Joint Universal Lossy Coding and Identification of Stationary Mixing 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, treated in the context of lossless codes by Rissanen, was recently generalized to fixed-rate lossy coding of finitely parametrized continuous-alphabet i.i.d. sources. We extend these results to variable-rate lossy block coding of stationary ergodic sources and show that, for bounded metric distortion measures, any finitely parametrized family of stationary sources satisfying suitable mixing, smoothness and Vapnik–Chervonenkis learnability conditions admits universal schemes for joint lossy source coding and identification. We also give several explicit examples of parametric sources satisfying the regularity conditions.**

I. Introduction

A universal source coding scheme is one that performs asymptotically optimally for all sources within a given class. Intuition suggests that a good universal coder should acquire a probabilistic model of the source from a sufficiently long data sequence and operate based on this model. For lossless codes, this intuition has been made rigorous by Rissanen [1]: the data are encoded via a two-part code which comprises (1) a suitably quantized maximum-likelihood estimate of the source parameters, and (2) an encoding of the data with the code optimized for the acquired model. The redundancy of this scheme converges to zero as $k \log n/n$, where n is the block length and k is the dimension of the parameter space.

Recently we have extended Rissanen's ideas to *lossy* block coding of finitely parametrized continuous-alphabet i.i.d. sources with bounded parameter spaces [2], [3]. We have shown that, under appropriate regularity conditions, there exist joint universal schemes for lossy coding and source identification whose distortion redundancy and source estimation fidelity both converge to zero as $O\left(\sqrt{\log n/n}\right)$ as the block length n tends to infinity. The code operates by coding each block with the code matched to the parameters estimated from the preceding block. Moreover, the constant hidden in the $O(\cdot)$ notation increases with the "richness" of the model class, as measured by the Vapnik–Chervonenkis (VC) dimension [4], [5] of a certain class of decision regions in the source alphabet.

The main limitation of the results of [2], [3] is the i.i.d. assumption, which excludes such practically relevant model classes as autoregressive sources or Markov and hidden Markov processes. Furthermore, the assumption of a bounded parameter space may not be always justified. In this paper we relax both of these assumptions. Because the parameter space is not bounded, we have to use variable-rate codes with countably infinite codebooks, whose performance is naturally quantified by Lagrangians [6], [7]. We show that, under certain regularity conditions, there are universal schemes for joint lossy source coding and modeling such that, as the block length n tends to infinity, both the Lagrangian redundancy relative to the best variable-rate code at each block length and the source estimation fidelity at the decoder converge to zero as $O(\sqrt{V_n \log n/n})$, where V_n is the VC dimension of a certain class of decision regions induced by the collection of all n-dimensional marginals of the source process distributions.

The key novel feature of our scheme is that, unlike most existing schemes for universal lossy coding, which rely on implicit identification of the active source, it learns an explicit probabilistic model. Moreover, our results clearly show that the "price of universality" of a modeling-based compression scheme grows with the combinatorial richness of the underlying model class, as captured by the VC dimension sequence $\{V_n\}$. The richer the model class, the harder it is to learn, which in turn affects the compression performance because we use the source parameters learned from past data in deciding how to encode the current block. These insights may prove useful in such settings as digital forensics or adaptive control under communication constraints, where trade-offs between the quality of parameter estimation and compression performance are of central importance.

II. Preliminaries

Let $\boldsymbol{X} = \{X_i\}_{i\in\mathbb{Z}}$ be a stationary, ergodic source with alphabet $\mathcal{X}$. All alphabets are assumed to be Polish spaces equipped with their Borel σ-fields. We adopt the usual setting of universal source coding: the process distribution of $\boldsymbol{X}$ is not known exactly, apart from being a member of some indexed class $\{P_\theta : \theta \in \Lambda\}$. We assume that the parameter space Λ is an open subset of $\mathbb{R}^k$ with nonempty interior. We also assume that there exists a σ-finite measure μ on $\mathcal{X}$, such that for every $\theta \in \Lambda$ the n-dimensional marginals P_θ^n of P_θ are absolutely continuous with respect to (w.r.t.) the product measure μ^n, for all n, denoting the corresponding densities $dP_\theta^n/d\mu^n$ by p_θ^n.

We wish to code $\boldsymbol{X}$ into a reproduction process $\widehat{\boldsymbol{X}} = \{\widehat{X}_i\}_{i\in\mathbb{Z}}$ with alphabet $\widehat{\mathcal{X}}$ by means of a finite-memory variable-rate lossy block code (vector quantizer). Such a code with block length n and memory length m [an (n,m)-block code, for short] is a pair $C^{n,m} = (f, \varphi)$, where $f : \mathcal{X}^n \times \mathcal{X}^m \to \mathcal{S}$ is the encoder, $\varphi : \mathcal{S} \to \widehat{\mathcal{X}}^n$ is the decoder, and $\mathcal{S} \subseteq \{0,1\}^*$ is a finite or countable collection of binary strings

1-4244-1429-6/07/$25.00 ©2007 IEEE

satisfying the prefix condition. The mapping of $\boldsymbol{X}$ into $\widehat{\boldsymbol{X}}$ is defined by $\widehat{X}_{nk+1}^{n(k+1)} = \varphi(f(X_{nk+1}^{n(k+1)}, X_{nk-m+1}^{nk}))$, $k \in \mathbb{Z}$, where $X_i^j \triangleq (X_i, X_{i+1}, \ldots, X_j)$, $i < j$. Thus, the encoding is done in blocks of length n, but the encoder is also allowed to view the m source symbols immediately preceding the current n-block. Abusing notation, we shall denote by $C^{n,m}$ both the composition $\varphi \circ f$ and the pair (f, φ); when $m = 0$, we shall use a more compact notation C^n and say "n-block code."

Let $\rho : \mathcal{X} \times \widehat{\mathcal{X}} \to \mathbb{R}^+$ be a measurable single-letter distortion function; $\rho_n(x^n, \widehat{x}^n) = n^{-1} \sum_{i=1}^n \rho(x_i, \widehat{x}_i)$ is the per-letter distortion due to reproducing $x^n \in \mathcal{X}^n$ by $\widehat{x}^n \in \widehat{\mathcal{X}}^n$. We assume that ρ is a metric on $\mathcal{X} \cup \widehat{\mathcal{X}}$, bounded from above by some $\rho_{\max} < \infty$. Suppose $\boldsymbol{X} \sim P_\theta$. Associated with the code $C^{n,m}$ are its expected distortion $D_\theta(C^{n,m}) \triangleq \mathbb{E}_\theta\{\rho_n(X_1^n, \widehat{X}_1^n)\}$ and its expected rate $R_\theta(C^{n,m}) \triangleq \mathbb{E}_\theta\{\ell_n(f(X_1^n, X_{-m+1}^0))\}$, where, for a binary string s, $\ell_n(s)$ is its length in bits, normalized by n. When working with variable-rate quantizers, it is convenient [6], [7] to absorb the distortion and the rate into a single performance measure, the *Lagrangian* $L_\theta(C^{n,m}, \lambda) \triangleq D_\theta(C^{n,m}) + \lambda R_\theta(C^{n,m})$, where $\lambda > 0$ is the *Lagrange multiplier* which controls the distortion-rate trade-off. The optimal Lagrangian performance achievable on P_θ by any zero-memory variable-rate quantizer with block length n is given by the *nth-order operational distortion-rate Lagrangian* $\widehat{L}_\theta^n(\lambda) \triangleq \inf_{C^n} L_\theta(C^n, \lambda)$ [6]. Allowing the codes to have nonzero memory does not improve optimal performance, because we can use memoryless nearest-neighbor encoders to convert any (n,m)-block code into an n-block code without increasing the Lagrangian. Thus, $\widehat{L}_\theta^n(\lambda) = \inf_m \inf_{C^{n,m}} L_\theta(C^{n,m}, \lambda)$, where the infimum is over all memory lengths m and all (n,m)-block codes $C^{n,m}$, for a fixed block length n. Because each P_θ is ergodic, $\widehat{L}_\theta^n(\lambda)$ converges, as $n \to \infty$, to the *distortion-rate Lagrangian* $L_\theta(\lambda) \triangleq \min_R \Big(D_\theta(R) + \lambda R\Big)$, where $D_\theta(R)$ is the Shannon distortion-rate function of P_θ [6].

III. The results

In this section we state our result on universal schemes for joint lossy compression and identification of stationary sources satisfying certain regularity conditions. We wish to design a sequence of variable-rate vector quantizers, such that the decoder can reliably reconstruct the source sequence $\boldsymbol{X}$ and reliably identify the active source in an asymptotically optimal manner for all $\theta \in \Lambda$. The identification performance will be judged in terms of the variational distance, which for any two probability measures P, Q on a measurable space $(\mathcal{Z}, \mathcal{A})$ is defined by $d(P,Q) \triangleq 2 \sup_{A \in \mathcal{A}} |P(A) - Q(A)|$. Denoting by p and q the respective densities of P and Q w.r.t. a dominating measure ν, we can also write $d(P,Q) = \int_{\mathcal{X}} |p(z) - q(z)| d\nu(z)$. The set of all Q satisfying $d(P,Q) \le \delta$ for a given P is called the *variational ball of radius* δ around P.

Our first condition ensures that each source in the class is sufficiently close to an i.i.d. source, in an asymptotic sense. Define the kth β-mixing coefficient of P_θ [5] by

$$\beta_\theta(k) \triangleq 2 \sup_{A \in \sigma(X_{-\infty}^0, X_k^\infty)} |P_\theta(A) - P_\theta^- \times P_\theta^+(A)|,$$

where $\sigma(X_{-\infty}^0, X_k^\infty)$ is the σ-field generated by $\{X_i\}_{i \le 0}$ and $\{X_i\}_{i \ge k}$, and P_θ^- and P_θ^+ are the marginal distributions of $\{X_i\}_{i \le 0}$ and $\{X_i\}_{i > 0}$, respectively. An i.i.d. source has $\beta(k) \equiv 0, \forall k$; if $\beta(k) \stackrel{k \to \infty}{\longrightarrow} 0$, the source is called *$\beta$-mixing*.

Condition 1. The sources in $\{P_\theta : \theta \in \Lambda\}$ are *algebraically β-mixing*:

$$\exists r > 0 \text{ such that } \beta_\theta(k) = O(k^{-r}), \forall \theta \in \Lambda.$$

The second condition ensures that the parametrization of the sources is sufficiently smooth.

Condition 2. Let $d_n(\theta, \theta')$ denote the variational distance between P_θ^n and $P_{\theta'}^n$. Then for every $\theta \in \Lambda$,

$$\exists \delta_\theta, c_\theta > 0 \text{ such that } \sup_n \frac{d_n(\theta, \theta')}{\sqrt{n}} \le c_\theta \|\theta - \theta'\|$$

for all θ' satisfying $\|\theta' - \theta\| < \delta_\theta$, where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^k$.

This condition is met, for instance, if the asymptotic Fisher information matrix $I(\theta)$ exists for all $\theta \in \Lambda$ (under some technical assumptions on the densities p_θ^n). It guarantees that, for every sequence $\{\delta_n\}_{n \in \mathbb{N}}$ of positive reals satisfying $\delta_n \to 0, \sqrt{n}\delta_n \to 0$ as $n \to \infty$, and for every sequence $\{\theta_n\}_{n \in \mathbb{N}}$ in Λ satisfying $\|\theta_n - \theta\| < \delta_n$ for a given $\theta \in \Lambda$, we have $d_n(\theta_n, \theta) \to 0$ as $n \to \infty$.

Finally, we impose a learnability condition. To state it we need some facts on Vapnik–Chervonenkis classes (see, e.g., [4], [5]). Let $(\mathcal{Z}, \mathcal{A})$ be a measurable space. Given a collection $\mathcal{C}$ of measurable subsets of $\mathcal{Z}$, its *Vapnik-Chervonenkis (VC) dimension* $\mathsf{V}(\mathcal{C})$ is defined as the largest integer n for which

$$\max_{x^n \in \mathcal{X}^n} |\{(1_{\{x_1 \in A\}}, \cdots, 1_{\{x_n \in A\}}) : A \in \mathcal{C}\}| = 2^n; \quad (1)$$

if (1) holds for all n, then $\mathsf{V}(\mathcal{C}) = \infty$. If $\mathsf{V}(\mathcal{C}) < \infty$, we say that $\mathcal{C}$ is a VC class. The Vapnik–Chervonenkis inequalities are finite-sample bounds on uniform deviations of probabilities of events in a VC class from their relative frequencies: if $X^n = (X_1, \cdots, X_n)$ is an i.i.d. sample from a probability measure P on $(\mathcal{Z}, \mathcal{A})$, and if $\mathcal{C}$ is a VC class with $\mathsf{V}(\mathcal{C}) \ge 2$, then

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

$$\text{and } \mathbb{E}\left\{\sup_{A \in \mathcal{C}} |P_{X^n}(A) - P(A)|\right\} \le c\sqrt{\mathsf{V}(\mathcal{C}) \log n / n},$$

where $c > 0$ is a universal constant[1], P_{X^n} is the empirical distribution of X^n, and the probabilities and expectations are w.r.t. the product measure P^n on $(\mathcal{Z}^n, \mathcal{A}^n)$.

Condition 3. For $n \in \mathbb{N}$, let $\mathcal{A}_n$ consist of all sets of the form

$$A_{\theta,\theta'} = \{x^n \in \mathcal{X}^n : p_\theta(x^n) > p_{\theta'}(x^n)\}, \ \theta \ne \theta'$$

($\mathcal{A}_n$ is the so-called *Yatracos class* defined by $\{p_\theta^n\}$, see [4] and references therein). Then we require that each $\mathcal{A}_n$ is a VC class, $V_n \equiv \mathsf{V}(\mathcal{A}_n) < \infty$, and that $V_n = o(n/\log n)$.

Theorem 1. Suppose Conditions 1–3 are satisfied. Then for every $\lambda, \eta > 0$ there exists a sequence $\{C_*^{n,m_n}\}_{n \in \mathbb{N}}$ of

[1] Using more refined techniques, the $c\sqrt{\mathsf{V}(\mathcal{C}) \log n / n}$ bound can be improved to $c'\sqrt{\mathsf{V}(\mathcal{C})/n}$, where c' is another constant. However, c' is much larger than c, so any benefit of the new bound shows only for "impractically" large values of n.

variable-rate vector quantizers with memory lengths $m_n = n(n + \lceil n^{(2+\eta)/r} \rceil)$, such that

$$L_\theta(C_*^{n,m_n}, \lambda) - \inf_m \inf_{C^{n,m}} L_\theta(C^{n,m}, \lambda) = O\left(\sqrt{\frac{V_n \log n}{n}}\right)$$

for all $\theta \in \Lambda$. Moreover, for each n, the binary description produced by the encoder is such that the decoder can identify the n-dimensional marginal of the active source up to a variational ball of radius $O(\sqrt{V_n \log n/n})$ almost surely.

That is, for each n, θ the code C_*^{n,m_n}, which is *independent of* θ, performs almost as well as the best finite-memory quantizer with block length n that can be designed with full knowledge of P_θ^n. Thus, as far as compression goes, our scheme can compete with all finite-memory variable-rate quantizers, with the additional bonus of allowing the decoder to identify the active source in an asymptotically optimal manner. Recalling the discussion of Lagrangian optimality in Section II, we see that Theorem 1 immediately implies the following:

Corollary 2. The sequence $\{C_*^{n,m_n}\}_{n\in\mathbb{N}}$ is *weakly minimax universal*[2] for $\{P_\theta : \theta \in \Lambda\}$, i.e., for every $\theta \in \Lambda$, $L_{\theta_0}(C_*^{n,m_n}, \lambda) \to L_\theta(\lambda)$ as $n \to \infty$.

IV. The proof of Theorem 1

The main idea. It suffices to construct a universal scheme that can compete with all *zero-memory* codes; that is, we need to show that there exists a sequence $\{C_*^{n,m_n}\}$ of codes, such that $L_\theta(C_*^{n,m_n}, \lambda) - \widehat{L}_\theta^n(\lambda) = O(\sqrt{V_n \log n/n})$ for all $\theta \in \Lambda$.

We assume throughout that the "true" source is P_{θ_0} for some $\theta_0 \in \Lambda$. Our code operates as follows. Suppose that both the encoder and the decoder have access to a countably infinite "database" $\boldsymbol{c} = \{\theta(i)\}_{i\in\mathbb{N}} \subset \Lambda$. Using Elias' universal representation of the integers [8], we can associate to each $\theta(i)$ a unique binary string $s(i)$ with $\ell(s(i)) = \log i + O(\log\log i)$ bits. Suppose also that for each n, θ there exists a zero-memory n-block code $C_\theta^n = (f_\theta, \varphi_\theta)$ that achieves the nth-order Lagrangian optimum for P_θ: $L_\theta(C_\theta^n, \lambda) = \widehat{L}_\theta^n(\lambda)$. The encoding of X_1^n into $\widehat{X}_1^n$ is done as follows:

1) The encoder estimates $P_{\theta_0}^n$ from the m_n-block $X_{-m_n+1}^0$ as $P_{\widetilde{\theta}}^n$, where $\widetilde{\theta} = \widetilde{\theta}(X_{-m_n+1}^0)$.
2) The encoder then computes the *waiting time*
$$T_n \triangleq \inf\left\{i \geq 1 : d_n\big(\theta(i), \widetilde{\theta}(X_{-m_n+1}^0)\big) \leq \sqrt{n}\delta_n\right\},$$
with the standard convention that the infimum of the empty set is equal to $+\infty$; $\{\delta_n\}$ is a sequence of positive reals to be specified later.
3) If $T_n < +\infty$, the encoder sets $\widehat{\theta} = \theta(T_n)$; otherwise, the encoder sets $\widehat{\theta} = \theta(1)$ (or some other default θ).
4) The description of X_1^n is a concatenation of three binary strings: (i) a 1-bit flag b to tell whether T_n is finite ($b = 0$) or infinite ($b = 1$); (ii) a binary string s_1 which is equal to $s(T_n)$ if $T_n < +\infty$ or is empty if $T_n = +\infty$; (iii) $s_2 = f_{\widehat{\theta}}(X_1^n)$. The string $\widetilde{s} = bs_1$ is the *first-stage description*, while s_2 is the *second-stage description*.

[2] See [6] for other notions of universality for lossy codes.

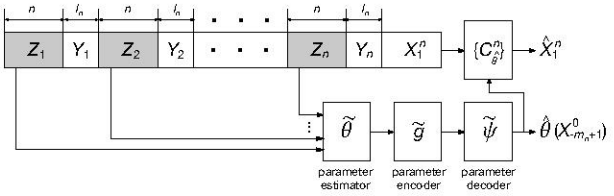

Fig. 1. The structure of the code C_*^{n,m_n}. The shaded blocks are those used for estimating the source parameters.

The decoder receives bs_1s_2, determines $\widehat{\theta}$ from $\widetilde{s}$, and produces $\widehat{X}_1^n = \varphi_{\widehat{\theta}}(s)$. If $b = 0$ (which, as we shall show, will happen eventually a.s.), then $P_{\widehat{\theta}}^n$ is in the variational ball of radius $\sqrt{n}\delta_n$ around the estimated $P_{\widetilde{\theta}}^n$. If the latter is a good estimate, i.e., $d_n(\theta_0, \widetilde{\theta}) \overset{\text{a.s.}}{\to} 0$ as $n \to \infty$, then the decoder's estimate of $P_{\theta_0}^n$ is only slightly worse. Moreover, the a.s. convergence of $d_n(\theta_0, \widehat{\theta})$ to zero as $n \to \infty$ implies that the performance of $C_{\widehat{\theta}}^n$ on P_{θ_0} is close to the optimum $L_{\theta_0}(C_{\theta_0}^n, \lambda) \equiv \widehat{L}_{\theta_0}^n(\lambda)$.

Formally, the code C_*^{n,m_n} is comprised by the following maps: (1) the *parameter estimator* $\widetilde{\theta} : \mathcal{X}^{m_n} \to \Lambda$; (2) the *parameter encoder* $\widetilde{g} : \Lambda \to \widetilde{\mathcal{S}}$, where $\widetilde{\mathcal{S}} = \{0s(i)\}_{i\in\mathbb{N}} \cup \{1\}$; (3) the *parameter decoder* $\widetilde{\psi} : \widetilde{\mathcal{S}} \to \Lambda$. Let $\widetilde{f}$ denote the composition $\widetilde{g} \circ \widetilde{\theta}$ of the parameter estimator and the parameter encoder, which we refer to as the *first-stage encoder*, and let $\widehat{\theta}$ denote the composition $\widetilde{\psi} \circ \widetilde{f}$ of the parameter decoder and the first-stage encoder. The decoder $\widetilde{\psi}$ is the *first-stage decoder*. The collection $\{C_\theta^n : \theta \in \Lambda\}$ defines the *second-stage codes*. The encoder $f_* : \mathcal{X}^n \times \mathcal{X}^{m_n} \to \widetilde{\mathcal{S}} \times \mathcal{S}$ and the decoder $\varphi_* : \widetilde{\mathcal{S}} \times \mathcal{S} \to \widehat{\mathcal{X}}^n$ of C_*^{n,m_n} are defined as $f_*(X_1^n, X_{-m_n+1}^0) = \widetilde{f}(X_{-m_n+1}^0) f_{\widehat{\theta}(X_{-m_n+1}^0)}(X_1^n)$ and $\varphi_*(\widetilde{s}s) = \varphi_{\widetilde{\psi}(\widetilde{s})}(s)$ for all $s \in \mathcal{S}, \widetilde{s} \in \widetilde{\mathcal{S}}$, respectively.

To assess the performance of the code, introduce the functions $g(x^n, y^{m_n}) = \rho_n(x^n, C_{\widehat{\theta}(y^{m_n})}^n(x^n)) + \lambda \ell_n(f_{\widehat{\theta}(y^{m_n})}(x^n))$ and $h(y^{m_n}) = \ell_n(\widetilde{f}(y^{m_n}))$. Then $h(X_{-m_n+1}^0)$ is the normalized length of the first-stage description, while $g(X_1^n, X_{-m_n+1}^0)$ is the instantaneous Lagrangian performance of the corresponding second-stage code. The expected Lagrangian performance of our code is

$$L_{\theta_0}(C_*^{n,m_n}, \lambda) = \mathbb{E}_{\theta_0}\, g(X_1^n, X_{-m_n+1}^0) + \lambda\, \mathbb{E}_{\theta_0}\, h(X_{-m_n+1}^0).$$

We prove the theorem by showing that, with proper choices for the memory length m_n, the "database" $\boldsymbol{c}$, the parameter estimator $\widetilde{\theta}$, and the sequence $\{\delta_n\}$, we can ensure that $\mathbb{E}_{\theta_0} h(X_{-m_n+1}^0) = O(k \log n/n) + O(\log\log n/n) + o(1)$, $\mathbb{E}_{\theta_0} g(X_1^n, X_{-m_n+1}^0) = \widehat{L}_\theta^n(\lambda) + O(\sqrt{V_n \log n/n})$, and $d_n(\theta_0, \widehat{\theta}(X_{-m_n+1}^0)) = O(\sqrt{V_n \log n/n})$ P_{θ_0}-almost surely.

Step 1: choice of memory length. Let $l_n = \lceil n^{(2+\eta)/r} \rceil$ and $m_n = n(n + l_n)$. Divide $X_{-m_n+1}^0$ into n blocks $Z_1, \ldots, Z_n$ of length n interleaved by n blocks $Y_1, \ldots, Y_n$ of length l_n (see Figure 1). The parameter estimator $\widetilde{\theta}$, although defined as acting on the entire $X_{-m_n+1}^0$, effectively will make use only of $Z^n = (Z_1, \ldots, Z_n)$. Each $Z_j \sim P_{\theta_0}^n$, but the Z_j's are not independent. Let $Q^{(n)}$ denote the marginal distribution of Z^n, and let $\widetilde{Q}^{(n)}$ denote the product of n copies of $P_{\theta_0}^n$. Using induction and the definition of the β-mixing coefficient, we can show that $d(Q^{(n)}, \widetilde{Q}^{(n)}) \leq (n-1)\beta_{\theta_0}(l_n) = O(1/n^{1+\eta})$,

which follows from Condition 1 and our choice of l_n. This "blocking technique" [9] allows us to approximate certain probabilities and expectations w.r.t. P_{θ_0} by probabilities and expectations w.r.t. suitably constructed i.i.d. processes.

Step 2: construction of the database. We proceed by random selection. Let W be some probability measure on Λ with a positive, everywhere continuous density $w(\theta)$. We generate $\boldsymbol{C} = \{\theta(i)\}_{i\in\mathbb{N}}$ as an i.i.d. sequence of vectors in Λ drawn according to W, independently of $\boldsymbol{X}$.

Step 3: estimation of the active source. We use the Devroye–Lugosi *minimum-distance estimator* (MDE) (see [4] and references therein). Namely, given the estimation blocks $Z^n = (Z_1,\ldots,Z_n)$, define $U_\theta(Z^n) \triangleq \sup_{A\in\mathcal{A}_n} |P^n_\theta(A) - P_{Z^n}(A)|$ for every $\theta\in\Lambda$, where the supremum is over all sets in the Yatracos class $\mathcal{A}_n$ and P_{Z^n} is the empirical distribution on $\mathcal{X}^n$ induced by Z^n. Then $\widetilde{\theta}(X^0_{-m_n+1})$ is any $\theta^*\in\Lambda$ satisfying $U_{\theta^*}(Z_1^n) < \inf_{\theta\in\Lambda} U_\theta(Z_1^n) + 1/n$ (the extra $1/n$ term ensures that at least one such θ^* exists). Note that $\widetilde{\theta}(X^0_{-m_n+1})$ only depends on Z^n. The key property of the MDE is [4]

$$d_n(\theta_0, \widetilde{\theta}(X^0_{-m_n+1})) \le 4U_{\theta_0}(Z_1^n) + 3/n, \qquad (2)$$

which holds regardless of whether Z^n is i.i.d. or not.

Step 4: expected first-stage description length. We follow the ideas of [10]. Let us assume that the sequence $\{\delta_n\}$ is such that $\delta_n \to 0$ as $n\to\infty$. Define the event $F_n = \{\theta\in\Lambda : d_n(\theta, \widetilde{\theta}(X^0_{-m_n+1})) \le \sqrt{n}\delta_n\}$ and note that if $q_n = W(F_n|X^0_{-m_n+1} = x^0_{-m_n+1}) > 0$, then the waiting time T_n is a geometric random variable with parameter q_n. Condition 2 ensures that, in fact, $q_n > 0$ for n sufficiently large, for P_{θ_0}-almost all realizations of $\boldsymbol{X}$. Using the Borel–Cantelli lemma, it is not hard to show that $\mathbb{E}_{\theta_0}\log T_n \le \log\log n + 2 - \mathbb{E}_{\theta_0}\log q_n$ for all realizations of $\boldsymbol{C}$, eventually P_{θ_0}-a.s. We now lower-bound q_n for large n. Using the triangle inequality, independence of $\boldsymbol{X}$ and $\boldsymbol{C}$, Condition 2 and the fact that $\delta_n\to 0$ as $n\to\infty$, we have, for n sufficiently large,

$$q_n \ge W\Big(\|\Theta-\theta_0\| \le \delta_n/2c_{\theta_0}\Big) P_{\theta_0}\Big(d_n(\theta_0,\widetilde{\theta}) \le \sqrt{n}\delta_n/2\Big),$$

where $\widetilde{\theta} = \widetilde{\theta}(X^0_{-m_n+1})$ and $\Theta \sim W$. Via simple volume bounding, $W\big(\|\Theta-\theta_0\| \le \delta_n/2c_{\theta_0}\big) \ge (1/2)w(\theta_0)v_k(\delta_n/2c_{\theta_0})^k$ for n sufficiently large, where v_k is the volume of the unit sphere in $\mathbb{R}^k$. Next, we use blocking to approximate P_{θ_0}-probabilities by $\widetilde{Q}^{(n)}$-probabilities, and then invoke the property (2) of the MDE and the Vapnik–Chervonenkis inequalities to obtain

$$P_{\theta_0}\Big(d_n(\theta_0,\widetilde{\theta}(X^0_{-m_n+1})) \le \sqrt{n}\delta_n/2\Big) \ge 1 - 8n^{\mathsf{V}(\mathcal{A}_n)} e^{-n(\sqrt{n}\delta_n - 6/n)^2/2048} - O(1/n^{1+\eta}).$$

Choosing $\delta_n = \frac{\sqrt{2048(V_n+1)\ln n}}{n} + \frac{6}{n^{3/2}}$, we get for the normalized expected first-stage description length[3]

$$\mathbb{E}_{\theta_0} h(X^0_{-m_n+1}) = O(k\log n/n) + O(\log\log n/n) + o(1).$$

The sequence δ_n indeed converges to 0 owing to Condition 3.

Step 5: expected second-stage Lagrangian performance. Using the fact that the distortion measure ρ is bounded, one can show via an argument similar to the proof of Lemma 9 in Section 10 of [7] that for every $\theta\in\Lambda$ there is no loss of generality in assuming that an n-block code $C^n_\theta = (f_\theta,\varphi_\theta)$ achieving $\widehat{L}^n_\theta(\lambda)$ satisfies $\ell_n(f_\theta(x^n)) \le 2\rho_{\max}/\lambda$ for all $x^n\in\mathcal{X}^n$. Thus, g is bounded by $3\rho_{\max}$. A straightforward application of Fubini's theorem and the definition of the β-mixing coefficient yields $\mathbb{E}_{\theta_0} g(X_1^n, X^0_{-m_n+1}) \le \mathbb{E}_{\theta_0} L_{\theta_0}(C^n_{\widehat{\theta}},\lambda) + O(1/n^{2+\eta})$, where $\widehat{\theta} = \widehat{\theta}(X^0_{-m_n+1})$. Thus, the Lagrangian performance of the second-stage code is determined by the behavior of the code $C^n_{\widehat{\theta}}$ (which depends on $X^0_{-m_n+1}$). Because ρ is a metric, a basic Lagrangian mismatch argument (see, e.g., Lemma 9 in Section 8 of [7]) shows that

$$\mathbb{E}_{\theta_0} L_{\theta_0}(C^n_{\widehat{\theta}},\lambda) \le \mathbb{E}_{\theta_0} L_{\theta_0}(C^n_{\theta_0},\lambda) + 4\rho_{\max}\,\mathbb{E}_{\theta_0} d_n(\theta_0,\widehat{\theta}).$$

By blocking, the expectation of $d_n(\theta_0,\widehat{\theta})$ w.r.t. P_{θ_0} can be approximated by expectation w.r.t. $\widetilde{Q}^{(n)}$. Followed by an application of the triangle inequality, this yields

$$\mathbb{E}_{\theta_0} d_n(\theta_0,\widehat{\theta}) \le \mathbb{E}_{\widetilde{Q}^{(n)}}\big\{d_n(\theta_0,\widetilde{\theta}) + d_n(\widetilde{\theta},\widehat{\theta})\big\} + O(1/n^{1+\eta}),$$

where $\widetilde{\theta} = \widetilde{\theta}(X^0_{-m_n+1})$ is the MD estimate of θ_0. Now, $d_n(\widetilde{\theta},\widehat{\theta}) \le \sqrt{n}\delta_n = O(\sqrt{V_n\log n/n})$ eventually almost surely, by construction of the first-stage encoder. The expectation $\mathbb{E}_{\widetilde{Q}^{(n)}} d_n(\theta_0,\widetilde{\theta})$ can be handled via (2) and the Vapnik–Chervonenkis inequalities, yielding

$$\mathbb{E}_{\theta_0} d_n(\theta_0,\widetilde{\theta}) = O\big(\sqrt{V_n\log n/n}\big) + O(1/n^{1+\eta}).$$

Thus, $\mathbb{E}_{\theta_0}\{g\} = \widehat{L}^n_{\theta_0}(\lambda) + O(\sqrt{V_n\log n/n}) + O(1/n^{1+\eta})$.

Step 6: the overall performance. Gathering together our estimates for the first stage and for the second stage, we get

$$L_{\theta_0}(C^{n,m_n}_*,\lambda) = \widehat{L}^n_{\theta_0}(\lambda) + O(\sqrt{V_n\log n/n}) + O(k\log n/n) + O(\log\log n/n) + o(1)$$

for almost every realization of the database $\boldsymbol{C}$. As for the performance of the scheme in identifying the active source, note that, with our choice of l_n, the sequence $n\beta_{\theta_0}(l_n)$ is summable in n. Then a straightforward application of the Borel–Cantelli lemma and the Vapnik–Chervonenkis inequalities yields

$$d_n\Big(\theta_0,\widehat{\theta}(X^0_{-m_n+1})\Big) = O\Big(\sqrt{V_n\log n/n}\Big),\ P_{\theta_0}-\text{a.s.}.$$

V. Examples

Here, we present three examples of parametric families satisfying the conditions of Theorem 1 and thus admitting joint universal lossy coding and identification schemes. The following result [5] will be used throughout: Let $\mathcal{C} = \{A_\xi : \xi\in\mathbb{R}^N\}$ be a collection of measurable subsets of $\mathbb{R}^d$, such that $A_\xi = \{z\in\mathbb{R}^d : \Pi(z,\xi) > 0\}$ for all ξ, where for each $z\in\mathbb{R}^d$, $\Pi(z,\cdot)$ is a polynomial of degree s in the components of ξ. Then $\mathcal{C}$ is a VC class with $\mathsf{V}(\mathcal{C}) \le 2N\log(4es)$.

Stationary memoryless sources. Let $\mathcal{X} = \mathbb{R}$, and let $\{P_\theta : \theta\in\Lambda\}$ be the collection of all Gaussian i.i.d. processes with mean $m\in\mathbb{R}$ and variance $\sigma\in(0,\infty)$. Thus $\Lambda = \{(m,\sigma) : m\in\mathbb{R}, 0<\sigma<\infty\} \subset \mathbb{R}^2$. This class of sources trivially satisfies Condition 1 with $r = +\infty$, and it remains to check Conditions 2 and 3. To check Condition 2, consider the normalized

[3] Note that, up to a constant, the first term on the right-hand side has the same form as in Rissanen [1]; additional terms are due to the unboundedness of Λ and the fact that the points $\theta(i)$ do not form a regular grid.

relative entropy (information divergence) $D_n(\theta\|\theta')$ between P_θ^n and $P_{\theta'}^n$, with $\theta = (m, \sigma)$ and $\theta' = (m', \sigma')$ (which is equal to $D_1(\theta\|\theta')$ because the sources are i.i.d.). It is not hard to get the bound $D_n(\theta\|\theta') \le (1 + \sigma'/\sigma)^2 \|\theta - \theta'\|^2 / 2\sigma'^2$. Now fix a small $\delta \in (0, \sigma)$ and suppose that $\|\theta - \theta'\| < \delta$. Then $|\sigma - \sigma'| < \delta$, so we can further upper-bound $D_n(\theta\|\theta')$ as $D_n(\theta\|\theta') \le \frac{c_\theta^2}{2}\|\theta - \theta'\|^2$ for all θ' in the open ball of radius δ around θ, with $c_\theta = 3/(\sigma - \delta)$. Using Pinsker's inequality [4], we have $d_n(\theta, \theta')/\sqrt{n} \le \sqrt{2D_n(\theta\|\theta')} \le c_\theta \|\theta - \theta'\|$ for all n. Thus, Condition 2 holds. To check Condition 3 note that, for each n, the Yatracos class $\mathcal{A}_n$ consists of all sets of the form $\{x^n \in \mathbb{R}^n : \Pi(x^n, \theta, \theta') > 0\}$, $\theta, \theta' \in \Lambda$, where for each $x^n \in \mathcal{X}^n$ $\Pi(x^n, \theta, \theta')$ is a third-degree polynomial in $(\ln\sigma^2, \ln\sigma'^2, 1/\sigma^2, 1/\sigma'^2, m, m')$. Thus, $\mathcal{A}_n$ is a VC class with $\mathrm{V}(\mathcal{A}_n) \le 12\log(12e)$, satisfying Condition 3.

Autoregressive (AR) sources. Let $\mathcal{X} = \mathbb{R}$ and let $\boldsymbol{X}$ be a Gaussian AR(p) source. That is, there exist p real parameters $a_1, \ldots, a_p$, such that $X_n = -\sum_{i=1}^p a_i X_{n-i} + Y_n$ for all n, where $\boldsymbol{Y} = \{Y_i\}_{i\in\mathbb{Z}}$ is an i.i.d. Gaussian process with zero mean and unit variance. Let $\Lambda \subset \mathbb{R}^p$ be the set of all $a_1, \ldots, a_p$, such that all roots of the polynomial $A(z) = \sum_{i=0}^p a_i z^i$, $a_0 \equiv 1$, lie outside the unit circle in the complex plane. Under these conditions, for each $\theta \in \Lambda$ the process $\boldsymbol{X}$ is *exponentially β-mixing* [11], i.e., there exists some $\gamma = \gamma(\theta) \in (0, 1)$, such that $\beta_\theta(k) = O(\gamma^k)$. Now, for any fixed $r > 0$, $\gamma^k \le k^{-r}$ for k sufficiently large, so Condition 1 holds. For Condition 2, it can be shown that, for each $\theta \in \Lambda$, the asymptotic Fisher information matrix $I(\theta)$ exists (and is nonsingular) [12]. Thus, Condition 2 can be met. To verify Condition 3, consider the n-dimensional marginal $P_\theta(x^n)$, which has the normal density $p_\theta(x^n) = \mathcal{N}(x^n; 0, R_n(\theta))$, where $R_n(\theta)$ is the nth-order autocorrelation matrix of $\boldsymbol{X}$. For every $\theta \in \Lambda$, let $\bar\theta = (\theta, \ln\det R_n^{-1}(\theta))$. Since $\ln\det R_n^{-1}(\theta)$ is uniquely determined by θ, we have $A_{\theta,\theta'} = A_{\bar\theta,\bar\theta'}$ for all sets in the Yatracos class $\mathcal{A}_n$. This, and the fact that the entries of $R_n^{-1}(\theta)$ are quadratic functions of $a_1, \ldots, a_p$, implies that, for each x^n, the condition $x^n \in A_{\theta,\theta'}$ can be expressed as $\Pi(x^n, \bar\theta, \bar\theta') > 0$, where $\Pi(x^n, \cdot)$ is quadratic in the $2p + 2$ real variables $\bar\theta_1, \ldots, \bar\theta_{p+1}, \bar\theta'_1, \ldots, \bar\theta'_{p+1}$. Thus, $\mathrm{V}(\mathcal{A}_n) \le (4p + 4)\log(8e)$. Therefore, Condition 3 is met.

Hidden Markov processes. A hidden Markov process is a discrete-time finite-state homogeneous Markov chain, observed through a discrete-time memoryless channel (see [13] and references therein). Let $\boldsymbol{S} = \{S_i\}_{i\in\mathbb{Z}}$ be a stationary ergodic Markov process with $M < \infty$ states and the (unique) stationary distribution $\pi = (\pi_1, \ldots, \pi_M)$. Let $a_{ij} = \mathbb{P}(S_{t+1} = j | S_t = i)$, $1 \le i, j \le M$, denote the corresponding one-step transition probabilities. Let $\mathcal{X} = \mathbb{R}^d$, and consider a discrete-time memoryless channel with input alphabet $\mathcal{S} \triangleq \{1, \ldots, M\}$ and output alphabet $\mathcal{X}$, specified by a collection $\{p(\cdot|s) : s \in \mathcal{S}\}$ of probability densities on $\mathbb{R}^d$ w.r.t. the Lebesgue measure. The output process $\boldsymbol{X} = \{X_i\}_{i\in\mathbb{Z}}$ is the source of interest.

Let us assume that the channel transition densities are known, and that the one-step transition probabilities of the underlying Markov chain $\boldsymbol{S}$ are known to be strictly positive and bounded from below by some $a_0 > 0$. Thus, our parameter space is the set $\Lambda = \{\theta = [a_{ij}] \in \mathbb{R}^{M\times M} : a_{ij} > a_0, \forall i, j\}$. Under these assumptions, for any $\theta \in \Lambda$ the underlying Markov process $\boldsymbol{S}$ is exponentially β-mixing [14]. It can also be shown [5] that for every $\theta \in \Lambda$ there exists a measurable map $F : \mathcal{S} \times [0, 1] \to \mathcal{X}$, such that $X_i = F(S_i, U_i)$ for all $i \in \mathbb{Z}$, where U_i are i.i.d. random variables with uniform distribution on $[0, 1]$, independent of $\boldsymbol{S}$. The pair process $\{(S_i, U_i)\}$ is exponentially β-mixing, and therefore so is $\boldsymbol{X}$. This establishes Condition 1. Under additional technical assumptions on the densities $\{p(\cdot|s)\}$ it can be shown that the asymptotic Fisher information matrix $I(\theta)$ exists for all $\theta \in \Lambda$ [15], which implies that Condition 2 holds as well. Finally, to show that Condition 3 is satisfied, note that the n-dimensional marginal of P_θ for a given $\theta = [a_{ij}]$ has the density $p_\theta(x^n) = \sum_{s^n\in\mathcal{S}^n} \prod_{i=1}^n a_{s_{i-1}s_i} p(x_i|s_i)$, where $a_{s_0 s} \equiv \pi_s$ for all s. Then it follows that the Yatracos class $\mathcal{A}_n$ consists of sets of the form $\{x^n \in \mathcal{X}^n : \Pi(x^n, \theta, \theta') > 0\}$, $\theta = [a_{ij}], \theta' = [a'_{ij}] \in \Lambda$, where $\Pi(x^n, \cdot)$ is a polynomial of degree n in the $2M^2$ parameters $\{a_{ij}, a'_{ij}\}$. Thus, $\mathrm{V}(\mathcal{A}_n) \le 4M^2\log(4en)$, so that Condition 3 holds as well.

Acknowledgment

The author would like to thank Andrew Barron, Ioannis Kontoyiannis and Mokshay Madiman for useful discussions. This work was supported by the Beckman 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] M. Raginsky, "Joint fixed-rate universal lossy coding and identification of continuous-alphabet memoryless sources," *IEEE Trans. Inform. Theory*, 2005, submitted.

[3] ——, "Joint universal lossy coding and identification of i.i.d. vector sources," in *Proc. IEEE Int. Symp. on Information Theory*, Seattle, July 2006, pp. 577–581.

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

[5] M. Vidyasagar, *Learning and Generalization*, 2nd ed. London: Springer-Verlag, 2003.

[6] 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.

[7] T. Linder, "Learning-theoretic methods in vector quantization," in *Principles of Nonparametric Learning*, L. Györfi, Ed. New York: Springer-Verlag, 2001.

[8] P. Elias, "Universal codeword sets and representations of the integers," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 2, pp. 194–203, March 1975.

[9] B. Yu, "Rates of convergence for empirical processes of stationary mixing sequences," *Ann. Probab.*, vol. 22, no. 1, pp. 94–116, 1994.

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

[11] A. Mokkadem, "Mixing properties of ARMA processes," *Stochastic Process. Appl.*, vol. 29, pp. 309–315, 1988.

[12] A. Klein and P. Spreij, "The Bezoutian, state space realizations and Fisher's information matrix of an ARMA process," *Lin. Algebra Appl.*, vol. 416, pp. 160–174, 2006.

[13] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1518–1569, June 2002.

[14] P. Billingsley, *Probability and Measure*, 3rd ed. New York: Wiley, 1995.

[15] R. Douc, É. Moulines, and T. Rydén, "Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime," *Ann. Statist.*, vol. 32, no. 5, pp. 2254–2304, 2004.