilde{\mathcal{S}} = \{\tilde{s}_j\}$. Then if a given $\theta \in \Theta$ is contained in $M_j^{(n)}$, map it to $\tilde{s}_j, \tilde{g}(\theta) = \tilde{s}_j$; this choice can be described by a string of no more than $k(\log n^{1/2} + \log J)$ bits. Finally, for each $M_j^{(n)}$ that intersects Θ, choose a representative $\hat{\theta}_j \in M_j^{(n)} \cap \Theta$ and define the corresponding n-block code $C_{\tilde{s}_j}^n$ to be $C_{\hat{\theta}_j}^n$. Thus, we can associate to $\tilde{g}$ the decoder $\tilde{\psi} : \tilde{\mathcal{S}} \to \Theta$ via $\tilde{\psi}(\tilde{s}_j) = \hat{\theta}_j$.

Now let us describe and analyze the operation of the resulting two-stage (n, n)-block code $C^{n,n}$. In order to keep the notation simple, we shall suppress the discrete time variable t and denote the current block $X^n(t)$ by X^n, while the preceding block $X^n(t-1)$ will be denoted by Z^n. The first-stage encoder $\tilde{f}$ computes the minimum-distance estimate $\tilde{\theta} = \tilde{\theta}(Z^n)$ and communicates its lossy binary description $\tilde{s} = \tilde{f}(Z^n) \equiv \tilde{g}(\tilde{\theta}(Z^n))$ to the second-stage encoder. The second-stage encoder then encodes X^n with the n-block code $C_{\tilde{s}}^n \equiv C_{\hat{\theta}}^n$, where $\hat{\theta} = \tilde{\psi}(\tilde{s})$ is the quantized version of the minimum-distance estimate $\tilde{\theta}$. The string transmitted to the decoder thus consists of two parts: the header $\tilde{s}$, which specifies the second-stage code $C_{\tilde{s}}^n$, and the body $s = f_{\tilde{s}}(X^n)$, which is the encoding of X^n under $C_{\tilde{s}}^n$. The decoder computes the reproduction $\hat{X}^n = \phi_{\tilde{s}}(s)$. Note, however, that the header $\tilde{s}$ not only instructs the decoder how to decode the body s, but also contains a binary description of the quantized minimum-distance estimate of the active source, which can be recovered by means of the rule $\hat{\theta} = \tilde{\psi}(\tilde{s})$.

In order to keep the notation simple, assume for now that $p = 1$, i.e., the distortion function ρ is a metric on $\mathcal{X} \cup \hat{\mathcal{X}}$ with the bound $\rho_{\max} \equiv d_{\max}$; the case of general p is similar. The rate of $C^{n,n}$ is clearly no more than

$$R + \frac{k(\log n^{1/2} + \log J)}{n}$$

bits per letter, which proves (4). The average per-letter distortion of $C^{n,n}$ on the source P_θ is, in accordance with (2), given by

$$D_\theta(C^{n,n}) = \mathbb{E}_\theta\left[D_\theta\left(C_{\tilde{f}(Z^n)}^n\right)\right]$$

where $C^n_{\tilde{f}(Z^n)} = C^n_{\hat{\theta}}$ with $\hat{\theta} = \tilde{\psi}(\tilde{f}(Z^n))$. Without loss of generality, we may assume that each C^n_θ is a nearest-neighbor quantizer, i.e.,

$$\rho\left(x^n, C^n_\theta(x^n)\right) = \min_{\hat{x}^n \in \Gamma_\theta} \rho(x^n, \hat{x}^n)$$

for all $x^n \in \mathcal{X}^n$, where Γ_θ is the codebook of C^n_θ. Then we have the following chain of estimates:

$$\begin{aligned} D_\theta(C^n_{\hat{\theta}}) &\overset{(a)}{\leq} D_{\hat{\theta}}\left(C^n_{\hat{\theta}}\right) + 2\rho_{\max} d_V\left(P_\theta, P_{\hat{\theta}}\right) \\ &\overset{(b)}{=} \hat{D}^n_{\hat{\theta}}(R) + 2\rho_{\max} d_V(P_\theta, P_{\hat{\theta}}) \\ &\overset{(c)}{\leq} \hat{D}^n_\theta(R) + 4\rho_{\max} d_V(P_\theta, P_{\hat{\theta}}) \\ &\overset{(d)}{\leq} \hat{D}^n_\theta(R) + 4\rho_{\max}\left[d_V(P_\theta, P_{\tilde{\theta}}) + d_V(P_{\tilde{\theta}}, P_{\hat{\theta}})\right] \end{aligned}$$

where (a) and (c) follow from a basic quantizer mismatch estimate (Lemma C.1 in the Appendix), (b) follows from the assumed nth-order optimality of $C^n_{\hat{\theta}}$ for $P_{\hat{\theta}}$, while (d) is a routine application of the triangle inequality. Taking expectations, we get

$$\begin{aligned} D_\theta(C^{n,n}) \leq \hat{D}^n_\theta(R) + 4\rho_{\max}\{&\mathbb{E}_\theta[d_V(P_\theta, P_{\tilde{\theta}})] \\ &+ \mathbb{E}_\theta[d_V(P_{\tilde{\theta}}, P_{\hat{\theta}})]\}. \end{aligned} \quad (11)$$

We now estimate separately each term in the curly brackets in (11). The first term can be bounded using the fact that $\tilde{\theta} = \tilde{\theta}(Z^n)$ is a minimum-distance estimate of θ, so by (10) we have

$$\mathbb{E}_\theta[d_V(P_\theta, P_{\tilde{\theta}})] \leq 2c_1\sqrt{\frac{\log n}{n}} + \frac{3}{2n}. \quad (12)$$

The second term involves $\tilde{\theta}$ and its quantized version $\hat{\theta}$, which satisfy $\|\tilde{\theta} - \hat{\theta}\| \leq \sqrt{k/n}$, by construction of the parameter space quantizer $(\tilde{g}, \tilde{\psi})$. By the uniform local Lipschitz property of the map $\theta \mapsto P_\theta$, there exist constants $r > 0$ and $m > 0$, such that

$$d_V(P_\eta, P_{\hat{\theta}}) \leq m\|\eta - \hat{\theta}\|$$

for all $\eta \in B_r(\hat{\theta})$. If $\tilde{\theta} \in B_r(\hat{\theta})$, this implies that $d_V(P_{\tilde{\theta}}, P_{\hat{\theta}}) \leq m\sqrt{k/n}$. Suppose, on the other hand, that $\tilde{\theta} \notin B_r(\hat{\theta})$. By assumption, $\|\tilde{\theta} - \hat{\theta}\| \geq r$. Therefore, since $d_V(\cdot, \cdot)$ is bounded from above by unity, we can write

$$d_V(P_{\tilde{\theta}}, P_{\hat{\theta}}) \leq \frac{1}{r}\|\tilde{\theta} - \hat{\theta}\| \leq \frac{1}{r}\sqrt{\frac{k}{n}}.$$

Let $b \overset{\triangle}{=} \max(m, 1/r)$. Then the above argument implies that

$$d_V(P_{\tilde{\theta}}, P_{\hat{\theta}}) \leq b\sqrt{\frac{k}{n}} \quad (13)$$

and consequently

$$\mathbb{E}_\theta[d_V(P_{\tilde{\theta}}, P_{\hat{\theta}})] \leq b\sqrt{\frac{k}{n}}. \quad (14)$$

Substituting the bounds (12) and (14) into (11) yields

$$D_\theta(C^{n,n}) \leq \hat{D}^n_\theta(R) + \rho_{\max}\left(8c_1\sqrt{\frac{\log n}{n}} + \frac{6}{n} + 4b\sqrt{\frac{k}{n}}\right)$$

whence it follows that the nth-order redundancy $\delta^n_\theta(C^n) = O(\sqrt{n^{-1}\log n})$ for every $\theta \in \Theta$. Then the decomposition

$$\delta_\theta(C^{n,n}) = \delta^n_\theta(C^{n,n}) + \hat{D}^n_\theta(R) - D_\theta(R)$$

and Proposition 3.1 imply that (5) holds for every $\theta \in \Theta$. The case of $p \neq 1$ is similar.

To prove (6), fix an $\epsilon > 0$ and note that by (8), (13), and the triangle inequality, $d_V(P_\theta, P_{\hat{\theta}(Z^n)}) > \epsilon$ implies that

$$2\Delta_\theta(Z^n) + \frac{3}{2n} + b\sqrt{\frac{k}{n}} > \epsilon.$$

Hence

$$\begin{aligned} &\mathbb{P}\left\{d_V\left(P_\theta, P_{\hat{\theta}(Z^n)}\right) > \epsilon\right\} \\ &\leq \mathbb{P}\left\{\Delta_\theta(Z^n) > \frac{1}{2}\left(\epsilon - \frac{3}{2n} - b\sqrt{\frac{k}{n}}\right)\right\} \\ &\leq \mathbb{P}\left\{\Delta_\theta(Z^n) > \frac{1}{2}\left(\epsilon - c_2\sqrt{\frac{1}{n}}\right)\right\} \end{aligned}$$

where $c_2 = 3/2 + b\sqrt{k}$. Therefore, by Lemma A.2

$$\mathbb{P}\left\{d_V\left(P_\theta, P_{\hat{\theta}(Z^n)}\right) > \epsilon\right\} \leq 8n^V e^{-n(\epsilon - c_2\sqrt{1/n})^2/128}. \quad (15)$$

If for each n we choose $\epsilon_n > \sqrt{128V\ln n/n} + c_2\sqrt{1/n}$, then the right-hand side of (15) will be summable in n, hence $d_V(P_\theta, P_{\hat{\theta}(Z^n)}) = O(\sqrt{n^{-1}\log n})$ a.s. by the Borel–Cantelli lemma. $\square$

Remark 3.3: Our proof combines the techniques of Rissanen [1], in that the second-stage code is selected through explicit estimation of the source parameters, and of Chou, Effros, and Gray [13], in that the parameter space is quantized and each θ is identified with its optimal code C^n_θ. The novel element here is the use of minimum-distance estimation instead of maximum-likelihood estimation, which is responsible for the appearance of the VC dimension.

Remark 3.4: The boundedness of the distortion measure has been assumed mostly in order to ensure that the main idea behind the proof is not obscured by technical details. In Section IV-D, we present an extension to distortion measures that satisfy a moment condition with respect to a fixed reference letter in the reproduction alphabet. In that case, the parameter estimation fidelity and the per-letter overhead rate still converge to zero as $O(\sqrt{n^{-1}\log n})$ and $O(n^{-1}\log n)$, respectively, but the distortion redundancy converges more slowly as $O(\sqrt[4]{n^{-1}\log n})$.

Remark 3.5: Essentially the same convergence rates, up to multiplicative and/or additive constants, can be obtained if the

memory length is taken to be some fraction of the block length n: $m = \alpha n$ for some $\alpha \in (0, 1)$.

Remark 3.6: Let us compare the local Lipschitz condition of Theorem 3.2 to the corresponding smoothness conditions of Rissanen [1] for lossless codes and of Chou *et al.* [13] for quantizers. In the lossless case, $\mathcal{X}$ is finite or countably infinite, and the smoothness condition is for the relative entropies $D(P_\eta \| P_\theta) \triangleq \sum_{x \in \mathcal{X}} p_\eta(x) \log(p_\eta(x)/p_\theta(x))$, where p_θ and p_η are the corresponding probability mass functions, to be locally quadratic in θ: $D(P_\eta \| P_\theta) \le m_\theta \|\theta - \eta\|^2$ for some constant m_θ and for all η in some open neighborhood of θ. Pinsker's inequality $D(P_\eta \| P_\theta) \ge d_V^2(P_\eta, P_\theta)/2 \ln 2$ [25, p. 58] then implies the local Lipschitz property for $d_V(P_\theta, P_\eta)$, although the magnitude of the Lipschitz constant is not uniform in θ. Now, $D(P_\eta \| P_\theta)$ is also the redundancy of the optimum lossless code for P_θ relative to P_η. Thus, Rissanen's smoothness condition can be interpreted either in the context of source models or in the context of coding schemes and their redundancies. The latter interpretation has been extended to quantizers in [13], where it was required that the redundancies $\delta_\eta^n(C_\theta^n)$ be locally quadratic in θ. However, because here we are interested in joint modeling and coding, we impose a smoothness condition on the source distributions, rather than on the codes. The variational distance is more appropriate here than the relative entropy because, for bounded distortion functions, it is a natural measure of redundancy for lossy codes [9].

Remark 3.7: The VC dimension of a given class of measurable subsets of $\mathbb{R}^k$ (provided it is finite) is, in a sense, a logarithmic measure of the combinatorial "richness" of the class for the purposes of learning from empirical data. For many parametric families of probability densities, the VC dimension of the corresponding Yatracos class is polynomial in k, the dimension of the parameter space (see [16] for detailed examples).

Remark 3.8: Instead of the VC condition, we could have required that the class of sources $\{P_\theta : \theta \in \Theta\}$ be *totally bounded* with respect to the variational distance. (Totally bounded classes, with respect to either the variational distance or its generalizations, such as the $\bar{\rho}$-distance [26], have, in fact, been extensively used in the theory of universal lossy codes [9].) This was precisely the assumption made in the paper of Yatracos [24] on density estimation, which in turn inspired the work of Devroye and Lugosi [17], [18]. The main result of Yatracos is that, if the class $\{P_\theta : \theta \in \Theta\}$ is totally bounded under the variational distance, then for any $\epsilon > 0$ there exists an estimator $\theta^* = \theta^*(X^n)$, where X^n is an i.i.d. sample from one of the P_θ's, such that

$$\mathbb{E}_\theta \left[d_V \left(P_\theta, P_{\theta^*(X^n)} \right) \right] \le 3\epsilon + \sqrt{\frac{32 H_\epsilon + 8}{n}}$$

where H_ϵ is the metric entropy, or Kolmogorov ϵ-entropy [27], of $\{P_\theta : \theta \in \Theta\}$, i.e., the logarithm of the cardinality of the minimal ϵ-net for $\{P_\theta\}$ under $d_V(\cdot, \cdot)$. Thus, if we choose $\epsilon = \epsilon_n$ such that $\sqrt{H_{\epsilon_n}/n} \to 0$ as $n \to \infty$, then $\theta^*(X^n)$ is a consistent estimator of θ. However, totally bounded classes have certain drawbacks. For example, depending on the structure and the complexity of the class, the Kolmogorov ϵ-entropy may vary

rather drastically from a polynomial in $\log(1/\epsilon)$ for "small" parametric families (e.g., finite mixture families) to a polynomial in $1/\epsilon$ for nonparametric families (e.g., monotone densities on the hypercube or smoothness classes such as Sobolev spaces). One can even construct extreme examples of nonparametric families with H_ϵ *exponential* in $1/\epsilon$. (For details, the reader is invited to consult [19, Ch. 7].) Thus, in sharp contrast to VC classes for which we can obtain $O(\sqrt{n^{-1} \log n})$ convergence rates both for parameter estimates and for distortion redundancies, the performance of joint universal coding and modeling schemes for totally bounded classes of sources will depend rather strongly on the metric properties of the class. Additionally, although in the totally bounded case there is no need for quantizing the parameter space, one has to construct an ϵ-net for each given class, which is often an intractable problem.

IV. COMPARISONS AND EXTENSIONS

A. Comparison With Nearest-Neighbor and Omniscient First-Stage Encoders

The two-stage universal quantizer of Chou, Effros, and Gray [13] has zero memory and works as follows. Given a collection $\{C_{\tilde{s}}^n\}$ of n-block codes, the first-stage encoder is given by the "nearest-neighbor" map

$$\tilde{f}_*(x^n) \triangleq \arg\min_{\tilde{s} \in \tilde{\mathcal{S}}} \rho(x^n, C_{\tilde{s}}^n(x^n))$$

where the term "nearest-neighbor" is used in the sense that the code $C_{\tilde{f}_*(x^n)}^n$ encodes x^n with the smallest instantaneous distortion among all $C_{\tilde{s}}^n$'s. Accordingly, the average per-letter distortion of the resulting two-stage code C_*^n on the source P_θ is given by

$$D_\theta(C_*^n) = \frac{1}{n} \int \min_{\tilde{s} \in \tilde{\mathcal{S}}} \rho(x^n, C_{\tilde{s}}^n(x^n)) \, dP_\theta(x^n).$$

Although such a code is easily implemented in practice, its theoretical analysis is quite complicated. However, the performance of C_*^n can be upper-bounded if the nearest-neighbor first-stage encoder is replaced by the so-called *omniscient* first-stage encoder, which has direct access to the source parameter $\theta \in \Theta$, rather than to x^n. This latter encoder is obviously not achievable in practice, but is easily seen to do no better than the nearest-neighbor one.

This approach can be straightforwardly adapted to the setting of our Theorem 3.2, except that we no longer require Condition 3. In that case, it is apparent that the sequence $\{C_*^n\}$ of the two-stage n-block (zero-memory) codes with nearest-neighbor (or omniscient) first-stage encoders is such that

$$R(C_*^n) = R + O\left(\frac{\log n}{n}\right) \tag{16}$$

and

$$\delta_\theta(C_*^n) = O\left(\sqrt{\frac{\log n}{n}}\right). \tag{17}$$

Comparing (16) and (17) with (4) and (5), we immediately see that the use of memory and direct parameter estimation has no

effect on rate or on distortion. However, our scheme uses the $O(\log n)$ overhead bits in a more efficient manner—indeed, the bits produced by the nearest-neighbor first-stage encoder merely tell the second-stage encoder and the decoder which quantizer to use, but there is, in general, no guarantee that the nearest-neighbor code for a given $x^n \in \mathcal{X}^n$ will be matched to the actual source in an average sense. By contrast, the first-stage description under our scheme, while requiring essentially the same number of extra bits, can be used to identify the active source up to a variational ball of radius $O(\sqrt{n^{-1}\log n})$, with probability arbitrarily close to one.

B. Comparison With Schemes Based on Codebook Transmission

Another two-stage scheme, due to Linder, Lugosi, and Zeger [10], [11], yields weakly minimax universal codes for all real i.i.d. sources with bounded support, with respect to the squared-error distortion. The main feature of their approach is that, instead of constraining the first-stage encoder to choose from a collection of preselected codes, they encode each n-block $x^n \in \mathcal{X}^n$ by designing, in real time, an optimal quantizer for the empirical distribution P_{x^n}, whose codevectors are then quantized to some carefully chosen resolution. Then, in the second stage, x^n is quantized with this "quantized quantizer," and a binary description of the quantized codevectors is transmitted together with the second-stage description of x^n. The overhead needed to transmit the quantized codewords is $O(n^{-1}\log n)$ bits per letter, while the distortion redundancy converges to zero at a rate $O(\sqrt{n^{-1}\log n})$.

In order to draw a comparison with the results presented here, let $\{P_\theta : \theta \in \Theta\}$ be a class of real i.i.d. sources satisfying Conditions 1)–3) of Theorem 3.2, and with support contained in some closed interval $[-B, B]$, i.e., $P_\theta\{|X| \leq B\} = 1$ for all $\theta \in \Theta$. Let also $\hat{\mathcal{X}} = \mathbb{R}$, and consider the squared-error distortion $\rho(x, \hat{x}) = |x - \hat{x}|^2$. Without loss of generality, we may assume that the optimal n-block quantizers C_θ^n have nearest-neighbor encoders, which in turn allows us to limit our consideration only to those quantizers whose codevectors have all their components in $[-B, B]$. Then ρ is bounded with $\rho_{\max} = 4B^2$, and Theorem 3.2 guarantees the existence of a weakly minimax universal sequence $\{C^{n,n}\}$ of (n, n)-block codes satisfying (4) and (5). Comparing this with the results of Linder *et al.* quoted in the preceding paragraph, we see that, as far as the rate and the distortion redundancy go, our scheme performs as well as that of [10], [11], but, again, in our case the extra $O(\log n)$ bits have been utilized more efficiently, enabling the decoder to identify the active source with good precision. However, the big difference between our code and that of Linder *et al.* is that the class of sources considered by them is *fully nonparametric*, whereas our development requires that the sources belong to a compactly parametrized family.

C. Extension to Curved Parametric Families

We can also consider parameter spaces that are more general than bounded subsets of $\mathbb{R}^k$. For instance, in information geometry [28], one often encounters *curved parametric families*, i.e., families $\{P_\theta : \theta \in \Theta\}$ of probability distributions where the parameter space Θ is a smooth compact manifold. Roughly

speaking, an abstract set Θ is a smooth compact manifold of dimension k if it admits a covering by finitely many sets $G_l \subset \Theta$, such that for each l there exists a one-to-one map ξ_l of G_l onto a pre-compact subset F_l of $\mathbb{R}^k$; the maps ξ_l are also required to satisfy a certain smooth compatibility condition, but we need not consider it here. The pairs (G_l, ξ_l) are called the *charts* of Θ.

In order to cover this case, we need to make the following modifications in the statement and in the proof of Theorem 3.2. First of all, let $\{P_\theta : \theta \in \Theta\}$ satisfy Condition 3) of the theorem, and replace Condition 2) with

2a) For each l, the map $u \mapsto P_{\xi_l^{-1}(u)}, u \in F_l$ is uniformly locally Lipschitz: there exist constants $r_l > 0$ and $m_l > 0$, such that for every $u \in F_l$

$$d_V\left(P_{\xi_l^{-1}(u)}, P_{\xi_l^{-1}(w)}\right) \leq m_l\|u - w\|$$

for all $w \in B_{r_l}(u)$.

[Note that $\xi_l^{-1}(u) \in G_l \subset \Theta$ for all $u \in F_l$.] Condition 1) is satisfied for each F_l by definition of Θ. Next, we need to modify the first-stage encoder. For each l, quantize F_l in cubes of side $1/\lceil n^{1/2} \rceil$, so that each $u \in F_l$ can be encoded into $k(\log n^{1/2} + \log J_l)$ bits, for some J_l, and reproduced by some $\hat{u} \in F_l$ satisfying $\|u - \hat{u}\| \leq \sqrt{k/n}$. Then $\theta = \xi_l^{-1}(u)$ and $\hat{\theta} = \xi_l^{-1}(\hat{u})$ both lie in $G_l \subset \Theta$. Now, when the first-stage encoder computes the minimum-distance estimate $\tilde{\theta}$ of the active source θ, it will prepend a fixed-length binary description of the index l such that $\tilde{\theta} \in G_l$ to the binary description of the cube in $\mathbb{R}^k$ containing $\tilde{u} = \xi_l(\tilde{\theta})$. Let $\hat{u}$ be the reproduction of $\tilde{u}$ under the cubic quantizer for F_l. The per-letter rate of the resulting two-stage code is

$$R(C^{n,n}) = R + O\left(\frac{\log n}{n}\right)$$

bits per letter. The nth-order distortion redundancy is bounded as

$$\delta_\theta^n(C^{n,n}) \leq 4\rho_{\max}\left\{\mathbb{E}_\theta[d_V(P_\theta, P_{\tilde{\theta}})] + \mathbb{E}_\theta[d_V(P_{\tilde{\theta}}, P_{\hat{\theta}})]\right\}$$

where $\hat{\theta} = \xi_l^{-1}(\hat{u})$. The first term in the brackets is upper-bounded by means of the usual VC estimate

$$\mathbb{E}_\theta[d_V(P_\theta, P_{\tilde{\theta}})] \leq 2c_1\sqrt{\frac{\log n}{n}} + \frac{3}{2n}$$

while the second term is handled using Condition 2a). Specifically, if $\tilde{\theta} \in G_l$, then $P_{\tilde{\theta}} = P_{\xi_l^{-1}(\tilde{u})}$ and $P_{\hat{\theta}} = P_{\xi_l^{-1}(\hat{u})}$. Then the same argument as in the proof of Theorem 3.2 can be used to show that there exists a constant $b_l > 0$ such that $d_V(P_{\tilde{\theta}}, P_{\hat{\theta}}) \leq b_l\sqrt{k/n}$, which can be further bounded by $b\sqrt{k/n}$ with $b \triangleq \max_l b_l$. Combining all these bounds and using Proposition 3.1, we get that the distortion redundancy is

$$\delta_\theta(C^{n,n}) = O\left(\sqrt{\frac{\log n}{n}}\right).$$

This establishes that $\{C^{n,n}\}$ is weakly minimax universal for the curved parametric family $\{P_\theta : \theta \in \Theta\}$. The fidelity of the source identification procedure is similar to that in the "flat"

case $\Theta \subset \mathbb{R}^k$, by the same Borel–Cantelli arguments as in the proof of Theorem 3.2.

D. Extension to Unbounded Distortion Measures

In this section, we show that the boundedness condition on the distortion measure can be relaxed, so that our approach can work with any distortion measure satisfying a certain moment condition, except that the distortion redundancy will converge to zero at a slower rate of $O(\sqrt[4]{n^{-1}\log n})$ instead of $O(\sqrt{n^{-1}\log n})$, as in the bounded case.

Specifically, let $\{P_\theta : \theta \in \Theta\}$ be a family of i.i.d. sources satysfing the conditions of Theorem 3.2, and let $\rho : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ be a single-letter distortion function for which there exists a *reference letter* $a_* \in \hat{\mathcal{X}}$ such that

$$\int_{\mathcal{X}} \rho^2(x, a_*) dP_\theta(x) \le G < \infty \qquad (18)$$

for all $\theta \in \Theta$, and which has the form $\rho(x, \hat{x}) = [d(x, \hat{x})]^p$ for some p, where $d(\cdot, \cdot)$ is a metric on $\mathcal{X} \cup \hat{\mathcal{X}}$. In the following, we shall show that for any rate $R > 0$ satisfying

$$D(R, \Theta) \triangleq \sup_{\theta \in \Theta} D_\theta(R) < \infty$$

and for any $\epsilon > 0$ there exists a sequence $\{C^{n,n}\}_{n=1}^\infty$ of two-stage (n, n)-block codes, such that

$$R(C^{n,n}) \le R + \epsilon + O\left(\frac{\log n}{n}\right)$$

and

$$\delta_\theta(C^{n,n}) \le \epsilon + O\left(\sqrt[4]{\frac{\log n}{n}}\right)$$

for every $\theta \in \Theta$. Taking a cue from García-Muñoz and Neuhoff [29], we shall call a sequence of codes $\{C^{n,n}\}$ satisfying

$$\lim_{n \to \infty} R(C^{n,n}) \le R + \epsilon$$

and

$$\lim_{n \to \infty} \delta_\theta(C^{n,n}) \le \epsilon, \quad \forall \theta \in \Theta$$

for a given $\epsilon > 0$ ϵ-*weakly minimax universal* for $\{P_\theta : \theta \in \Theta\}$. By continuity, the existence of ϵ-weakly minimax universal codes for all $\epsilon > 0$ then implies the existence of weakly minimax universal codes in the sense of Definition 2.2. Moreover, we shall show that the convergence rate of the source identification procedure is the same as in the case of a bounded distortion function, namely, $O(\sqrt{n^{-1}\log n})$; in particular, the constant implicit in the $O(\cdot)$ notation depends neither on ϵ nor on the behavior of ρ.

The following proof draws upon some ideas of Dobrushin [30], the difference being that he considered *robust*, rather than universal, codes.[2] Let $M > 0$ be a constant to be specified later,

---

[2] A sequence of lossy codes is (strongly) robust for a given class of information sources at rate R (see, e.g., [31]–[33]) if its asymptotic performance on each source in the class is no worse than the supremum of the distortion-rate functions of all the sources in the class at R. Neuhoff and García-Muñoz [33] have shown that strongly robust codes occur more widely than strongly minimax universal codes, but less widely than weakly minimax universal ones.

and define a single-letter distortion function $\rho_M : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ by

$$\rho_M(x, \hat{x}) \triangleq \begin{cases} \rho(x, \hat{x}), & \text{if } \rho(x, \hat{x}) \le M \\ M, & \text{if } \rho(x, \hat{x}) > M. \end{cases}$$

Let $\bar{D}_\theta(C^{n,m})$ denote the average per-letter ρ_M-distortion of an (n, m)-block code $C^{n,m}$ with respect to P_θ, and let $\bar{D}_\theta(R)$ denote the corresponding Shannon DRF. Then Theorem 3.2 guarantees that for every $R > 0$ there exists a weakly minimax universal sequence $\{\bar{C}^{n,n}\}_{n=1}^\infty$ of two-stage (n, n)-block codes, such that

$$R(\bar{C}^{n,n}) = R + O\left(\frac{\log n}{n}\right)$$

and

$$\bar{D}_\theta(\bar{C}^{n,n}) = \bar{D}_\theta(R) + O\left(\sqrt{\frac{\log n}{n}}\right) \qquad (19)$$

for all $\theta \in \Theta$.

We shall now modify $\bar{C}^{n,n}$ to obtain a new code $C^{n,n}$. Fix some $\delta > 0$, to be chosen later. Let $\{\bar{C}_{\tilde{s}}^n : \tilde{s} \in \tilde{\mathcal{S}}\}$ be the collection of the second-stage codes of $\bar{C}^{n,n}$. Fix $\tilde{s} \in \tilde{\mathcal{S}}$ and let $\bar{\Gamma}_{\tilde{s}}$ be the reproduction codebook of $\bar{C}_{\tilde{s}}^n$. Let $\Gamma_{\tilde{s}} \subset \hat{\mathcal{X}}^n$ be the set consisting of a) all codevectors in $\bar{\Gamma}_{\tilde{s}}$, b) all vectors obtained by replacing $\lfloor \delta n \rfloor$ or fewer components of each codevector in $\bar{\Gamma}_{\tilde{s}}$ with a_*, and c) the vector a_*^n. The size of $\Gamma_{\tilde{s}}$ can be estimated by means of Stirling's formula as

$$|\Gamma_{\tilde{s}}| = |\bar{\Gamma}_{\tilde{s}}| \sum_{i=0}^{\lfloor \delta n \rfloor} \binom{n}{i} + 1 \le |\bar{\Gamma}_{\tilde{s}}| 2^{n[h(\delta) + o(1)]} + 1$$

where $h(\delta) \triangleq -\delta \log \delta - (1 - \delta) \log(1 - \delta)$ is the binary entropy function. Since $|\bar{\Gamma}_{\tilde{s}}| = 2^{nR}$, we can choose δ small enough so that

$$|\Gamma_{\tilde{s}}| \le 2^{n(R+\epsilon)}. \qquad (20)$$

Now, if $\bar{C}_{\tilde{s}}^n$ maps a given $x^n \in \mathcal{X}^n$ to $\hat{\tilde{x}}^n = (\hat{\tilde{x}}_1, \dots, \hat{\tilde{x}}_n) \in \hat{\mathcal{X}}^n$, define a new string $\hat{x}^n = (\hat{x}_1, \dots, \hat{x}_n) \in \hat{\mathcal{X}}^n$ as follows. If $|\{1 \le i \le n : \rho(x_i, \hat{\tilde{x}}_i) > M\}| \le \delta n$, let

$$\hat{x}_i \triangleq \begin{cases} \hat{\tilde{x}}_i, & \text{if } \rho(x_i, \hat{\tilde{x}}_i) \le M \\ a_*, & \text{if } \rho(x_i, \hat{\tilde{x}}_i) > M; \end{cases}$$

otherwise, let $\hat{x}^n = a_*^n$. Now, construct a new code $C_{\tilde{s}}^n$ with the codebook $\Gamma_{\tilde{s}}$, and with the encoder and the decoder defined in such a way that $C_{\tilde{s}}^n(x^n) = \hat{x}^n$ whenever $\bar{C}_{\tilde{s}}^n(x^n) = \hat{\tilde{x}}^n$. Finally, let $C^{n,n}$ be a two-stage code with the same first-stage encoder as $\bar{C}^{n,n}$, but with the collection of the second-stage codes replaced by $\{C_{\tilde{s}}^n\}$. From (20) it follows that $R(C^{n,n}) \le R(\bar{C}^{n,n}) + \epsilon$. Since $R(\bar{C}^{n,n}) = R + O(n^{-1}\log n)$, we have that

$$R(C^{n,n}) \le R + \epsilon + O\left(\frac{\log n}{n}\right). \qquad (21)$$

Furthermore, the code $C^{n,n}$ has the following property.

Lemma 4.1: Let $G' = G(1 + 2/\delta)$. Then for any $\theta \in \Theta$,

$$D_\theta(C^{n,n}) \le \bar{D}_\theta(\bar{C}^{n,n}) + \sqrt{\frac{G' \bar{D}_\theta(\bar{C}^{n,n})}{M}}. \qquad (22)$$

Proof: See Appendix D. $\square$

Substituting (19) into (22), we have that

$$D_\theta(C^{n,n}) \leq \bar{D}_\theta(R) + O\left(\sqrt{\frac{\log n}{n}}\right)$$
$$+ \sqrt{\frac{G'}{M}\bar{D}_\theta(R) + O\left(\sqrt{\frac{\log n}{n}}\right)}.$$

Now, since $\rho_M(x,\hat{x}) \leq \rho(x,\hat{x})$ for all $(x,\hat{x}) \in \mathcal{X} \times \hat{\mathcal{X}}, \bar{D}_\theta(R) \leq D_\theta(R)$ for all $\theta \in \Theta$. Using this fact and the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we can write

$$D_\theta(C^{n,n}) \leq D_\theta(R) + O\left(\sqrt{\frac{\log n}{n}}\right)$$
$$+ \sqrt{\frac{G'}{M}D_\theta(R)} + O\left(\sqrt[4]{\frac{\log n}{n}}\right).$$

Upon choosing M so that $\sqrt{G'D(R,\Theta)/M} < \epsilon$, we get

$$\delta_\theta(C^{n,n}) \leq \epsilon + O\left(\sqrt[4]{\frac{\log n}{n}}\right). \qquad (23)$$

Thus, (21) and (23) prove the claim made at the beginning of the section. Moreover, because the first-stage encoder of $C^{n,n}$ is the same as in $\bar{C}^{n,n}$, our code modification procedure has no effect on parameter estimation, so the same arguments as in the end of the proof of Theorem 3.2 can be used to show that the decoder can identify the source in effect up to a variational ball of radius $O(\sqrt{n^{-1}\log n})$ asymptotically almost surely.

V. EXAMPLES

In this section, we present a detailed analysis of two classes of parametric sources that meet Conditions 1)–3) of Theorem 3.2, and thus admit schemes for joint universal lossy coding and modeling. These are finite mixture classes and exponential families, which are widely used in statistical modeling, both in theory and in practice (see, e.g., [34]–[37]).

A. Mixture Classes

Let $p_1,\ldots,p_k$ be fixed probability densities on a measurable $\mathcal{X} \subseteq \mathbb{R}^d$, and let

$$\Theta \triangleq \left\{ \theta = (\theta_1,\ldots,\theta_k) \in \mathbb{R}^k : \right.$$
$$\left. 0 \leq \theta_i \leq 1, i = 1,\ldots,k; \sum_{i=1}^k \theta_i = 1 \right\}$$

be the probability k-simplex. Then the mixture class defined by the p_i's consists of all densities of the form

$$p_\theta(x) = \sum_{i=1}^k \theta_i p_i(x), \quad \theta \in \Theta.$$

The parameter space Θ is obviously compact, which establishes Condition 1) of Theorem 3.2. In order to show that Condition 2) holds, fix any $\theta, \eta \in \Theta$. Then

$$d_V(P_\theta, P_\eta) = \frac{1}{2}\int_\mathcal{X} |p_\theta(x) - p_\eta(x)|dx$$
$$\leq \frac{1}{2}\sum_{i=1}^k \int_\mathcal{X} |\theta_i - \eta_i|p_i(x)dx$$

$$= \frac{1}{2}\sum_{i=1}^k |\theta_i - \eta_i|$$
$$\leq \frac{\sqrt{k}}{2}\sqrt{\sum_{i=1}^k (\theta_i - \eta_i)^2}$$
$$= \frac{\sqrt{k}}{2}\|\theta - \eta\|$$

where the last inequality is a consequence of the concavity of the square root. This implies that the map $\theta \mapsto P_\theta$ is everywhere Lipschitz with Lipschitz constant $\sqrt{k}/2$. We have left to show that Condition 3) of Theorem 3.2 holds as well, i.e., that the Yatracos class

$$\mathcal{A}_\Theta = \{A_{\theta,\eta} = \{x \in \mathcal{X} : p_\theta(x) > p_\eta(x)\} : \theta,\eta \in \Theta; \theta \neq \eta\}$$

has finite VC dimension. To this end, observe that $x \in A_{\theta,\eta}$ if and only if

$$\sum_{i=1}^k (\theta_i - \eta_i)p_i(x) > 0.$$

Thus, $\mathcal{A}_\Theta$ consists of sets of the form

$$\left\{ x \in \mathcal{X} : \sum_{i=1}^k \alpha_i p_i(x) > 0, \alpha = (\alpha_1,\ldots,\alpha_k) \in \mathbb{R}^k \right\}.$$

Since the functions $p_1,\ldots,p_k$ span a linear space whose dimension is not larger than k, Lemma A.3 in the Appendx guarantees that $\mathsf{V}(\mathcal{A}_\Theta) \leq k$, which establishes Condition 3).

B. Exponential Families

Let $\mathcal{X}$ be a measurable subset of $\mathbb{R}^d$, and let Θ be a compact subset of $\mathbb{R}^k$. A family $\{p_\theta : \theta \in \Theta\}$ of probability densities on $\mathcal{X}$ is an *exponential family* [28], [35] if each p_θ has the form

$$p_\theta(x) = p(x)\exp\left(\sum_{i=1}^k \theta_i h_i(x) - g(\theta)\right)$$
$$\equiv p(x)e^{\theta \cdot h(x) - g(\theta)}$$

where p is a fixed reference density, $h_1,\ldots,h_k$ are fixed real-valued functions on $\mathcal{X}$, and

$$g(\theta) = \ln \int_\mathcal{X} e^{\theta \cdot h(x)}p(x)dx$$

is the normalization constant. By way of notation

$$h(x) \triangleq (h_1(x),\ldots,h_k(x)) \quad \text{and} \quad \theta \cdot h(x) \triangleq \sum_{i=1}^k \theta_i h_i(x).$$

Given the densities p and p_θ, let P and P_θ denote the corresponding distributions. The assumed compactness of Θ guarantees that the family $\{P_\theta : \theta \in \Theta\}$ satisfies Condition 1) of Theorem 3.2. In the following, we shall demonstrate that Conditions 2) and 3) can also be met under certain regularity assumptions.

It is customary to choose the functions h_i in such a way that $\{1, h_1,\ldots,h_k\}$ is a linearly independent set. This guarantees

that the map $\theta \mapsto P_\theta$ is one-to-one. We shall also assume that each h_i is square-integrable with respect to P

$$\int_{\mathcal{X}} h_i^2 dP \equiv \int_{\mathcal{X}} h_i^2(x) p(x) dx < \infty, \quad 1 \leq i \leq k.$$

Then, the $(k+1)$-dimensional real linear space $\mathcal{F} \subset L^2(\mathcal{X}, P)$ spanned by $\{1, h_1, \ldots, h_k\}$ can be equipped with an inner product

$$\langle f, g \rangle \triangleq \int_{\mathcal{X}} fg \, dP, \quad f, g \in \mathcal{F}$$

and the corresponding L_2 norm

$$\|f\|_2 \triangleq \sqrt{\langle f, f \rangle} \equiv \sqrt{\int_{\mathcal{X}} f^2 dP}, \quad f \in \mathcal{F}.$$

Also let

$$\|f\|_\infty \triangleq \inf\{M : |f(x)| \leq M \, P\text{-a.e.}\}$$

denote the L_∞ norm of f. Since $\mathcal{F}$ is finite-dimensional, there exists a constant $A_k > 0$ such that

$$\|f\|_\infty \leq A_k \|f\|_2.$$

Finally, assume that the logarithms of the Radon–Nikodym derivatives $dP/dP_\theta \equiv p/p_\theta$ are uniformly bounded P-a.e.: $\sup_{\theta \in \Theta} \|\log p/p_\theta\|_\infty < \infty$. Let

$$D(P_\theta \| P_\eta) \triangleq \int_{\mathcal{X}} \frac{dP_\theta}{dP_\eta} \ln \frac{dP_\theta}{dP_\eta} dP_\eta \equiv \int_{\mathcal{X}} p_\theta \ln \frac{p_\theta}{p_\eta} dx$$

denote the relative entropy (information divergence) between P_θ and P_η. Then we have the following basic estimate:

Lemma 5.1: For all $\theta, \eta \in \Theta$

$$D(P_\theta \| P_\eta) \leq \frac{1}{2} e^{\|\ln p/p_\theta\|_\infty} e^{2A_k \|\theta - \eta\|} \|\theta - \eta\|^2$$

where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^k$.

Proof: The proof is along the lines of Barron and Sheu [35, Lemma 5]. Without loss of generality, we may assume that the functions $\{h_0, h_1, \ldots, h_k\}, h_0 \equiv 1$, form an orthonormal set with respect to P

$$\langle h_i, h_j \rangle = \int_{\mathcal{X}} h_i h_j dP = \delta_{ij}, \quad 0 \leq i, j \leq k.$$

Then

$$\|(\theta - \eta) \cdot h\|_2 = \left\| \sum_{i=1}^{k} (\theta_i - \eta_i) h_i \right\|_2 = \|\theta - \eta\|.$$

Now, since

$$g(\eta) - g(\theta) = \ln \int_{\mathcal{X}} e^{(\eta - \theta) \cdot h} dP_\theta$$
$$= \ln \int_{\mathcal{X}} e^{(\eta - \theta) \cdot h(x)} p_\theta(x) dx$$

we have

$$|g(\eta) - g(\theta)| \leq \|(\eta - \theta) \cdot h\|_\infty$$
$$\leq A_k \|(\eta - \theta) \cdot h\|_2$$
$$= A_k \|\eta - \theta\|.$$

Furthermore

$$\ln \frac{p_\theta}{p_\eta} = (\theta - \eta) \cdot h + g(\eta) - g(\theta)$$

whence it follows that the logarithm of the Radon–Nikodym derivative $dP_\theta/dP_\eta = p_\theta/p_\eta$ is bounded P-a.e.: $\|\ln p_\theta/p_\eta\|_\infty \leq 2\|(\eta - \theta) \cdot h\|_\infty \leq 2A_k \|\eta - \theta\|$. In this case, the relative entropy $D(P_\theta \| P_\eta)$ satisfies [35, Lemma 1]

$$D(P_\theta \| P_\eta) \leq \frac{1}{2} e^{\|\ln p_\theta/p_\eta - c\|_\infty} \int_{\mathcal{X}} \left(\ln \frac{p_\theta}{p_\eta} - c \right)^2 dP_\theta$$

for any constant c. Choosing $c = g(\eta) - g(\theta)$ and using the orthonormality of the h_i, we get

$$D(P_\theta \| P_\eta) \leq \frac{1}{2} e^{\|(\theta - \eta) \cdot h\|_\infty} \int_{\mathcal{X}} ((\theta - \eta) \cdot h)^2 dP_\theta$$
$$= \frac{1}{2} e^{\|(\theta - \eta) \cdot h\|_\infty} \int_{\mathcal{X}} \frac{p_\theta}{p} ((\theta - \eta) \cdot h)^2 dP$$
$$\leq \frac{1}{2} e^{\|\ln p/p_\theta\|_\infty} e^{2A_k \|\theta - \eta\|} \|(\theta - \eta) \cdot h\|_2^2$$
$$= \frac{1}{2} e^{\|\ln p/p_\theta\|_\infty} e^{2A_k \|\theta - \eta\|} \|\theta - \eta\|^2$$

and the lemma is proved. $\qquad \square$

Now, using Pinsker's inequality [38, Lemma 5.2.8]

$$d_V(P_\theta, P_\eta) \leq \sqrt{(1/2) D(P_\theta \| P_\eta)}$$

together with the above lemma and the assumed uniform boundedness of $\ln p/p_\theta$, we get the bound

$$d_V(P_\theta, P_\eta) \leq m_0 e^{A_k \|\theta - \eta\|} \|\theta - \eta\|, \quad \theta, \eta \in \Theta \qquad (24)$$

where $m_0 \triangleq \frac{1}{2} \exp\left(\frac{1}{2} \sup_{\theta \in \Theta} \|\ln p/p_\theta\|_\infty\right)$. If we fix $\theta \in \Theta$, then from (24) it follows that, for any $r > 0$

$$d_V(P_\theta, P_\eta) \leq m_0 e^{A_k r} \|\theta - \eta\|$$

for all η satisfying $\|\eta - \theta\| \leq r$. That is, the family $\{P_\theta : \theta \in \Theta\}$ satisfies the uniform local Lipschitz condition (Condition 2) of Theorem 3.2, and the magnitude of the Lipschitz constant can be controlled by tuning r.

All we have left to show is that the VC condition (Condition 3) of Theorem 3.2 is satisfied. Let $\theta, \eta \in \Theta$ be distinct; then $p_\theta(x) > p_\eta(x)$ if and only if $(\theta - \eta) \cdot h(x) > g(\theta) - g(\eta)$. Thus, the corresponding Yatracos class $\mathcal{A}_\Theta$ consists of sets of the form

$$\left\{ x \in \mathcal{X} : \alpha_0 + \sum_{i=1}^{k} \alpha_i h_i(x) > 0, \right.$$
$$\left. \alpha = (\alpha_0, \alpha_1, \ldots, \alpha_k) \in \mathbb{R}^{k+1} \right\}.$$

Since the functions $1, h_1, \ldots, h_k$ span a $(k+1)$-dimensional linear space, $\mathsf{V}(\mathcal{A}_\Theta) \leq k+1$ by Lemma A.3.

VI. Summary and Discussion

We have constructed and analyzed a scheme for universal fixed-rate lossy coding of continuous-alphabet i.i.d. sources based on a forward relation between statistical modeling and universal coding, in the spirit of Rissanen's achievability theorem [1, Theorem 1b] (see also [13, Theorem 2]). To the best of our knowledge, such a joint universal source-coding and source-modeling scheme has not been constructed before, although Chou *et al.* [13] have demonstrated the existence of universal vector quantizers whose Lagrangian redundancies converge to zero at the same rate as the corresponding redundancies in Rissanen's achievability theorem for the lossless case. What we have shown is that, for a wide class of bounded distortion measures and for any compactly parametrized family of i.i.d. sources with absolutely continuous distributions satisfying a smoothness condition and a VC learnability condition, the tasks of parameter estimation (statistical modeling) and universal lossy coding can be accomplished jointly in a two-stage setup, with the overhead per-letter rate and the distortion redundancy converging to zero as $O(n^{-1} \log n)$ and $O(\sqrt{n^{-1} \log n})$, respectively, as the block length n tends to infinity, and the extra bits generated by the first-stage encoder can be used to identify the active source up to a variational ball of radius $O(\sqrt{n^{-1} \log n})$ (a.s.). We have compared our scheme with several existing schemes for universal vector quantization and demonstrated that our approach offers essentially similar performance in terms of rate and distortion, while also allowing the decoder to reconstruct the statistics of the source with good precision. We have described an extension of our scheme to unbounded distortion measures satisfying a moment condition with respect to a reference letter, which suffers no change in overhead rate or in source estimation fidelity, although it gives a slower, $O(\sqrt[4]{n^{-1} \log n})$, convergence rate for distortion redundancies. Finally, we have presented detailed examples of parametric sources satisfying the conditions of our Theorem 3.2 (namely, finite mixture classes and exponential families) and thus admitting schemes for joint universal quantization and modeling.

As mentioned in the Introduction, in treating universal lossy source coding as a statistical problem the term "model" can refer either to a probabilistic description of the source or to a probabilistic description of a rate–distortion codebook. In fact, as shown by Kontoyiannis and Zhang [39], for variable-rate lossy codes operating under a fixed distortion constraint, there is a one-to-one correspondence between codes and discrete distributions over sequences in the reproduction space, which they dubbed the "lossy Kraft inequality." The same paper also demonstrated the existence of variable-rate universal lossy codes for finite-alphabet memoryless sources with rate redundancy converging to zero as $(k/2) \log n / n$, where k is the dimension of the simplex of probability distributions on the reproduction alphabet. Yang and Zhang [40] proved an analogous result for fixed-rate universal lossy codes and showed furthermore that the $(k/2) \log n / n$ convergence rate

is *optimal* in a certain sense. (The redundancies in our scheme are therefore suboptimal, as can be seen from comparing them to these bounds, as well as to those of Chou *et al.* [13]. It is certainly an interesting open problem to determine lower bounds on the redundancies in the setting of joint source coding and identification.) These papers, together with the work of Madiman, Harrison, and Kontoyiannis [41], [42] can be thought of as generalizing Rissanen's MDL principle to lossy setting, provided that the term "model" is understood to refer to probability distributions over codebooks in the reproduction space.

We close by outlining several potential directions for further research. First of all, it would be of both theoretical and practical interest to extend the results presented here to sources with memory, to allow more realistic source models such as autoregressive or Markov sources, and to variable-rate codes, so that unbounded parameter spaces could be accommodated. We have made some initial progress in this direction in [43], [44], where we constructed joint schemes for variable-rate universal lossy coding and identification of stationary ergodic sources satisfying a certain mixing condition. Moreover, the theory presented here needs to be tested in practical settings, one promising area for applications being *media forensics* [45], where the parameter θ could represent traces or "evidence" of some prior processing performed, say, on an image or on a video sequence, and where the goal is to design an efficient system for compressing the data for the purposes of transmission or storage in such a way that the evidence can be later recovered from the compressed signal with minimal degradation in fidelity.

Appendix A
Vapnik–Chervonenkis (VC) Theory

In this appendix, we summarize, for the reader's convenience, some basic concepts and results of the VC theory. A detailed treatment can be found, e.g., in [19].

Definition A.1 (Shatter Coefficient): Let $\mathcal{A}$ be an arbitrary collection of measurable subsets of $\mathbb{R}^d$. Given an n-tuple $x^n = (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$, let $\mathcal{A}(x^n)$ be the subset of $\{0,1\}^n$ obtained by listing all distinct binary strings of the form $(1_{\{x_1 \in A\}}, \ldots, 1_{\{x_n \in A\}})$ as A is varied over $\mathcal{A}$. Then

$$\mathsf{S}_{\mathcal{A}}(n) \triangleq \max_{x^n \in (\mathbb{R}^d)^n} |\mathcal{A}(x^n)|$$

is called the *nth shatter coefficient* of $\mathcal{A}$.

Definition A.2 (VC Dimension; VC Class): The largest integer n for which $\mathsf{S}_{\mathcal{A}}(n) = 2^n$ is called the *Vapnik–Chervonenkis dimension* (or the VC dimension, for short) of $\mathcal{A}$ and denoted by $\mathsf{V}(\mathcal{A})$. If $\mathsf{S}_{\mathcal{A}}(n) = 2^n$ for all $n = 1, 2, \ldots$, then we define $\mathsf{V}(\mathcal{A}) = \infty$. If $\mathsf{V}(\mathcal{A}) < \infty$, we say that $\mathcal{A}$ is a *Vapnik–Chervonenkis class* (or VC class).

The basic result of VC theory relates the shatter coefficient $\mathsf{S}_{\mathcal{A}}(n)$ to uniform deviations of the probabilities of events in $\mathcal{A}$ from their relative frequencies with respect to an i.i.d. sample of size n.

Lemma A.1 (The Vapnik–Chervonenkis Inequalities): Let $\mathcal{A}$ be an arbitrary collection of measurable subsets of $\mathbb{R}^d$, and let $X^n = (X_1, \ldots, X_n)$ be an n-tuple of i.i.d. random variables in $\mathbb{R}^d$ with the common distribution P. Then

$$\mathbb{P}\left\{\sup_{A \in \mathcal{A}} |P_{X^n}(A) - P(A)| > \epsilon\right\} \leq 8\mathsf{S}_{\mathcal{A}}(n)e^{-n\epsilon^2/32} \quad \text{(A1)}$$

for any $\epsilon > 0$, and

$$\mathbb{E}\left\{\sup_{A \in \mathcal{A}} |P_{X^n}(A) - P(A)|\right\} \leq 2\sqrt{\frac{\log 2\mathsf{S}_{\mathcal{A}}(n)}{n}} \quad \text{(A2)}$$

where P_{X^n} is the empirical distribution of X^n:

$$P_{X^n}(B) \triangleq \frac{1}{n}\sum_{i=1}^n 1_{\{X_i \in B\}}$$

for all Borel sets $B \subset \mathbb{R}^d$. The probabilities and expectations are with respect to P.

Now, if $\mathcal{A}$ is a VC class and $\mathsf{V}(\mathcal{A}) \geq 2$, then the results of Vapnik and Chervonenkis [46] and Sauer [47] imply that $\mathsf{S}_{\mathcal{A}}(n) \leq n^{\mathsf{V}(\mathcal{A})}$. Plugging this bound into (A1) and (A2), we obtain the following.

Lemma A.2: If $\mathcal{A}$ is a VC class with $\mathsf{V}(\mathcal{A}) \geq 2$, then

$$\mathbb{P}\left\{\sup_{A \in \mathcal{A}} |P_{X^n}(A) - P(A)| > \epsilon\right\} \leq 8n^{\mathsf{V}(\mathcal{A})}e^{-n\epsilon^2/32}$$

for any $\epsilon > 0$, and

$$\mathbb{E}\{\sup_{A \in \mathcal{A}} |P_{X^n}(A) - P(A)|\} \leq c\sqrt{\frac{\log n}{n}} \quad \text{(A3)}$$

where c is a constant that depends only on $\mathsf{V}(\mathcal{A})$.

Remark A.1: One can use more delicate arguments involving metric entropies and covering numbers, along the lines of Dudley [48], to improve the bound in (A3) to $c'\sqrt{1/n}$, where $c' = c'(\mathsf{V}(\mathcal{A}))$ is another constant. However, c' turns out to be much larger than c, so that, for all "practical" values of n, the "improved" $O(\sqrt{1/n})$ bound is much worse than the original $O(\sqrt{n^{-1}\log n})$ bound.

Lemma A.3: Let $\mathcal{F}$ be an m-dimensional linear space of real-valued functions on $\mathbb{R}^d$. Then the class

$$\mathcal{A} = \{\{x : f(x) \geq 0\} : f \in \mathcal{F}\}$$

is a VC class, and $\mathsf{V}(\mathcal{A}) \leq m$.

APPENDIX B
PROOF OF PROPOSITION 3.1

Fix $\theta \in \Theta$, and let X be distributed according to P_θ. Let the distortion function ρ satisfy Condition 1) of Proposition 3.1. Then a result of Csiszár [23] says that, for each point $(R, D_\theta(R))$ on the distortion–rate curve for P_θ, there exists a random variable Y with values in the reproduction alphabet $\hat{\mathcal{X}}$, where the joint distribution of X and Y is such that

$$I(X, Y) = R \quad \text{and} \quad \mathbb{E}_{XY}[\rho(X, Y)] = D_\theta(R) \quad \text{(B1)}$$

and the Radon–Nikodym derivative

$$a(x, y) \triangleq \frac{dP_{XY}}{d(P_X \times P_Y)}(x, y)$$

where $P_X \equiv P_\theta$ has the parametric form

$$a(x, y) = \alpha(x)2^{-s\rho(x,y)} \quad \text{(B2)}$$

where $s \geq 0$ and $\alpha(x) \geq 1$ satisfy

$$\int_{\mathcal{X}} \alpha(x)2^{-\rho(x,y)}dP_\theta(x) \leq 1, \quad \forall y \in \hat{\mathcal{X}}$$

and $-1/s = D_\theta'(R)$, the derivative of the DRF $D_\theta(R)$ at R, i.e., $-1/s$ is the slope of the tangent to the graph of $D_\theta(R)$ at R.

Next, let $N = \lfloor 2^{n(R+\delta)} \rfloor$, where $\delta > 0$ will be specified later, and generate a random codebook $\mathcal{W}$ as a vector $\mathcal{W} = (W_1, \ldots, W_N)$, where each $W_i = (W_{i1}, \ldots, W_{in}) \in \hat{\mathcal{X}}^n$, and the W_{ij}'s are i.i.d. according to P_Y. Thus

$$P_{\mathcal{W}} = \mathsf{X}_{i=1}^N P_Y^n$$

is the probability distribution for the randomly selected codebook. We also assume that $\mathcal{W}$ is independent from $X^n = (X_1, \ldots, X_n)$. Now, let $C_{\mathcal{W}}$ be a (random) n-block code with the reproduction codebook $\mathcal{W}$ and the minimum-distortion encoder, so that $\rho(x^n, C_{\mathcal{W}}(x^n)) = \rho(x^n, \mathcal{W})$, where

$$\rho(x^n, \mathcal{W}) \triangleq \min_{1 \leq i \leq N} \rho(x^n, W_i).$$

Then the average per-letter distortion of this random code over the codebook generation and the source sequence is

$$\Delta_n = \int D_\theta(C_{\boldsymbol{w}})dP_{\mathcal{W}}(\boldsymbol{w})$$
$$= \frac{1}{n}\int \mathbb{E}_\theta\left[\rho(X^n, \boldsymbol{w})\right]dP_{\mathcal{W}}(\boldsymbol{w})$$
$$= \frac{1}{n}\int (\int_{\mathcal{X}^n} \rho(x^n, \boldsymbol{w})dP_\theta(x^n))dP_{\mathcal{W}}(\boldsymbol{w}).$$

Using standard arguments (see, e.g., Gallager's proof of the source coding theorem [22, Ch. 9]), we can bound Δ_n from above as

$$\Delta_n \leq D_\theta(R) + \delta + \rho_{\max}P_{XY}(Y^n \notin \mathcal{S}_{X^n})$$
$$+ \rho_{\max}e^{-N2^{-n(R+\delta/2)}} \quad \text{(B3)}$$

where

$$\mathcal{S}_{x^n} \triangleq \Big\{y^n \in \hat{\mathcal{X}}^n : \rho(x^n, y^n) \leq n(D_\theta(R) + \delta)$$
$$\text{and } i_n(x^n, y^n) \leq n(R + \delta/2)\Big\}$$

and

$$i_n(x^n, y^n) \triangleq \sum_{i=1}^n \log a(x_i, y_i).$$

is the sample mutual information. Here, the pairs (X_i, Y_i) are i.i.d. according to P_{XY}. Now, by the union bound

$$P_{XY}(Y^n \notin \mathcal{S}_{X^n})$$
$$\leq P_{XY}\left(\frac{1}{n}i_n(X^n, Y^n) \geq R + \delta/2\right)$$
$$+ P_{XY}\left(\frac{1}{n}\rho(X^n, Y^n) \geq D_\theta(R) + \delta\right). \quad \text{(B4)}$$

Note that both $i_n(X^n, Y^n)$ and $\rho(X^n, Y^n)$ are sums of i.i.d. random variables, and that from (B1) we have $\mathbb{E}_{XY}[\log a(X, Y)] = I(X, Y) = R$ and $\mathbb{E}_{XY}[\rho(X, Y)] =$

$D_\theta(R)$. Since $0 \leq \rho(X,Y) \leq \rho_{\max}$, the second probability on the right-hand side of (B4) can be bounded using Hoeffding's inequality [49], which states that for i.i.d. random variables $S_1, \ldots, S_n$ satisfying $a \leq S_i \leq b$ a.s.

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} S_i \geq \mathbb{E}[S_1] + \delta\right) \leq e^{-2n\delta^2/(b-a)^2}.$$

This yields the estimate

$$P_{XY}\left(\frac{1}{n}\sum_{i=1}^{n}\rho(X_i,Y_i) \geq D_\theta(R) + \delta\right) \leq e^{-2n\delta^2/\rho_{\max}^2}.$$
(B5)

In order to apply Hoeffding's inequality to the first probability on the right-hand side of (B4), we have to show that $\log a(X,Y)$ is bounded. From (B2) we have that $\log a(x,y) = \log \alpha(x) - s\rho(x,y)$. On the other hand, integrating both sides of (B2) with respect to P_Y, we get

$$\alpha(x) = \frac{1}{\int 2^{-s\rho(x,y)}dP_Y(y)} \qquad P_\theta\text{-a.e.}$$

Since $2^{-s\rho_{\max}} \leq 2^{-s\rho(x,y)} \leq 1$, we have that $1 \leq \alpha(x) \leq 2^{s\rho_{\max}}$, whence it follows that $-s\rho_{\max} \leq \log a(x,y) \leq s\rho_{\max}$. Thus, by Hoeffding's inequality

$$P_{XY}\left(\frac{1}{n}\sum_{i=1}^{n}\log a(X_i,Y_i) \geq R + \delta/2\right) \leq e^{-n\delta^2/8s^2\rho_{\max}^2}.$$
(B6)

Putting together (B3), (B5), and (B6), and using the fact that $N \geq 2^{n(R+\delta)} - 1$, we obtain

$$\Delta_n \leq D_\theta(R) + \delta$$
$$+ \rho_{\max}\left(e^{-2n\delta^2/\rho_{\max}^2} + e^{-n\delta^2/8s^2\rho_{\max}^2} + e^{-2^{n\delta/2}}\right).$$

Since Δ_n is the average of the expected distortion over the random choice of codes, it follows that there exists at least one code whose average distortion with respect to P_θ is smaller than Δ_n. Thus

$$\hat{D}_\theta^n(R+\delta) \leq D_\theta(R) + \delta$$
$$+ \rho_{\max}\left(e^{-2n\delta^2/\rho_{\max}^2} + e^{-n\delta^2/8s^2\rho_{\max}^2} + e^{-2^{n\delta/2}}\right).$$

Now, let $c_s = \max(\rho_{\max}/2, 2s\rho_{\max})$ and put $\delta = c_s\sqrt{n^{-1}\ln n}$ to get

$$\hat{D}_\theta^n\left(R + c_s\sqrt{\frac{\ln n}{n}}\right) - D_\theta(R) = (c_s + o(1))\sqrt{\frac{\ln n}{n}}. \quad \text{(B7)}$$

Because $-1/s$ is the slope of the tangent to the distortion–rate curve at the point $(R, D_\theta(R))$, and because $D_\theta(R)$ is nonincreasing in R, we have $-1/s' \leq -1/s$ for s' corresponding to another point $(R', D_\theta(R'))$ with $R' < R$. Thus, $c_{s'} \leq c_s$, and (B7) remains valid for all $R' \leq R$. Thus, let $R' = R - c_s\sqrt{n^{-1}\ln n}$ to get

$$\hat{D}_\theta^n(R) - D_\theta\left(R - c_s\sqrt{\frac{\ln n}{n}}\right) \leq (c_s + o(1))\sqrt{\frac{\ln n}{n}}.$$

Therefore, expanding $D_\theta(R)$ in a Taylor series to first order and recalling that $-1/s = D_\theta'(R)$, we see that

$$\hat{D}_\theta^n(R) = c_s\left(1 + \frac{1}{s} + o(1)\right)\sqrt{\frac{\ln n}{n}},$$

and the proposition is proved.

APPENDIX C
QUANTIZER MISTMATCH LEMMA

Lemma C.1: Let P and Q be two absolutely continuous probability distributions on $\mathcal{X} \subseteq \mathbb{R}^d$, with respective densities p and q, and let $\rho : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ be a single-letter distortion measure having the form $\rho(x,\hat{x}) = [d(x,\hat{x})]^p$, where $p > 0$ and $d(\cdot,\cdot)$ is a bounded metric on $\mathcal{X} \cup \hat{\mathcal{X}}$. Consider an n-block lossy code C^n with the nearest-neighbor encoder, and let

$$D_P(C^n) \triangleq \frac{1}{n}\mathbb{E}_P[\rho(X^n, C^n(X^n))]$$
$$= \frac{1}{n}\int_{\mathcal{X}^n}\rho(x^n, C^n(x^n))dP(x^n)$$

be the average per-letter distortion of C^n with respect to P. Define $D_Q(C^n)$ similarly. Then

$$|D_P(C^n)^{1/p} - D_Q(C^n)^{1/p}| \leq 2^{1/p}d_{\max}d_V(P,Q). \quad \text{(C8)}$$

Furthermore, the corresponding nth-order operational DRFs $\hat{D}_P^n(R)$ and $\hat{D}_Q^n(R)$ satisfy

$$\left|\hat{D}_P^n(R)^{1/p} - \hat{D}_Q^n(R)^{1/p}\right| \leq 2^{1/p}d_{\max}d_V(P,Q). \quad \text{(C9)}$$

Proof: The proof closely follows Gray, Neuhoff, and Shields [26]. Let $\mathcal{P}_n$ denote the set of all probability measures on $\mathcal{X}^n \times \mathcal{X}^n$ having P^n and Q^n as marginals, and let $\bar{\mu}$ achieve (or come arbitrarily close to) the infimum in the *Wasserstein metric*

$$\bar{\rho}_n(P,Q) \triangleq \inf_{\mu \in \mathcal{P}_n}\left(\frac{1}{n}\int_{\mathcal{X}^n \times \mathcal{X}^n}\rho(x^n, y^n)d\mu(x^n, y^n)\right)^{1/p}.$$

Suppose that $D_P(C^n) \leq D_Q(C^n)$. Then, using the fact that d is a metric, Minkowski's inequality, and the nearest-neighbor property of C^n, we have

$$D_P(C^n)^{1/p}$$
$$= \left(\frac{1}{n}\int_{\mathcal{X}^n}\rho(x^n, C^n(x^n))dP^n(x^n)\right)^{1/p}$$
$$= \left(\frac{1}{n}\int_{\mathcal{X}^n \times \mathcal{X}^n}\rho(x^n, C^n(x^n))d\bar{\mu}(x^n, y^n)\right)^{1/p}$$
$$\leq \left(\frac{1}{n}\int_{\mathcal{X}^n \times \mathcal{X}^n}\rho(x^n, y^n)d\bar{\mu}(x^n, y^n)\right)^{1/p}$$
$$+ \left(\frac{1}{n}\int_{\mathcal{X}^n \times \mathcal{X}^n}\rho(y^n, C^n(y^n))d\bar{\mu}(x^n, y^n)\right)^{1/p}$$
$$= \bar{\rho}_n(P,Q) + D_Q(C^n)^{1/p}$$

Now

$$\bar{\rho}_n(P,Q) = \bar{\rho}_1(P,Q)$$
$$= \left(\inf_{\mu \in \mathcal{P}_1} \int_{\mathcal{X} \times \mathcal{X}} \rho(x,y) d\mu(x,y) \right)^{1/p}$$

(see, e.g., [26, Sec. 2]), and $\rho(x,y) \le d_{\max}^p 1_{\{x \neq y\}}$, so

$$\inf_{\mu \in \mathcal{P}_1} \int_{\mathcal{X} \times \mathcal{X}} \rho(x,y) d\mu(x,y)$$
$$\le d_{\max}^p \inf_{\mu \in \mathcal{P}_1} \int_{\mathcal{X} \times \mathcal{X}} 1_{\{x \neq y\}} d\mu(x,y).$$

The right-hand side of this expression is the well-known coupling characterization of twice the variational distance $d_V(P,Q)$ (see, e.g., Lindvall [50, Sec. I.5]), so we obtain

$$D_P(C^n)^{1/p} \le D_Q(C^n)^{1/p} + 2^{1/p} d_{\max} d_V(P,Q).$$

Interchanging the roles of P and Q, we obtain (C8).

To prove (C9), let C_*^n achieve the nth-order optimum for P: $D_P(C_*^n) = \hat{D}_P^n(R)$. Without loss of generality, we can assume that C_*^n has a nearest-neighbor encoder. Then

$$\hat{D}_Q^n(R)^{1/p} \le D_Q(C_*^n)^{1/p}$$
$$\le D_P(C_*^n)^{1/p} + 2^{1/p} d_{\max} d_V(P,Q)$$
$$= \hat{D}_P^n(R)^{1/p} + 2^{1/p} d_{\max} d_V(P,Q).$$

The other direction is proved similarly. □

APPENDIX D
PROOF OF LEMMA 4.1

Fix a $\theta \in \Theta$. Define the measurable set

$$\mathcal{U} \triangleq \{(x^n, z^n) \in \mathcal{X}^n \times \mathcal{X}^n : C^{n,n}(x^n, z^n) = a_*^n\}.$$

Then the distortion $D_\theta(C^{n,n})$ can be split into two terms as

$$D_\theta(C^{n,n}) = \frac{1}{n} \int_{\mathcal{X}^n \times \mathcal{X}^n} \rho(x^n, C^{n,n}(x^n, z^n)) dP_\theta(x^n, z^n)$$
$$= \frac{1}{n} \int_{\mathcal{U}} \rho(x^n, C^{n,n}(x^n, z^n)) dP_\theta(x^n, z^n)$$
$$+ \frac{1}{n} \int_{\mathcal{U}^c} \rho(x^n, C^{n,n}(x^n, z^n)) dP_\theta(x^n, z^n) \quad \text{(D1)}$$

where the superscript c denotes set-theoretic complement. We shall prove the lemma by upper-bounding separately each of the two terms on the right-hand side of (D1).

First of all, we have

$$\frac{1}{n} \int_{\mathcal{U}} \rho_M(x^n, \bar{C}^{n,n}(x^n, z^n)) dP_\theta(x^n, z^n)$$
$$\le \frac{1}{n} \int_{\mathcal{X}^n \times \mathcal{X}^n} \rho_M(x^n, \bar{C}^{n,n}(x^n, z^n)) dP_\theta(x^n, z^n)$$
$$\equiv \bar{D}_\theta(\bar{C}^{m,n}). \quad \text{(D2)}$$

By construction of $\mathcal{U}$ and $C^{n,n}$, $(x^n, z^n) \in \mathcal{U}$ implies that at least δn components of $\hat{\bar{x}}^n \equiv \bar{C}^{n,n}(x^n, z^n)$ satisfy

$\rho(x_i, \hat{\bar{x}}_i) > M$, so by definition of ρ_M it follows that $\rho_M(x^n, \bar{C}^{n,n}(x^n, z^n)) \ge n\delta M$ for all $(x^n, z^n) \in \mathcal{U}$. Thus

$$\frac{1}{n} \int_{\mathcal{U}} \rho_M(x^n, \bar{C}^{n,n}(x^n, z^n)) dP_\theta(x^n, z^n) \ge \delta M \cdot P_\theta^n \times P_\theta^n(\mathcal{U})$$

which, together with (D2), implies that

$$P_\theta^n \times P_\theta^n(\mathcal{U}) \le \frac{\bar{D}_\theta(\bar{C}^{n,n})}{\delta M}. \quad \text{(D3)}$$

Using the Cauchy–Schwarz inequality, (18), and (D3), we can write

$$\frac{1}{n} \int_{\mathcal{U}} \rho(x^n, C^{n,n}(x^n, z^n)) dP_\theta(x^n, z^n)$$
$$= \frac{1}{n} \mathbb{E}_\theta [\rho(X^n, a_*^n) \cdot 1_{\mathcal{U}}]$$
$$\le \sqrt{P_\theta^n \times P_\theta^n(\mathcal{U}) \mathbb{E}_\theta[\rho^2(X^n, a_*^n)/n^2]}$$
$$\le \sqrt{\frac{2G\bar{D}_\theta(\bar{C}^{n,n})}{\delta M}} \quad \text{(D4)}$$

where the last inequality follows from the easily established fact that, for any n independent random variables $V_1, \ldots, V_n$ satisfying $\mathbb{E}[V_i] \le G, \mathbb{E}[(\frac{1}{n}\sum_{i=1}^n V_i)^2] \le 2G$.

Now, $(x^n, z^n) \in \mathcal{U}^c$ implies that for each $i = 1, \ldots, n$, either $\hat{x}_i = \hat{\bar{x}}_i$ and $\rho_M(x_i, \hat{x}_i) = \rho(x_i, \hat{x}_i)$, or $\hat{x}_i = a_*$ and $\rho(x_i, \hat{\bar{x}}_i) > M$, where $\hat{\bar{x}}^n \equiv (\hat{\bar{x}}_1, \ldots, \hat{\bar{x}}_n) = \bar{C}^{n,n}(x^n, z^n)$ and $\hat{x}^n \equiv (\hat{x}_1, \ldots, \hat{x}_n) = C^{n,n}(x^n, z^n)$. Then, by the union bound

$$\frac{1}{n} \int_{\mathcal{U}^c} \rho(x^n, C^{n,n}(x^n, z^n)) dP_\theta(x^n, z^n)$$
$$= \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{U}^c} \rho(x_i, \hat{x}_i) dP_\theta(x^n, z^n)$$
$$\le \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{U}^c} \rho_M(x_i, \hat{\bar{x}}_i) dP_\theta(x^n, z^n)$$
$$+ \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{K}_i} \rho(x_i, a_*) dP_\theta(x^n, z^n) \quad \text{(D5)}$$

where, for each i, $\mathcal{K}_i \triangleq \{(x^n, z^n) : \rho(x_i, \hat{\bar{x}}_i) > M\}$. The first term on the right-hand side of (D5) is bounded as

$$\frac{1}{n} \sum_{i=1}^n \int_{\mathcal{U}^c} \rho_M(x_i, \hat{\bar{x}}_i) dP_\theta(x^n, z^n) \le \bar{D}_\theta(\bar{C}^{n,n}). \quad \text{(D6)}$$

As for the second term, we can once again invoke the Cauchy–Schwarz inequality and (18) to write

$$\frac{1}{n} \sum_{i=1}^n \int_{\mathcal{K}_i} \rho(x_i, a_*) dP_\theta(x^n, z^n)$$
$$\le \frac{1}{n} \sum_{i=1}^n \sqrt{P_\theta(\mathcal{K}_i) \mathbb{E}_\theta[\rho^2(X, a_*)]}$$
$$\le \frac{\sqrt{G}}{n} \sum_{i=1}^n \sqrt{P_\theta(\mathcal{K}_i)}. \quad \text{(D7)}$$

Let us estimate the summation on the right-hand side of (D7). First of all, note that

$$\frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{K}_i} \rho_M(x_i, \hat{\bar{x}}_i) dP_\theta(x^n, z^n) \leq \bar{D}_\theta(\bar{C}^{n,n}). \qquad \text{(D8)}$$

Now, $\rho(x_i, \hat{\bar{x}}_i) > M$ implies that $\rho_M(x_i, \hat{\bar{x}}_i) = M$, so that

$$\int_{\mathcal{K}_i} \rho_M(x_i, \hat{\bar{x}}_i) dP_\theta(x^n, z^n) = M P_\theta(\mathcal{K}_i)$$

which, together with (D8), yields the estimate

$$\sum_{i=1}^{n} P_\theta(\mathcal{K}_i) \leq \frac{n \bar{D}_\theta(\bar{C}^{n,n})}{M}$$

whence, by the concavity of the square root, it follows that

$$\sum_{i=1}^{n} \sqrt{P_\theta(\mathcal{K}_i)} \leq n \sqrt{\frac{\bar{D}_\theta(\bar{C}^{n,n})}{M}}.$$

Substituting this bound into (D7) yields

$$\frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{K}_i} \rho(x_i, a_*) dP_\theta(x^n, z^n) \leq \sqrt{\frac{G \bar{D}_\theta(\bar{C}^{n,n})}{M}}. \qquad \text{(D9)}$$

The lemma is proved by combining (D1), (D4), (D6), and (D9).

ACKNOWLEDGMENT

REFERENCES

[1] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inf. Theory*, vol. IT-30, no. 4, pp. 629–636, Jul. 1984.

[2] J. Rissanen, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, no. 3, pp. 1080–1100, Sept. 1986.

[3] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 40–47, Jan. 1996.

[4] A. Barron, J. Rissanen, and B. Yu, "Minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2743–2760, Oct. 1998.

[5] P. D. Grünwald, *The Minimum Description Length Principle*. Cambridge, MA: MIT Press, 2007.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[7] G. Tao, *Adaptive Control Design and Analysis*. Hoboken, NJ: Wiley, 2003.

[8] J. Ziv, "Coding of sources with unknown statistics—Part II: Distortion relative to a fidelity criterion," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 3, pp. 389–394, May 1972.

[9] D. L. Neuhoff, R. M. Gray, and L. D. Davisson, "Fixed rate universal block source coding with a fidelity criterion," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 5, pp. 511–523, Sep. 1975.

[10] T. Linder, G. Lugosi, and K. Zeger, "Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1728–1740, Nov. 1994.

[11] T. Linder, G. Lugosi, and K. Zeger, "Fixed-rate universal lossy source coding coding and rates of convergence for memoryless sources," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 665–676, May 1995.

[12] Z. Zhang and V. K. Wei, "An on-line universal lossy data compression algorithm via continuous codebook refinement—Part I: Basic results," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 803–821, May 1996.

[13] P. A. Chou, M. Effros, and R. M. Gray, "A vector quantization approach to universal noiseless coding and quantization," *IEEE Trans. Inf. Theory*, vol. 42, no. 4, pp. 1109–1138, Jul. 1996.

[14] R. F. Rice and J. R. Plaunt, "The Rice machine: Television data compression," Jet Propulsion Lab., Pasadena, CA, Tech. Rep. 900-408, Sep. 1970.

[15] R. F. Rice and J. R. Plaunt, "Adaptive variable-length coding for efficient compression of spacecraft television data," *IEEE Trans. Commun.*, vol. COM-19, no. 12, pp. 889–897, Dec. 1971.

[16] L. Devroye and L. Györfi, "Distribution and density estimation," in *Principles of Nonparametric Learning*, L. Györfi, Ed. New York: Springer-Verlag, 2001.

[17] L. Devroye and G. Lugosi, "A universally acceptable smoothing factor for kernel density estimation," *Ann. Statist.*, vol. 24, pp. 2499–2512, 1996.

[18] L. Devroye and G. Lugosi, "Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes," *Ann. Statist.*, vol. 25, pp. 2626–2637, 1997.

[19] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. New York: Springer-Verlag, 2001.

[20] D. L. Neuhoff and R. K. Gilbert, "Causal source codes," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 5, pp. 701–713, Sep. 1982.

[21] R. M. Gray, D. L. Neuhoff, and J. K. Omura, "Process definitions of distortion-rate functions and source coding theorems," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 5, pp. 524–532, Sep. 1975.

[22] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[23] I. Csiszár, "On an extremum problem in information theory," *Stud. Sci. Math. Hung.*, vol. 9, pp. 57–70, 1974.

[24] Y. G. Yatracos, "Rates of convergence of minimum distance estimates and Kolmogorov's entropy," *Ann. Math. Statist.*, vol. 13, pp. 768–774, 1985.

[25] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Sources*. Budapest, Hungary: Akadémiai Kiadó, 1981.

[26] R. M. Gray, D. L. Neuhoff, and P. S. Shields, "A generalization of Ornstein's $\bar{d}$ distance with applications to information theory," *Ann. Probab.*, vol. 3, no. 2, pp. 315–328, 1975.

[27] A. N. Kolmogorov and V. M. Tihomirov, "ϵ-entropy and ϵ-capacity of sets in function spaces," *Amer. Math. Soc. Transl.*, ser. 2, vol. 17, pp. 277–364, 1961.

[28] S. Amari and H. Nagaoka, *Methods of Information Geometry*. Providence, RI: Amer. Math. Soc., 2000.

[29] R. García-Muñoz and D. L. Neuhoff, "Strong universal source coding subject to a rate-distortion constraint," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 285–295, Mar. 1982.

[30] R. L. Dobrushin, "Unified methods for optimal quantization of messages," in *Problemy Kibernetiki* (in Russian), A. A. Lyapunov, Ed. Moscow, U.S.S.R.: Nauka, 1970, vol. 22, pp. 107–156.

[31] D. J. Sakrison, "The rate distortion function for a class of sources," *Inf. Contr.*, vol. 15, pp. 165–195, 1969.

[32] D. J. Sakrison, "Worst sources and robust codes for difference distortion measures," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 3, pp. 301–309, May 1975.

[33] D. L. Neuhoff and R. García-Muñoz, "Robust source coding of weakly compact classes," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 4, pp. 522–530, Jul. 1987.

[34] H. W. Sorenson and D. L. Alspach, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, vol. 7, no. 4, pp. 465–479, Jul. 1971.

[35] A. R. Barron and C.-H. Sheu, "Approximation of density functions by sequences of exponential families," *Ann. Statist.*, vol. 19, no. 3, pp. 1347–1369, 1991.

[36] H. Zhuang, Y. Huang, K. Palaniappan, and Y. Zhao, "Gaussian mixture density modeling, decomposition, and applications," *IEEE Trans. Image Process.*, vol. 5, no. 9, pp. 1293–1302, Sep. 1996.

[37] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 3, pp. 381–396, Mar. 2002.

[38] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.

[39] I. Kontoyiannis and J. Zhang, "Arbitrary source models and Bayesian codebooks in rate-distortion theory," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2276–2290, Aug. 2002.

[40] E.-H. Yang and Z. Zhang, "The redundancy of source coding with a fidelity criterion—Part II: Coding at a fixed rate level with unknown statistics," *IEEE Trans. Inf. Theory*, vol. 47, no. 1, pp. 126–145, Jan. 2002.

[41] M. Madiman and I. Kontoyiannis, Second-order properties of lossy likelihoods and the MLE/MDL dichotomy in lossy compression Brown Univ., Providence, RI, APPTS Rep. 04-5, May 2004 [Online]. Available: http://www.dam.brown.edu/ptg/REPORTS/04-5.pdf

[42] M. Madiman, M. Harrison, and I. Kontoyiannis, "Minimum description length vs. maximum likelihood in lossy data compression," in *Proc. IEEE Int. Symp. Information Theory*, Chicago, IL, Jun./Jul. 2004, p. 461.

[43] M. Raginsky, "Joint universal lossy coding and identification of stationary mixing sources with general alphabets," *IEEE Trans. Inf. Theory*, submitted for publication.

[44] M. Raginsky, "Joint universal lossy coding and identification of stationary mixing sources," in *Proc. IEEE Int. Symp. Information Theory*, Nice, France, Jun. 2007, pp. 1961–1965.

[45] P. Moulin and R. Koetter, "Data-hiding codes," *Proc. IEEE*, vol. 93, no. 12, pp. 2085–2127, Dec. 2005.

[46] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Appl.*, vol. 16, pp. 264–280, 1971.

[47] N. Sauer, "On the density of families of sets," *J. Combin. Theory Ser. A*, vol. 13, pp. 145–147, 1972.

[48] R. M. Dudley, "Central limit theorems for empirical measures," *Ann. Probab.*, vol. 6, pp. 898–929, 1978.

[49] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Soc.*, vol. 58, pp. 13–30, 1963.

[50] T. Lindvall, *Lectures on the Coupling Method*. New York: Dover, 2002.