Model Selection via Rate-Distortion Theory

(Invited Paper)

I. Kontoyiannis[1]
Department of Statistics
Purdue University
1399 Math Sciences Building
West Lafayette, IN 47906
Email: yiannis@stat.purdue.edu
Web: www.stat.purdue.edu/∼yiannis

Abstract — **Rissanen's Minimum Description Length (MDL) principle for model selection proposes that, among a predetermined collection of models, we choose the one which assigns the shortest description to the data at hand. In this context, a "description" is a lossless representation of the data that also takes into account the cost of describing the chosen model itself. We examine how the MDL principle might extend to the case when the requirement for lossless coding is relaxed (lossy compression), and we outline some of the mathematical and conceptual ingredients that facilitate this extension**

I. Introduction

Rissanen's Minimum Description Length (MDL) principle, as well as several other prominent model selection criteria, are based on the idea that among a predetermined collection of models (or model classes), the one which best captures the characteristics of the data is the one which can be used to encode the data using the smallest number of bits.[2]

In applications, it is often the case that we are willing to tolerate less accurate but simpler models. Consider for example the case of lossy data compression. Given a data string $x_1^n = (x_1, x_2, \ldots, x_n)$, the objective is to efficiently represent x_1^n by a corresponding string $y_1^n = (y_1, y_2, \ldots, y_n)$ that agrees with x_1^n to within some fixed level of accuracy. More precisely, we require that the distortion $\rho_n(x_1^n, y_1^n)$ between the original string x_1^n and its representation y_1^n is within some fixed bound, say

$$\rho_n(x_1^n, y_1^n) \leq D$$

(see below for more precise definitions).

Now suppose that x_1^n is generated as a realization of a random process $\{X_n\}$ with nth order distributions P_n. From rate-distortion theory [14][5] we know that the most efficient way to represent X_1^n is by an element Y_1^n of Shannon's random codebook – that is, a suitably chosen realization of the optimum reproduction distribution Q_n^*. As is well known, unless we restrict our selves to *lossless* compression, the distribution Q_n^* will typically be very different from P_n. Therefore, in the sense of rate-distortion theory, if we are willing to tolerate distortion up to level D in the description of our data, the most efficient "model" is that provided by Q_n^*.

In lossless compression, there is a natural correspondence between models for the data (that is, probability distributions) and prefix-free codes. This correspondence is briefly summarized in Section II. In Sections III-V we show that there is a similar (although more complicated) correspondence between probability distributions on the reproduction alphabet and lossy compression codes at a fixed distortion level, we propose a lossy analog to the lossless idealized Shannon code, and we show that it is competitively optimal. In Section VI we provide some of the technical details, and in Section VII we briefly mention some recent related work.

II. Background: Lossless Compression

Suppose that the data string $x_1^n = (x_1, x_2, \ldots, x_n)$ takes values in a a finite set A, called the *source alphabet*, that is, $x_1^n \in A^n$. A *prefix-free code*, or simply a *lossless code* on A^n is a map $\psi_n : A^n \to \{0,1\}^*$ with the property that no codeword in the range of ψ_n is a prefix of another. Every such code induces a length function $L_n : A^n \to \mathbb{N}$, defined by

$$L_n(x_1^n) = \text{ length of } [\psi_n(x_1^n)] \qquad \text{(in bits)}.$$

As is well-known, the prefix-free property is characterized by the following inequality; see, e.g., [6, Chapter 5]:

> **Kraft Inequality.** If ψ_n is a is a prefix-free code with length function L_n, then
>
> $$\sum_{x_1^n \in A^n} 2^{-L_n(x_1^n)} \leq 1. \qquad \text{(K)}$$
>
> Conversely, if L_n is a length function satisfying (K), then there is a prefix-free code ψ_n with length function L_n.

The Kraft inequality provides a natural and very useful correspondence between lossless codes and (sub-)probability distributions: From any length function satisfying (K) we get a "model" on A^n by defining

$$Q_n(x_1^n) \stackrel{\triangle}{=} 2^{-L_n(x_1^n)}, \quad x_1^n \in A^n. \qquad (1)$$

Conversely, if we ignore the constraint that the code-lengths $L_n(x_1^n)$ have to be integers, then for any probability distribution Q_n on A^n there is a code with code-lengths given by

$$L_n(x_1^n) \stackrel{\triangle}{=} -\log Q_n(x_1^n), \quad x_1^n \in A^n. \qquad (2)$$

[1] This work was supported in part by a grant from the Purdue Research Foundation.

[2] See, e.g., [13] or [4] for extensive discussions of the MDL principle, and the paper [15] for a description of the minimum message length (MML) principle; Akaike's information criterion (AIC) also admits a similar information-theoretic interpretation [1].

[Throughout, log denotes the logarithm taken to base 2.] The code with length function (2) is often called the *Shannon code* (or Shannon-Fano code) *with respect to the distribution* Q_n.

Now suppose that data are generated by a random process $\{X_n\}$, and let P_n be its nth order marginal, that is, P_n denotes the distribution of $X_1^n = (X_1, X_2, \ldots, X_n)$ on A^n. Assuming we know the true distribution P_n, the best code for describing X_1^n is the Shannon code with respect to P_n, i.e., the prefix-free code with code-lengths given by (2) with P_n in place of Q_n.

The following well-known result makes more precise the sense in which the Shannon code with respect to P_n is the "best" code.

Theorem 1. [2][3] *Shannon Code Competitive Optimality* Let $\{L_n\}$ be the length functions of an arbitrary sequence of prefix-free codes on A^n. If X_1^n are data generated by an arbitrary random process $\{X_n\}$ with marginal distributions P_n, we have:

(a) For all n,

$$E_{P_n}[L_n(X_1^n)] \geq E_{P_n}[-\log P_n(X_1^n)] = H(X_1^n),$$

where $H(X_1^n)$ is the entropy of X_1^n.

(b) For all n and all $K \geq 1$, the probability

$$\begin{aligned} &\Pr\{L_n \text{ beats the Shannon code by } K \text{ bits or more}\} \\ &\quad = \Pr\{L_n(X_1^n) \leq -\log P_n(X_1^n) - K\} \\ &\quad \leq 2^{-K}. \end{aligned}$$

(c) If $\{c_n\}$ is a sequence of nonnegative constants such that $\sum_{n\geq 1} 2^{-c_n} < \infty$, then, with probability one,

$$L_n(X_1^n) \geq -\log P_n(X_1^n) - c_n, \quad \text{eventually.}$$

This result justifies, to some extent, our identification of the Shannon code as the "best" code, and it further encourages us to think of codes and models as being interchangeable. In the remainder of this note, we identify a corresponding "Shannon code" for lossy compression, we argue that it leads to a natural correspondence between codes and models in the lossy case, and we demonstrate its competitive optimality.

III. Lossy Compression: Random Coding & a Lossy Shannon Code

Let $\{X_n\}$ be a random source taking values in the source alphabet A, and for $1 \leq i \leq j \leq \infty$, write X_i^j for the vector of random variables $(X_i, X_{i+1}, \ldots, X_j)$ and similarly write $x_i^j = (x_i, x_{i+1}, \ldots, x_j) \in A^{j-i+1}$ for a realization of X_i^j. For most of this section we will assume that $\{X_n\}$ is a memoryless source, that is, that the X_n are independent and identically distributed (i.i.d.) random variables with common distribution P on A.

Let $\hat{A}$ denote the *reproduction alphabet*. For the sake of simplicity, we assume throughout that both A and $\hat{A}$ are finite sets. For an arbitrary nonnegative function ρ on $A \times \hat{A}$, we define a sequence of single-letter distortion measures ρ_n on $A^n \times \hat{A}^n$ by

$$\rho_n(x_1^n, y_1^n) = \frac{1}{n}\sum_{i=1}^n \rho(x_i, y_i) \quad x_1^n \in A^n, \;\; y_1^n \in \hat{A}^n,$$

and we also make the customary assumption that

$$\max_{x\in A} \min_{y\in\hat{A}} \rho(x,y) = 0.$$

We consider *variable-length codes operating at a fixed distortion level*, that is, codes C_n defined by triplets (B_n, ϕ_n, ψ_n) where:

(a) B_n is a subset of $\hat{A}^n$ called the *codebook*;

(b) $\phi_n : A^n \to B_n$ is the *quantizer*;

(c) $\psi_n : B_n \to \{0,1\}^*$ is a prefix-free map on B_n.

For $D \geq 0$, the code $C_n = (B_n, \phi_n, \psi_n)$ is said to *operate at distortion level* D, if it encodes each source string with distortion D or less:

$$\rho_n(x_1^n, \phi_n(x_1^n)) \leq D, \qquad \text{for all } \; x_1^n \in A^n.$$

As ψ_n is a prefix-free lossless code, it induces a length function L_n on B_n,

$$L_n(y_1^n) = \text{ length of } [\psi_n(y_1^n)]. \quad y_1^n \in B_n,$$

Moreover, the code C_n induces a length function ℓ_n on A^n,

$$\ell_n(x_1^n) = \text{ length of } [\psi_n(\phi_n(x_1^n))],$$

and he functions L_n and ℓ_n are clearly related by

$$\ell_n(x_1^n) = L_n(\phi_n(x_1^n)). \tag{3}$$

Shannon's celebrated source coding theorem [14] characterizes the best achievable compression ratio among codes operating at distortion level D. In particular, suppose $\{X_n\}$ is a memoryless source with distribution P. Then, for any sequence of codes $\{C_n = (B_n, \phi_n, \psi_n) \; ; \; n \geq 1\}$ operating at distortion level D, the expected compression ratio $E[\ell_n(X_1^n)]/n$ is asymptotically bounded below by the *rate-distortion function* $R(D)$,

$$\liminf_{n\to\infty} \frac{E[\ell_n(X_1^n)]}{n} \geq R(D) \quad \text{bits per symbol} \tag{4}$$

where $R(D) = R(P,D)$ is defined by the well-known formula

$$R(D) = \inf_{(X,Y)} I(X;Y) \tag{5}$$

and the infimum is over all pairs (X,Y) such that $X \sim P$ and $E[\rho(X,Y)] \leq D$. Moreover, Shannon showed that the the lower bound in (4) is (asymptotically) achievable by a sequence of random codes. Next we outline their construction (the presentation is along the lines of [10]).

Let (X^*, Y^*) be a pair of random variables achieving the infimum in (5), and let Q^* denote the distribution of Y^*. Following Shannon, we generate i.i.d. codewords $Y(i) = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,n})$, $i = 1, 2, \ldots,$ each drawn according to the product distribution $Q_n^* \stackrel{\triangle}{=} (Q^*)^n$ on $\hat{A}^n$. Now, given a source string X_1^n, we can encode it by specifying the index $i = W_n$ of the first codeword $Y(i)$ for which the distortion $\rho_n(X_1^n, Y(i))$ is D or less:

$$W_n = \inf\{i \geq 1 \; : \; \rho_n(X_1^n, Y(i)) \leq D\}.$$

The description of W_n takes no more than

$$\log W_n + \log\log W_n + O(\log\log\log W_n) \quad \text{bits,}$$

and the representation of X_1^n by $Y(i)$ is always within distortion D or less. But W_n, the "waiting time" until the first D-close match for X_1^n, is approximately equal to the reciprocal of the probability of finding such a match. More precisely,

$$\log W_n \leq \log[1/Q_n^*(B(X_1^n, D))] + \log n + 2 \log\log n,$$

eventually, with probability one, where the "distortion balls" $B(x_1^n, D)$ are defined by

$$B(x_1^n, D) = \{y_1^n \in \hat{A}^n \ : \ \rho_n(x_1^n, y_1^n) \leq D\}, \quad x_1^n \in A^n. \tag{6}$$

Also, from the recent results in [7] and [16] we know that the probabilities $\log[1/Q_n^*(B(X_1^n, D))]$ are equal to

$$nR(D) + \sum_{i=1}^n f(X_i) \ + \frac{1}{2}\log n + O(\log\log n)$$

eventually, with probability one (see Proposition 3 in [10]), where $f : A \to \mathbb{R}$ is a function depending on D and P, such that $E_P[f(X)] = 0$ (see Section VI for more details and a precise definition of f). Putting these estimates together, we get the following upper bound on the description length of Shannon's random code (Theorem 2 below is a slight refinement of Theorem 5 (a) in [10]):

Theorem 2. *Pointwise Performance of Random Coding* Suppose $\{X_n\}$ is a memoryless source with rate-distortion function $R(D)$. The description lengths of Shannon's random code are bounded above by

$$nR(D) + \sum_{i=1}^n f(X_i) + \frac{5}{2}\log n + O(\log\log n) \quad \text{bits}$$

eventually, with probability one.

Is this the best we can do? The answer is "yes, up to a few $(\log n)$ terms." Theorem 3 below follows from [10, Theorem 4].

Theorem 3. *Pointwise Optimality of Random Coding* Suppose $\{X_n\}$ is a memoryless source with rate-distortion function $R(D)$. For any sequence of codes $\{C_n\}$ with corresponding length functions $\{\ell_n\}$, operating at distortion level D, we have

$$\ell_n(X_1^n) \geq nR(D) + \sum_{i=1}^n f(X_i) - (1+\epsilon)\log n \quad \text{bits}$$

eventually, with probability one.

Observing that the upper and lower bounds provided in Theorems 2 and 3, respectively, agree in their first- and second-order terms, we decide to ignore the terms of order $O(\log n)$ and smaller, and we define the (idealized) *code-lengths of the Shannon code at distortion level D* by

$$\boxed{\ell_n^{(S)}(x_1^n) \ \stackrel{\triangle}{=} \ nR(D) + \sum_{i=1}^n f(x_i), \qquad x_1^n \in A^n.}$$

These "idealized" code-lengths turn out to have essentially the same strong optimality properties that the Shannon code enjoys in the lossless case. Theorem 4 generalizes Theorem 1 above to the lossy case (see [10, Corollary 2]).

Theorem 4. *Shannon Code Competitive Optimality* Suppose $\{X_n\}$ is a memoryless source with rate-distortion function $R(D)$, and let $\{\ell_n\}$ be the length functions of an arbitrary sequence of codes $\{C_n\}$ operating at distortion level D. Then we have:

(a) For all n,

$$E_{P^n}[\ell_n(X_1^n)] \geq E_{P^n}[\ell_n^{(S)}(X_1^n)] = nR(D).$$

(b) For all n and all $K \geq 1$, the probability

$$\begin{aligned} &\Pr\{\ell_n \text{ beats } \ell_n^{(S)} \text{ by } K \text{ bits or more}\} \\ &\quad = \ \Pr\{\ell_n(X_1^n) \leq \ell_n^{(S)}(X_1^n) - K\} \\ &\quad \leq \ 2^{-K}. \end{aligned}$$

(c) If $\{c_n\}$ is a sequence of nonnegative constants such that $\sum_{n\geq 1} 2^{-c_n} < \infty$, then, with probability one,

$$\ell_n(X_1^n) \geq \ell_n^{(S)}(X_1^n) - c_n, \quad \text{eventually.}$$

IV. An Example

The competitive optimality property of the Shannon code (Theorem 1 (b) in the lossless case and Theorem 4 (b) in the lossy case) may seem somewhat remarkable at first sight. It says that there is a sequence of codes (the Shannon codes) that, no matter how hard we try, we can can only beat them by K bits with probability at most 2^{-K} – and this holds for any block-length n.

Although this statement is precisely accurate in the lossless case, in the lossy case we ignored some $(\log n)$ terms, so it takes a little more work to make this into an honest bound. This is carried out in the following example.

Suppose we want to compress a 300x300 pixel grayscale image, call it $\mathbf{X}$. The source and reproduction alphabets each have size 256, and the block-length n is equal to 300x300 symbols, or approximately 90 Kbytes. Then for each fixed distortion level D, a slightly modified version of the random code described above gives a code operating at that distortion level, whose description lengths $\ell^*(\mathbf{X})$ have the following property: For *any* other code operating at distortion level D, the probability that its description lengths are significantly shorter than ℓ^* is negligible. For example, the following is precisely true:

For any *code C with code-lengths ℓ operating at distortion level D:*

$$\begin{aligned} &\Pr\{\ell \ \textit{ beats } \ell^* \textit{ by 11 bytes or more}\} \\ &\quad = \ \Pr\{\ell(\mathbf{X}) \leq \ell^*(\mathbf{X}) - 88 \ \textit{ bits}\} \\ &\quad \leq \ 2^{-39}. \end{aligned}$$

The exact assumptions under which the above bound holds are that: $\mathbf{X}$ is a 300x300 i.i.d. vector taking values in $A = \{0, 1, \ldots 255\}$, that the distortion level $D > 0$ is chosen outside the uninteresting region where $R(D) = 0$, and that n (in this case 90,000) is large enough so that the optimum reproduction distribution satisfies the (mild) condition that $Q_n^*(B(x_1^n, D)) \in (0, 1/2]$ for all x_1^n.

V. The Codes/Distributions Correspondence

The strong optimality of the idealized lossy Shannon code gives us a clear "target" to aim for, at least in the sense of data compression. We want to come as close as possible to the performance of Shannon's random code as described above.

Next we argue that achieving compression performance close to $\ell_n^{(S)}$ is, to some extent, equivalent to obtaining accurate estimates of the optimal reproduction distribution Q_n^*.

We now step back to the general case when the source $\{X_n\}$ is not necessarily memoryless, and outline the correspondence between codes at a fixed distortion level and distributions Q_n on $\hat{A}^n$.

A. Distributions ⇒ (Random) Codes. This is the more straightforward of the two directions. Given an arbitrary sequence $\{Q_n\}$ of probability distributions Q_n on $\hat{A}^n$, we can repeat Shannon's random coding argument, this time with respect to the (typically suboptimal) distributions Q_n. Recall that, in the case of a memoryless source, and with Q_n being the optimal distribution Q_n^*, the random coding argument produced a sequence of codes operating at distortion level D, with code-lengths

$$\ell_n(X_1^n) \;\approx\; -\log Q_n^*(B(X_1^n,D)) \;\approx\; \ell_n^{(S)}(X_1^n) \quad \text{bits},$$

where "≈" means that the difference between successive terms is at most of order $O(\log n)$, with probability 1. Similarly, repeating the random coding argument for an arbitrary sequence of distributions Q_n, we obtain a sequence of codes with code-lengths

$$\ell_n(X_1^n) \;\approx\; -\log Q_n(B(X_1^n,D)) \quad \text{bits.} \qquad (7)$$

This is the natural analog of the relation (2) in the lossless case. Note that relation (7) holds for an *arbitrary* sequence of distributions Q_n, as long as they produce random codes with finite rate, i.e., as long as

$$\limsup_{n\to\infty} -\frac{1}{n}\log Q_n(B(X_1^n,D)) < \infty$$

with probability one.

B. Codes ⇒ Distributions. Recall that a code C_n operating at distortion level D is defined by a triplet (B_n,ϕ_n,ψ_n), and the compression performance of C_n is described by its length function ℓ_n. In this notation, ψ_n is a prefix-free map defined on the codebook $B_n \subset \hat{A}^n$, and it has a corresponding length function L_n defined on B_n. Given such a code, we can define a (sub-)probability distribution on $\hat{A}^n$ by

$$Q_n(y_1^n) \;\stackrel{\triangle}{=}\; \begin{cases} 2^{-L_n(y_1^n)} & \text{if } y_1^n \in B_n \\ 0 & \text{otherwise.} \end{cases} \qquad (8)$$

This is the lossy analog of relation (1) in the lossy case. With this definition, we can get a simple but useful lower bound on the performance of C_n (cf. [10]): Since C_n operates at distortion level D, for any $x_1^n \in A^n$ we have (recalling (3)):

$$\begin{aligned} \ell_n(x_1^n) &= L_n(\phi_n(x_1^n)) \\ &= -\log Q_n(\phi_n(x_1^n)) \\ &\geq -\log Q_n(B(x_1^n,D)). \end{aligned}$$

Comparing this with (7) further reinforces our interpretation of the quantities

$$-\log Q(B(x_1^n,D))$$

as the natural lossy analogs of the Shannon code-lengths

$$-\log Q(x_1^n).$$

VI. Some of the Details

Here we give some of the more technical details about the function $f(\cdot)$ appearing in Theorems 2 and 3.

For a memoryless source with distribution P and rate-distortion function $R(D)$, let $D > 0$ be a distortion level such that $R(D) > 0$. As before, write Q^* for the distribution of the random variable Y^* achieving the minimum in (5). For each $x \in A$ and each $\lambda \in \mathbb{R}$ let

$$\Lambda_x(\lambda) = \log_e\left(\sum_{y\in\hat{A}} Q^*(y)e^{\lambda\rho(x,y)}\right)$$

(where $\log_e$ denotes the natural logarithm), and write λ^* for the unique $\lambda < 0$ such that

$$\frac{d}{d\lambda}\left[E_P(\Lambda_X(\lambda))\right] = D.$$

Then, for $x \in A$, define

$$f(x) = f_D(x) \stackrel{\triangle}{=} (\log e)\big[-\Lambda_x(\lambda^*) - E_P(-\Lambda_X(\lambda^*))\big].$$

Clearly $f(X)$ has mean zero (with respect to P), and in general it is non-degenerate, that is, it is usually not identically equal to zero. This statement was recently made precise in [8], where, among other things, the following is proved (under some mild conditions):

Theorem 5. *Nondegeneracy of f*
Suppose that $\{X_n\}$ is a memoryless source, that $A = \hat{A}$, and that $\rho(x,y)$ is a symmetric distortion measure. Only two possibilities exist:

A. Either $f(x) = f_D(x)$ is identically equal to zero for only finitely many values of D, or

B. The source distribution P is uniform and $\rho(x,y)$ is a "permutation" distortion measure, in which case $f(x) = f_D(x)$ is identically equal to zero for *all* D.

VII. Related Work

Here we briefly mention some recent work that explores connections between rate-distortion theory and model selection.

Jun Muramatsu's PhD thesis [11] contains very interesting results connecting lossy data compression with algorithmic complexity (in the sense of Kolmogorov). Dave Donoho gave a seminar in the spring of 1998 at Stanford on the "Shannon estimator" [9], a denoising technique based on random coding. Amir Najmi's PhD thesis [12] proposes a different model-selection criterion motivated by data compression ideas, and he also draws some connections with rate-distortion theory and with Donoho's Shannon estimator.

References

[1] H. Akaike. Prediction and entropy. In *A celebration of statistics*, pages 1–24. Springer, New York, 1985.

[2] P.H. Algoet. *Log-Optimal Investment.* PhD thesis, Dept. of Electrical Engineering, Stanford University, 1985.

[3] A.R. Barron. *Logically Smooth Density Estimation.* PhD thesis, Dept. of Electrical Engineering, Stanford University, 1985.

[4] A.R. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. (Information theory: 1948–1998). *IEEE Trans. Inform. Theory*, 44(6):2743–2760, 1998.

[5] T. Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Prentice-Hall Inc., Englewood Cliffs, NJ, 1971.

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

[7] A. Dembo and I. Kontoyiannis. The asymptotics of waiting times between stationary processes, allowing distortion. *Ann. Appl. Probab.*, 9:413–429, 1999.

[8] A. Dembo and I. Kontoyiannis. Critical redundancy in lossy source coding. *Submitted for publication*, December 1999. Available from `www.stat.purdue.edu/people/yiannis`.

[9] D. Donoho. The Shannon estimator. Statistics Seminar, Dept. of Statistics, Stanford University, April 1998.

[10] I. Kontoyiannis. Pointwise redundancy in lossy data compression and universal lossy data compression. *IEEE Trans. Inform. Theory*, 46(1):136–152, January 2000.

[11] J. Muramatsu. *Universal Data Compression Algorithms for Stationary Ergodic Sources Based on the Complexity of Sequences.* PhD thesis, Nagoya University, 1998.

[12] A. Najmi. *Data Compression, Model Selection and Statistical Inference.* PhD thesis, Dept. of Electrical Engineering, Stanford University, December 1999.

[13] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific, Singapore, 1989.

[14] C.E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec.*, part 4:142–163, 1959. Reprinted in D. Slepian (ed.), *Key Papers in the Development of Information Theory*, IEEE Press, 1974.

[15] C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *J. Roy. Statist. Soc. Ser. B*, 49(3):240–265, 1987. With discussion.

[16] E.-h. Yang and Z. Zhang. On the redundancy of lossy source coding with abstract alphabets. *IEEE Trans. Inform. Theory*, 45(4):1092–1110, 1999.