

# Joint Universal Lossy Coding and Identification of I.I.D. Vector Sources

Maxim Raginsky

Beckman Institute and the University of Illinois  
405 N Mathews Ave, Urbana, IL 61801, USA  
Email: maxim@uiuc.edu

**Abstract**—The problem of joint universal source coding and modeling, addressed by Rissanen in the context of lossless codes, is generalized to fixed-rate lossy coding of continuous-alphabet memoryless sources. We show that, for bounded distortion measures, any compactly parametrized family of i.i.d. real vector sources with absolutely continuous marginals (satisfying appropriate smoothness and Vapnik–Chervonenkis learnability conditions) admits a joint scheme for universal lossy block coding and parameter estimation, and give nonasymptotic estimates of convergence rates for distortion redundancies and variational distances between the active source and the estimated source. We also present explicit examples of parametric sources admitting such joint universal compression and modeling schemes.

## I. INTRODUCTION

In universal data compression, a single code achieves asymptotically optimal performance on all sources within a given family. Intuition suggests that a good universal coder should acquire an accurate model of the source statistics from a sufficiently long data sequence and incorporate this knowledge in its operation. For lossless codes, this intuition has been made rigorous by Rissanen [1]. Under his scheme, the data are encoded in a *two-stage* set-up, in which the binary representation of each source block consists of two parts: (1) a suitably quantized maximum-likelihood estimate of the source parameters, and (2) lossless encoding of the data matched to the acquired model; the redundancy of the resulting code converges to zero as  $\log n/n$ , where  $n$  is the block length.

In this paper, we extend Rissanen’s idea to *lossy* block coding (vector quantization) of i.i.d. sources with values in  $\mathbb{R}^d$  for some finite  $d$ . Specifically, let  $\{X_i\}_{i=-\infty}^{\infty}$  be an i.i.d. source with the marginal distribution of  $X_1$  belonging to some indexed class  $\{P_\theta : \theta \in \Theta\}$  of absolutely continuous distributions on  $\mathbb{R}^d$ , where  $\Theta$  is a bounded subset of  $\mathbb{R}^k$  for some  $k$ . For bounded distortion measures, our main result, Theorem 1, states that if the class  $\{P_\theta\}$  satisfies certain smoothness and learnability conditions, then there exists a sequence of finite-memory lossy block codes that achieves asymptotically optimal compression of each source in the class and permits asymptotically exact identification of the active source with respect to the *variational distance*, defined as  $d_V(P, Q) \triangleq \sup_B |P(B) - Q(B)|$ , where the supremum is over all Borel subsets of  $\mathbb{R}^d$ . The overhead rate and the distortion redundancy of the scheme converge to zero as  $O(\log n/n)$  and  $O(\sqrt{\log n/n})$ , respectively, where  $n$  is the block length, while the active source can be identified

up to a variational ball of radius  $O(\sqrt{\log n/n})$  eventually almost surely. We also describe an extension of our scheme to unbounded distortion measures satisfying a certain moment condition, and present two examples of parametric families satisfying the regularity conditions of Theorem 1.

While most existing schemes for universal lossy coding rely on *implicit* identification of the active source (e.g., through topological covering arguments [2], Glivenko–Cantelli uniform laws of large numbers [3], or nearest-neighbor code clustering [4]), our code builds an *explicit model* of the mechanism responsible for generating the data and then selects an appropriate code for the data on the basis of the model. This ability to simultaneously model and compress the data may prove useful in such applications as *media forensics* [5], where the parameter  $\theta$  could represent evidence of tampering, and the aim is to compress the data in such a way that the evidence can be later extracted with high fidelity from the compressed version. Another key feature of our approach is the use of Vapnik–Chervonenkis theory [6] in order to connect universal encodability of a class of sources to the combinatorial “richness” of a certain collection of decision regions associated with the sources. In a way, Vapnik–Chervonenkis estimates can be thought of as an (imperfect) analogue of the combinatorial method of types for finite alphabets [7].

## II. PRELIMINARIES

Let  $\{X_i\}_{i=-\infty}^{\infty}$  be an i.i.d. source with alphabet  $\mathcal{X}$ , such that the marginal distribution of  $X_1$  comes from an indexed class  $\{P_\theta : \theta \in \Theta\}$ . For any  $t \in \mathbb{Z}$  and any  $m, n \geq 0$ , let  $X_m^n(t)$  denote the segment  $(X_{tn-m+1}, X_{tn-m+2}, \dots, X_{tn})$  of  $\{X_i\}$ , with  $X_0^n(t)$  understood to denote an empty string for all  $n, t$ . We shall abbreviate  $X_n^n(t)$  to  $X^n(t)$ .

Consider coding  $\{X_i\}$  into the reproduction process  $\{\widehat{X}_i\}$  with alphabet  $\widehat{\mathcal{X}}$  by means of a stationary lossy code with block length  $n$  and memory length  $m$  [an  $(n, m)$ -block code, for brevity]. Such a code consists of an encoder  $f : \mathcal{X}^n \times \mathcal{X}^m \rightarrow \mathcal{S}$  and a decoder  $\phi : \mathcal{S} \rightarrow \widehat{\mathcal{X}}^n$ , where  $\mathcal{S}$  is a collection of fixed-length binary strings:  $\widehat{X}^n(t) = \phi(f(X^n(t), X_m^n(t-1)))$ ,  $\forall t \in \mathbb{Z}$ . That is, the encoding is done in blocks of length  $n$ , but the encoder is also allowed to observe a fixed finite amount of past data. Abusing notation, we shall denote by  $C^{n,m}$  both the composition  $\phi \circ f$  and the encoder-decoder pair  $(f, \phi)$ ; when  $m = 0$ , we shall use a more compact notation  $C^n$ . The number  $R(C^{n,m}) \triangleq n^{-1} \log |\mathcal{S}|$  is the rate of  $C^{n,m}$  in bits

per letter. The distortion between the source  $n$ -block  $X^n = (X_1, \dots, X_n)$  and its reproduction  $\hat{X}^n = (\hat{X}_1, \dots, \hat{X}_n)$  is given by  $\rho(X^n, \hat{X}^n) = \sum_{i=1}^n \rho(X_i, \hat{X}_i)$ , where  $\rho: \mathcal{X} \times \hat{\mathcal{X}} \rightarrow \mathbb{R}^+$  is a single-letter distortion measure.

Suppose  $X_1 \sim P_\theta$  for some  $\theta \in \Theta$ . Since the source is i.i.d., and the code  $C^{n,m}$  does not vary with time, the process  $\{(X_i, \hat{X}_i)\}$  is  $n$ -stationary, and the average distortion of  $C^{n,m}$  is  $D_\theta(C^{n,m}) = n^{-1} \mathbb{E}_\theta [\rho(X^n(1), \hat{X}^n(1))]$ , where  $\hat{X}^n(1) \equiv \phi(f(X^n(1), X_m^n(0)))$ . The optimal performance achievable on  $P_\theta$  by any finite-memory  $n$ -block code at rate  $R$  is given by the  $n$ th-order operational distortion-rate function (DRF)

$$\hat{D}_\theta^{n,*}(R) \triangleq \inf_{m \geq 0} \inf_{C^{n,m}: R(C^{n,m}) \leq R} D_\theta(C^{n,m}),$$

where the asterisk denotes the fact that we allow any finite memory length. For zero-memory  $n$ -block codes the corresponding DRF is

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

Clearly,  $\hat{D}_\theta^{n,*}(R) \leq \hat{D}_\theta^n(R)$ . Conversely, we can use memory-less minimum-distortion encoders to convert any  $(n, m)$ -block code into a zero-memory  $n$ -block code without increasing either distortion or rate, so  $\hat{D}_\theta^n(R) = \hat{D}_\theta^{n,*}(R)$ . Finally, the best performance achievable by any block code, with or without memory, is given by the operational DRF  $\hat{D}_\theta(R) \triangleq \inf_{n \geq 1} \hat{D}_\theta^n(R) = \lim_{n \rightarrow \infty} \hat{D}_\theta^n(R)$ ; since the source is i.i.d.,  $\hat{D}_\theta(R)$  is equal to the Shannon DRF  $D_\theta(R)$  by the source coding theorem and its converse.

We are interested in sequences of codes that asymptotically achieve optimal performance across the entire class  $\{P_\theta\}$ . Let  $\{C^{n,m}\}_{n=1}^\infty$  be a sequence of  $(n, m)$ -block codes, where the memory length  $m$  may depend on  $n$ , such that  $R(C^{n,m}) \rightarrow R$ . Then  $\{C^{n,m}\}$  is *weakly minimax universal* for  $\{P_\theta\}$  if the *distortion redundancy*  $\delta_\theta(C^{n,m}) \triangleq D_\theta(C^{n,m}) - D_\theta(R)$  converges to zero as  $n \rightarrow \infty$  for every  $\theta \in \Theta$ .<sup>1</sup> We shall follow Chou *et al.* [4] and split  $\delta_\theta(C^{n,m})$  into two terms:

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

The first term, which we shall call the  $n$ th-order redundancy and denote by  $\delta_\theta^n(C^{n,m})$ , is the excess distortion of  $C^{n,m}$  relative to the best  $n$ -block code for  $P_\theta$ , while the second term gives the extent to which the best  $n$ -block code falls short of the Shannon optimum. Note that  $\delta_\theta(C^{n,m}) \rightarrow 0$  if and only if  $\delta_\theta^n(C^{n,m}) \rightarrow 0$ , since  $\hat{D}_\theta^n(R) \rightarrow D_\theta(R)$  by the source coding theorem.

### III. INFORMAL DESCRIPTION OF THE SYSTEM

As stated in the Introduction, we are after a sequence of lossy block codes that would not only be universally optimal for a given class  $\{P_\theta\}$  of i.i.d. sources with values in  $\mathbb{R}^d$ , but would also permit asymptotically reliable identification of the source parameter  $\theta \in \Theta$ . We formally state and prove our result in Section IV; here we outline the main idea behind it.

<sup>1</sup>See [2] for other notions of universality for source codes.

Fix the block length  $n$ , and denote by  $X^n$  the current  $n$ -block  $X^n(t)$  and by  $Z^n$  the preceding  $n$ -block  $X^n(t-1)$ . Let us assume that we can find for each  $\theta \in \Theta$  an  $n$ -block code  $C_\theta^n$  at the desired rate  $R$  which achieves the  $n$ th-order DRF for  $P_\theta$ :  $D_\theta(C_\theta^n) = \hat{D}_\theta^n(R)$ . The basic idea is to construct an  $(n, n)$ -block code  $C^{n,n}$  that first estimates the parameter  $\theta$  of the active source from  $Z^n$  and then codes  $X^n$  with the code  $C_{\hat{\theta}(Z^n)}^n$ , where  $\hat{\theta}(Z^n)$  is the estimate of  $\theta$ , suitably quantized.

Suppose the encoder can use  $Z^n$  to identify the active source up to a variational ball of radius  $O(\sqrt{\log n/n})$ . Next, suppose that the parameters of the estimated source (assumed to belong to a bounded subset of  $\mathbb{R}^k$  for some  $k$ ) are quantized to  $O(\log n)$  bits in such a way that the variational distance between any two sources whose parameters lie in the same quantizer cell is  $O(\sqrt{1/n})$ . If  $\hat{\theta} = \hat{\theta}(Z^n)$  is the quantized parameter estimate, then the variational distance between  $P_{\hat{\theta}}$  and the “true” source  $P_\theta$  is  $O(\sqrt{\log n/n})$ , which for bounded distortion functions implies an  $O(\sqrt{\log n/n})$  upper bound on the distortion redundancy  $\delta_\theta^n(C_{\hat{\theta}}^n)$ .

More formally, let  $\tilde{f}: \mathcal{X}^n \rightarrow \tilde{\mathcal{S}}$  denote the map that sends each  $Z^n$  to the binary representation of  $\hat{\theta}(Z^n)$ , where  $\tilde{\mathcal{S}}$  is a collection of fixed-length binary strings with  $\log |\tilde{\mathcal{S}}| = O(\log n)$ , and let  $\tilde{\psi}: \tilde{\mathcal{S}} \rightarrow \Theta$  be the parameter decoder that maps each  $\tilde{s} \in \tilde{\mathcal{S}}$  to its reproduction:  $\hat{\theta}(Z^n) = \tilde{\psi}(\tilde{f}(Z^n))$ . Thus, to each  $\tilde{s} \in \tilde{\mathcal{S}}$  there corresponds an  $n$ -block code  $C_{\tilde{s}}^n$ , which we denote more compactly by  $C_{\tilde{s}}^n = (f_{\tilde{s}}, \phi_{\tilde{s}})$ . Our  $(n, n)$ -block code  $C^{n,n}$  thus has the encoder  $f(X^n, Z^n) \triangleq \tilde{f}(Z^n) f_{\tilde{f}(Z^n)}(X^n)$  and the corresponding decoder  $\phi(\tilde{f}(Z^n) f_{\tilde{f}(Z^n)}(X^n)) \triangleq \phi_{\tilde{f}(Z^n)}(f_{\tilde{f}(Z^n)}(X^n))$ . That is, the binary string emitted by the encoder consists of two parts: (a) the *header* containing a binary description of the chosen code [equivalently, of the estimated source  $P_{\hat{\theta}(Z^n)}$ ], and (b) the *body* containing the binary description of the data  $X^n$  using the chosen code at rate  $R$ . The combined rate of  $C^{n,n}$  is  $R + n^{-1} \log |\tilde{\mathcal{S}}| = R + O(\log n/n)$  bits per letter, while the expected distortion with respect to  $P_\theta$  is

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

This scheme is universal because the map  $\tilde{f}$  and the subcodes  $C_{\tilde{s}}^n$  are chosen so that  $\mathbb{E}_\theta [D_\theta(C_{\tilde{f}(Z^n)}^n)] - \hat{D}_\theta^n(R) = O(\sqrt{\log n/n})$  for each  $\theta \in \Theta$ . Note that decoder can not only decode the data in a near-optimal fashion, but also identify the active source up to a variational ball of radius  $O(\sqrt{\log n/n})$ .

We remark that our scheme is a modification of the two-stage code of Chou *et al.* [4], the difference being that here the subcode  $C_{\tilde{s}}^n$ , used to encode the current  $n$ -block  $X^n$ , is selected on the basis of the preceding  $n$ -block  $Z^n$ . Nonetheless, we shall adopt the terminology of [4] and refer to  $\tilde{f}$  as the *first-stage encoder*. The structure of the encoder and the decoder in our scheme is displayed in Fig. 1.

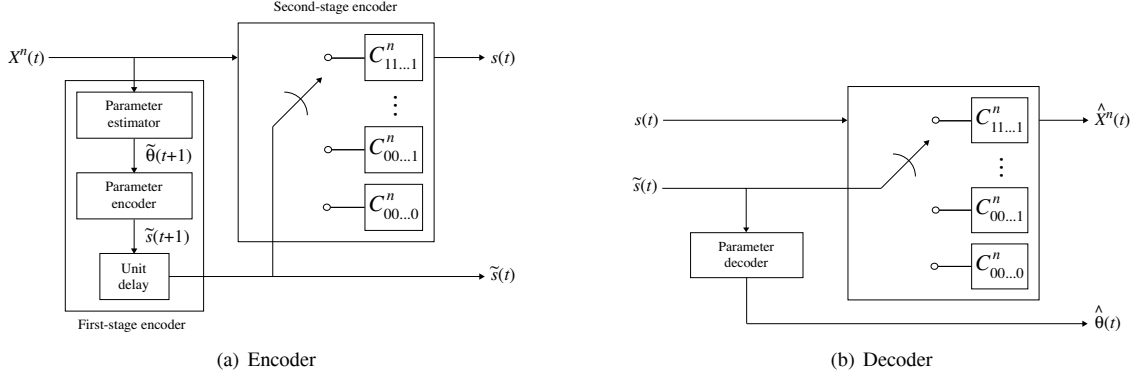


Fig. 1. The two-stage scheme for joint universal lossy source coding and identification.

#### IV. THE MAIN THEOREM

Before stating and proving our result, let us list our assumptions, as well as fix some auxiliary results and notation.

**The source models.** Let  $\{X_i\}_{i=-\infty}^{\infty}$  be an i.i.d. source with alphabet  $\mathcal{X}$  and with the marginal distribution of  $X_1$  belonging to an indexed class  $\{P_\theta : \theta \in \Theta\}$ , such that the following conditions are satisfied:

- (S.1)  $\mathcal{X}$  is a measurable subset of  $\mathbb{R}^d$ .
- (S.2) Each  $P_\theta$  is absolutely continuous with pdf  $p_\theta$ .

**Distortion function.** The distortion function  $\rho$  is assumed to satisfy the following requirements:

- (D.1)  $\inf_{\hat{x} \in \tilde{\mathcal{X}}} \rho(x, \hat{x}) = 0$  for all  $x \in \mathcal{X}$ .
- (D.2)  $\sup_{x \in \mathcal{X}, \hat{x} \in \tilde{\mathcal{X}}} \rho(x, \hat{x}) = K < \infty$ .

Under these conditions, it can be proved (in a manner similar to the proof of Thm. 2 of Linder *et al.* [3]) that

$$\hat{D}_\theta^n(R) = D_\theta(R) + O(\sqrt{\log n/n}) \quad (2)$$

for each  $\theta \in \Theta$  and for all rates  $R$  such that  $D_\theta(R) > 0$  (this condition is automatically satisfied in our case since all the  $P_\theta$ 's are absolutely continuous). The constant implicit in  $O(\cdot)$  depends on both  $\theta$  and  $R$ .

**Vapnik-Chervonenkis theory.** Given a collection  $\mathcal{A}$  of measurable subsets of  $\mathbb{R}^d$ , its *Vapnik-Chervonenkis (VC) dimension*  $V(\mathcal{A})$  is defined as the largest integer  $n$  for which

$$\max_{x^n \in (\mathbb{R}^d)^n} |\{(1_{\{x_1 \in A\}}, \dots, 1_{\{x_n \in A\}}) : A \in \mathcal{A}\}| = 2^n; \quad (3)$$

if (3) holds for all  $n$ , then  $V(\mathcal{A}) = \infty$ . If  $V(\mathcal{A}) < \infty$ , we say that  $\mathcal{A}$  is a VC class. For any such class, one can give finite-sample bounds on uniform deviations of probabilities of events in that class from their relative frequencies. That is, if  $X^n = (X_1, \dots, X_n)$  is an i.i.d. sample from a distribution  $P$ , and if  $\mathcal{A}$  is a VC class with  $V(\mathcal{A}) \geq 2$ , then

$$\Pr \left\{ \sup_{A \in \mathcal{A}} |P_{X^n}(A) - P(A)| > \epsilon \right\} \leq 8n^{V(\mathcal{A})} e^{-n\epsilon^2/32}, \forall \epsilon > 0$$

and

$$\mathbb{E} \left\{ \sup_{A \in \mathcal{A}} |P_{X^n}(A) - P(A)| \right\} \leq c\sqrt{\log n/n},$$

where  $P_{X^n}$  is the empirical distribution of  $X^n$  and the constant  $c$  depends on  $V(\mathcal{A})$  but not on  $P$ .<sup>2</sup> (See, e.g., [6] for details.)

**Theorem 1.** Let  $\{X_i\}_{i=-\infty}^{\infty}$  be an i.i.d. source satisfying Conditions (S.1) and (S.2), and let  $\rho$  be a distortion function satisfying Conditions (D.1) and (D.2). Assume the following:

- 1)  $\Theta$  is a bounded subset of  $\mathbb{R}^k$  for some  $k$ .
- 2) The map  $\theta \mapsto P_\theta$  is *uniformly locally Lipschitz*: there exist constants  $r, \beta > 0$  such that, for each  $\theta \in \Theta$ ,  $d_V(P_\theta, P_\eta) \leq \beta \|\theta - \eta\|$  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  $\theta$ .
- 3) The collection  $\mathcal{A}_\Theta$  of all sets of the form  $A_{\theta, \eta} = \{x \in \mathcal{X} : p_\theta(x) > p_\eta(x)\}$  with  $\theta \neq \eta$  (the so-called *Yatracos class* associated with  $\{P_\theta\}$  [8], [9], [10]) is a VC class.

Also, suppose that for each  $n, \theta$  there exists an  $n$ -block code  $C_\theta^n = (f_\theta, \phi_\theta)$  at rate of  $R$  bits per letter achieving the  $n$ th-order operational DRF for  $\theta$ :  $D_\theta(C_\theta^n) = \hat{D}_\theta^n(R)$ . Then there exists an  $(n, n)$ -block code  $C^{n,n}$  with

$$R(C^{n,n}) = R + O(\log n/n), \quad (4)$$

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

$$\delta_\theta(C^{n,n}) = O(\sqrt{\log n/n}). \quad (5)$$

Therefore, the sequence of codes  $\{C^{n,n}\}_{n=1}^{\infty}$  is 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

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

The constants implicit in the  $O(\cdot)$  notation in (4) and (6) are independent of  $\theta$ .

*Proof:* The proof is by construction of a two-stage  $(n, n)$ -block code as outlined in Sec. III. As before, let  $X^n$  denote the current  $n$ -block  $X^n(t)$ , and let  $Z^n$  be the preceding  $n$ -block  $X^n(t-1)$ . We first define our first-stage encoder  $\tilde{f}$ , which we shall realize as the composition  $\tilde{g} \circ \tilde{\theta}$  of a parameter estimator  $\tilde{\theta}$

<sup>2</sup>Using more refined techniques, the  $c\sqrt{\log n/n}$  bound can be improved to  $c'/\sqrt{n}$ , where  $c'$  is another constant, but  $c'$  is much larger than  $c$ , so any benefit of the new bound shows only for "impractically" large values of  $n$ .

and a lossy parameter encoder  $\tilde{g}$  (cf. Fig. 1). For any  $z^n \in \mathcal{X}^n$  and for any  $\theta \in \Theta$ , let  $\Delta_\theta(z^n) \triangleq \sup_{A \in \mathcal{A}_\Theta} |P_\theta(A) - P_{z^n}(A)|$ ,

where  $P_\theta(A) \equiv \int_A p_\theta(x) dx$ . Define  $\hat{\theta}(z^n)$  as any  $\theta^* \in \Theta$  such that  $\Delta_{\theta^*}(z^n) < \inf_{\theta \in \Theta} \Delta_\theta(z^n) + 1/n$ , where the extra  $1/n$  ensures that at least one such  $\theta^*$  exists. The map  $z^n \mapsto \hat{\theta}(z^n)$  is the so-called *minimum-distance density estimator* of Devroye and Lugosi [9], [10], which satisfies

$$d_V(P_\theta, P_{\hat{\theta}(z^n)}) \leq 2\Delta_\theta(Z^n) + 3/2n. \quad (7)$$

Since  $\mathcal{A}_\Theta$  is a VC class,

$$\mathbb{E}_\theta[d_V(P_\theta, P_{\hat{\theta}(Z^n)})] \leq c\sqrt{\log n/n} + 3/2n, \quad (8)$$

for each  $\theta \in \Theta$ , where  $c > 0$  depends on  $V(\mathcal{A}_\Theta)$ .

Next, we construct the lossy encoder  $\tilde{g}$ . Since  $\Theta$  is bounded, it is contained in some cube  $M$  of side  $J \in \mathbb{N}$ . Let  $\{M_1^{(n)}, M_2^{(n)}, \dots, M_K^{(n)}\}$  be a partitioning of  $M$  into contiguous cubes of side  $1/\lceil n^{1/2} \rceil$ , so that  $K \leq (Jn^{1/2})^k$ . Represent each  $M_j^{(n)}$  that intersects  $\Theta$  by a unique fixed-length binary string  $\tilde{s}_j$ , and let  $\tilde{S} = \{\tilde{s}_j\}$ . 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 can be described by a string of no more than  $k(\log n^{1/2} + \log J)$  bits. For each  $M_j^{(n)}$  that intersects  $\Theta$ , choose a reproduction  $\hat{\theta}_j \in M_j^{(n)} \cap \Theta$  and designate  $C_{\hat{\theta}_j}^n$  as the corresponding  $n$ -block code  $C_{\tilde{s}_j}^n$ . The parameter decoder  $\tilde{\psi} : \tilde{S} \rightarrow \Theta$  is then given by  $\tilde{\psi}(\tilde{s}_j) = \hat{\theta}_j$ .

The rate of the resulting  $(n, n)$ -block code  $C^{n,n}$  does not exceed  $R + n^{-1}k(\log n^{1/2} + \log J)$  bits per letter, which proves (4). By (1), the average distortion of  $C^{n,n}$  on the source  $P_\theta$  is given by  $D_\theta(C^{n,n}) = \mathbb{E}_\theta[D_\theta(C_{\hat{\theta}}^n)]$ , where  $\hat{\theta} = \hat{\theta}(Z^n) \equiv \tilde{\psi}(\tilde{f}(Z^n))$ . From standard quantizer mismatch arguments and the triangle inequality,

$$\delta_\theta^n(C_{\hat{\theta}}^n) \leq 4K[d_V(P_\theta, P_{\hat{\theta}}) + d_V(P_{\hat{\theta}}, P_{\hat{\theta}})].$$

Taking expectations, we get

$$\delta_\theta^n(C^{n,n}) \leq 4K\{\mathbb{E}_\theta[d_V(P_\theta, P_{\hat{\theta}})] + \mathbb{E}_\theta[d_V(P_{\hat{\theta}}, P_{\hat{\theta}})]\}. \quad (9)$$

We now estimate separately each term in the curly brackets in (9). Using (8), we can bound the first term by

$$\mathbb{E}_\theta[d_V(P_\theta, P_{\hat{\theta}})] \leq c\sqrt{\log n/n} + 3/2n. \quad (10)$$

The second term involves  $\tilde{\theta} = \tilde{\theta}(Z^n)$  and its quantized version  $\hat{\theta}$ , where  $\|\tilde{\theta} - \hat{\theta}\| \leq \sqrt{k/n}$  by construction of  $\tilde{g}, \tilde{\psi}$ . Using the assumption that the map  $\theta \mapsto P_\theta$  is uniformly locally Lipschitz, as well as the fact that  $d_V(P, Q) \leq 1$  for any two distributions  $P, Q$ , it is not hard to show that there exists a constant  $\beta'$  such that

$$d_V(P_{\tilde{\theta}}, P_{\hat{\theta}}) \leq \beta'\sqrt{k/n} \quad (11)$$

and consequently that

$$\mathbb{E}_\theta[d_V(P_{\tilde{\theta}}, P_{\hat{\theta}})] \leq \beta'\sqrt{k/n}. \quad (12)$$

Substituting the bounds (10) and (12) into (9) yields

$$\delta_\theta^n(C^{n,n}) \leq K\left(4c\sqrt{\log n/n} + 6/n + 4\beta'\sqrt{k/n}\right),$$

whence it follows that the  $n$ th-order redundancy  $\delta_\theta^n(C^{n,n}) = O(\sqrt{\log n/n})$  for every  $\theta \in \Theta$ . Then the decomposition

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

and (2) imply that (5) holds for every  $\theta \in \Theta$ .

To prove (6), fix an  $\epsilon > 0$  and note that by (7), (11) and the triangle inequality,  $d_V(P_\theta, P_{\hat{\theta}}) > \epsilon$  implies that  $2\Delta_\theta(Z^n) + 3/2n + \beta'\sqrt{k/n} > \epsilon$ . Hence,

$$\Pr\{d_V(P_\theta, P_{\hat{\theta}}) > \epsilon\} \leq \Pr\{\Delta_\theta(Z^n) > (\epsilon - \gamma\sqrt{1/n})/2\},$$

where  $\gamma = 3/2 + \beta'\sqrt{k}$ . Since  $\mathcal{A}_\Theta$  is a VC class,

$$\Pr\{d_V(P_\theta, P_{\hat{\theta}}) > \epsilon\} \leq 8n^{V(\mathcal{A}_\Theta)} e^{-n(\epsilon - \gamma\sqrt{1/n})^2/128}. \quad (13)$$

If for each  $n$  we choose  $\epsilon_n > \sqrt{128V(\mathcal{A}_\Theta) \ln n/n} + \gamma\sqrt{1/n}$ , then the right-hand side of (13) will be summable in  $n$ , hence  $d_V(P_\theta, P_{\hat{\theta}(Z^n)}) = O(\sqrt{\log n/n})$   $P_\theta$ -a.s. by the Borel-Cantelli lemma. ■

The above proof combines techniques of Rissanen [1] (namely, explicit identification of the source parameters) with the parameter-space quantization idea of Chou *et al.* [4]. The VC condition on the Yatracos class  $\mathcal{A}_\Theta$  is needed to control the  $L_1$  convergence rate of the density estimators, which bounds the convergence rate of the distortion redundancies. We remark also that the boundedness condition on the distortion function can be relaxed in favor of a uniform moment condition with respect to a reference letter, but at the expense of a quadratic slowdown of the rate at which the distortion redundancy converges to zero (the proof is omitted for lack of space):

*Theorem 2.* Let  $\{P_\theta : \theta \in \Theta\}$  be a family of i.i.d. sources satisfying the conditions of Theorem 1, and let  $\rho$  be a distortion function for which there exists a *reference letter*  $a_* \in \hat{\mathcal{X}}$  such that  $\sup_{\theta \in \Theta} \mathbb{E}_\theta[\rho(X, a_*)^2] < \infty$ , and which satisfies Condition (D.1). Then for any rate  $R > 0$  satisfying  $\sup_{\theta \in \Theta} D_\theta(R) < \infty$  there exists a sequence  $\{C^{n,n}\}_{n=1}^\infty$  of  $(n, n)$ -block codes with  $R(C^{n,n}) = R + O(\log n/n)$  and  $\delta_\theta(C^{n,n}) = O(\sqrt[4]{\log n/n})$  for every  $\theta \in \Theta$ . The source identification performance is the same as in Theorem 1.

## V. EXAMPLES

Here, we present two explicit examples of parametric families satisfying the conditions of Theorem 1 and thus admitting joint universal lossy coding and identification schemes.

**Mixture classes.** Let  $p_1, \dots, p_k$  be fixed pdf's over a measurable domain  $\mathcal{X} \subseteq \mathbb{R}^d$ , and let  $\Theta$  be the simplex of probability distributions on  $\{1, \dots, k\}$ . Then the mixture class defined by the  $p_i$ 's consists of all densities of the form  $p_\theta(x) = \sum_{i=1}^k \theta_i p_i(x)$ ,  $\theta = (\theta_1, \dots, \theta_k) \in \Theta$ . The parameter space  $\Theta$  is compact and thus satisfies Condition 1) of Theorem 1. In order to show that Condition 2) holds, fix any  $\theta, \eta \in \Theta$ . Then

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

where the last inequality follows by concavity of the square root. Therefore, the map  $\theta \mapsto P_\theta$  is everywhere Lipschitz with Lipschitz constant  $\sqrt{k}/2$ . It remains to show that the Yatracos class  $\mathcal{A}_\Theta$  is VC. To this end, observe that  $x \in A_{\theta, \eta}$  if and only if  $\sum_i (\theta_i - \eta_i) p_i(x) > 0$ . Thus  $\mathcal{A}_\Theta$  consists of sets of the form  $\{x \in \mathcal{X} : \sum_i \alpha_i p_i(x) > 0, (\alpha_1, \dots, \alpha_k) \in \mathbb{R}^k\}$ . Since the functions  $p_1, \dots, p_k$  span a linear space of dimension not larger than  $k$ , Lemma 4.2 in [6] guarantees that  $V(\mathcal{A}_\Theta) \leq k$ .

**Exponential families.** Let  $\mathcal{X}$  be a measurable subset of  $\mathbb{R}^d$ , and let  $\Theta$  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* [11] if there exist a probability density  $p$  and  $k$  real-valued functions  $h_1, \dots, h_k$  on  $\mathcal{X}$ , such that each  $p_\theta$  has the form  $p_\theta(x) = p(x)e^{\theta \cdot h(x) - g(\theta)}$ , where  $h(x) \triangleq (h_1(x), \dots, h_k(x))$ ,  $\theta \cdot h(x) \triangleq \sum_{i=1}^k \theta_i h_i(x)$ , and  $g(\theta) = \ln \int_{\mathcal{X}} e^{\theta \cdot h(x)} p(x) dx$  is the normalization constant. Given the densities  $p$  and  $p_\theta$ , let  $P$  and  $P_\theta$  denote the corresponding distributions. By the compactness of  $\Theta$ , Condition 1) of Theorem 1 is satisfied. Next we demonstrate that Conditions 2) and 3) can also be met under certain regularity assumptions.

Namely, suppose that  $\{1, h_1, \dots, h_k\}$  is a linearly independent set of functions. This guarantees that the map  $\theta \mapsto P_\theta$  is one-to-one. We also assume that each  $h_i$  is square-integrable with respect to  $P$ :  $\int_{\mathcal{X}} h_i^2 dP < \infty$ ,  $1 \leq i \leq k$ . Then the  $(k+1)$ -dimensional real linear space  $\mathcal{F} \subset L^2(\mathcal{X}, P)$  spanned by  $\{1, h_1, \dots, h_k\}$  can be equipped with an inner product  $\langle f, g \rangle \triangleq \int_{\mathcal{X}} f g dP$  and the corresponding  $L_2$  norm  $\|f\|_2 \triangleq \sqrt{\langle f, f \rangle} \equiv \sqrt{\int_{\mathcal{X}} f^2 dP}$ . Also let  $\|f\|_\infty \triangleq \inf \{M : |f(x)| \leq M \text{ } P\text{-a.e.}\}$  denote the  $L_\infty$  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 Radon-Nikodym derivatives  $dP/dP_\theta \equiv p/p_\theta$  are uniformly bounded  $P$ -a.e.:  $\sup_{\theta \in \Theta} \|\ln(p/p_\theta)\|_\infty = L < \infty$ . These conditions are satisfied, for example, by truncated Gaussian densities over a compact domain in  $\mathbb{R}^d$  with suitably bounded means and covariance matrices.

Let  $D(P_\theta \| P_\eta)$  denote the relative entropy (information divergence) between  $P_\theta$  and  $P_\eta$ . With the above conditions in place, we can prove the following result along the lines of Lemma 4 of Barron and Sheu [11]:

$$D(P_\theta \| P_\eta) \leq \frac{1}{2} e^{\|\ln(p/p_\theta)\|_\infty} e^{2A_k \|\theta - \eta\|} \|\theta - \eta\|^2, \quad (14)$$

where  $\|\cdot\|$  is the Euclidean norm on  $\mathbb{R}^k$ . From Pinsker's inequality  $d_V(P_\theta, P_\eta) \leq \sqrt{D(P_\theta \| P_\eta)}/2$  [12, Lemma 5.2.8], (14) and the uniform boundedness of  $\ln p/p_\theta$ , we get

$$d_V(P_\theta, P_\eta) \leq \beta_0 e^{A_k \|\theta - \eta\|} \|\theta - \eta\|, \quad \theta, \eta \in \Theta, \quad (15)$$

where  $\beta_0 \triangleq e^{L/2}/2$ . If we fix  $\theta \in \Theta$ , then from (15) it follows that for any  $r > 0$ ,  $d_V(P_\theta, P_\eta) \leq \beta_0 e^{A_k r} \|\theta - \eta\|$  for all  $\eta$  satisfying  $\|\eta - \theta\| \leq r$ . That is, the family  $\{P_\theta : \theta \in \Theta\}$  satisfies the uniform local Lipschitz condition of Theorem 1, and the magnitude of the Lipschitz constant can be controlled by tuning  $r$ .

All we have left to show is that the Yatracos class  $\mathcal{A}_\Theta$  is a VC class. Since  $p_\theta(x) > p_\eta(x)$  if and only if  $(\theta - \eta) \cdot h(x) > g(\theta) - g(\eta)$ ,  $\mathcal{A}_\Theta$  consists of sets of the form

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

Since the functions  $1, h_1, \dots, h_k$  span a  $(k+1)$ -dimensional linear space, the same argument as that used for mixture classes shows that  $V(\mathcal{A}_\Theta) \leq k+1$ .

## VI. CONCLUSION AND FUTURE WORK

We have presented a constructive proof of the existence of a scheme for joint universal lossy block coding and identification of real i.i.d. vector sources with parametric marginal distributions satisfying certain regularity conditions. Our main motivation was to show that the connection between universal coding and source identification, exhibited by Rissanen for lossless coding of discrete-alphabet sources [1], carries over to the domain of lossy codes and continuous alphabets. As far as future work is concerned, it would be of both theoretical and practical interest to extend the approach described here to variable-rate codes (thus lifting the boundedness requirement for the parameter space) and to general (not necessarily memoryless) stationary sources.

## ACKNOWLEDGMENT

The author would like to thank Pierre Moulin and Ioannis Kontoyiannis for useful discussions. This work was supported by the Beckman Institute Postdoctoral Fellowship.

## REFERENCES

- [1] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 4, pp. 629–636, July 1984.
- [2] D. L. Neuhoff, R. M. Gray, and L. D. Davison, "Fixed rate universal block source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 5, pp. 511–523, September 1975.
- [3] 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. Inform. Theory*, vol. 40, no. 6, pp. 1728–1740, November 1994.
- [4] P. A. Chou, M. Effros, and R. M. Gray, "A vector quantization approach to universal noiseless coding and quantization," *IEEE Trans. Inform. Theory*, vol. 42, no. 4, pp. 1109–1138, July 1996.
- [5] P. Moulin and R. Koetter, "Data-hiding codes," *Proc. IEEE*, vol. 93, no. 12, pp. 2085–2127, December 2005.
- [6] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. New York: Springer-Verlag, 2001.
- [7] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Sources*. Budapest: Akadémiai Kiadó, 1981.
- [8] Y. G. Yatracos, "Rates of convergence of minimum distance estimates and Kolmogorov's entropy," *Ann. Math. Statist.*, vol. 13, pp. 768–774, 1985.
- [9] L. Devroye and G. Lugosi, "A universally acceptable smoothing factor for kernel density estimation," *Ann. Statist.*, vol. 24, pp. 2499–2512, 1996.
- [10] —, "Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes," *Ann. Statist.*, vol. 25, pp. 2626–2637, 1997.
- [11] 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.
- [12] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.