A Minimum Description Length Proposal for Lossy Data Compression

M. Madiman M. Harrison I. Kontoyiannis

June 2004

Abstract

We give a development of the theory of lossy data compression from the point of view of statistics. This is partly motivated by the enormous success of the statistical approach in lossless compression, in particular Rissanen's celebrated Minimum Description Length (MDL) principle. A precise characterization of the fundamental limits of compression performance is given, for arbitrary data sources and with respect to general distortion measures.

The starting point for this development is the observation that there is a precise correspondence between compression algorithms and probability distributions (in analogy with the Kraft inequality in lossless compression). This leads us to formulate a version of the MDL principle for lossy data compression. We discuss the consequences of the lossy MDL principle and explain how it leads to potential practical design lessons for vector-quantizer design.

We introduce two methods for selecting efficient compression algorithms, the *lossy Maximum Likelihood Estimate* (LMLE) and the *lossy Minimum Description Length Estimate* (LMDLE). We describe their theoretical performance and give examples illustrating how the LMDLE has superior performance to the LMLE.

1 Introduction

Formally and somewhat roughly speaking, the central problem of universal data compression is that of selecting an appropriate code among a given family, in order to obtain good compression performance. Following [3], we identify compression algorithms with probability distributions on the reproduction space.

More precisely, consider a source $\{X_n\}$ with values in the alphabet A, which is to be compressed with distortion no more than D with respect to an arbitrary sequence of distortion measures $\rho_n : A^n \times \hat{A}^n \to [0, \infty)$, where $\hat{A}$ is the reproduction alphabet. For a source string $x_1^n = (x_1, x_2, \ldots, x_n) \in A^n$, let $B(x_1^n, D)$ denote the distortion-ball of radius D around x_1^n:

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

[*] Mokshay Madiman (corresponding author) and Matthew Harrison are with the Division of Applied Mathematics, Brown University, Box F, 182 George Street, Providence, RI 02912, USA. Email: mokshay@dam.brown.edu, Matthew_Harrison@brown.edu

[†] Ioannis Kontoyiannis is with the Division of Applied Mathematics and the Department of Computer Science, Brown University, Box F, 182 George Street, Providence, RI 02912, USA. Email: yiannis@dam.brown.edu

[‡] This research for this work was supported in part by NSF grant #0073378-CCR, and by USDA-IFAFS grant #00-52100-9615

We consider the class of codes $C_n : A^n \to \{0,1\}^*$ that operate at distortion level D. Any such code C_n is the composition of a quantizer ϕ_n that maps A^n to a (finite or countably infinite) codebook $B_n \subset \hat{A}^n$, followed by a prefix-free encoder $\psi_n : B_n \to \{0,1\}^*$. The code C_n operates at distortion level D, if $\rho_n(x_1^n, \phi_n(x_1^n)) \leq D$ for all $x_1^n \in A^n$. The figure of merit here is, of course, given by the length function $L_n^{(\text{true})}$ of the code C_n:

$$L_n^{(\text{true})}(x_1^n) = \text{length of}\, \psi_n(\phi_n(x_1^n)) \quad \text{bits.}$$

As shown in [3], there is a precise correspondence between compression algorithms and probability distributions Q on $\hat{A}^n$. Similar to the lossless case, this correspondence is expressed in terms of the *idealized lossy Shannon code-lengths*

$$L_n(x_1^n) = -\log Q(B(x_1^n, D)) \ \text{ bits.} \tag{1}$$

More specifically, for any code C_n with length-function L_n operating at distortion level D, there is a probability distribution Q_n on $\hat{A}^n$ such that $L_n^{(\text{true})}(x_1^n) \geq -\log Q_n(B(x_1^n, D))$ for all $x_1^n \in A^n$. Conversely, for any source $\{X_1^n\}$ and any sequence of "admissible" distributions $\{Q_n\}$, there is a sequence of codes $\{C_n, L_n\}$ operating at distortion level D, such that, for (almost) any realization of the source, $L_n^{(\text{true})}(X_1^n) \leq -\log Q_n(B(X_1^n, D)) + O(\log n)$, eventually with probability 1 (henceforth we will write "w.p.1").

Thus motivated, we pose the problem of selecting a "good" code among a given family, as the statistical estimation problem of selecting one of the available probability distributions $\{Q_\theta; \theta \in \Theta\}$ on the reproduction space.

2 Inference and Lossy Compression

In the lossless case the problem optimal compression is, theoretically at least, equivalent to finding a probability distribution Q that in some sense minimizes the code-lengths $-\log Q(x_1^n)$. In that case the optimal choice is simply to take Q to be the true source distribution.

As it turns out [3], in the *lossy* case the best choice is to take Q to be the optimal reproduction distribution Q^* on $\hat{A}^n$, i.e., the optimal output distribution in Shannon's rate-distortion problem. Thus, our goal here is to do statistical inference with the goal of estimating this distribution Q^*, and *not* the true source distribution, from the data.

More generally, given a family of probability distributions $\{Q_\theta; \theta \in \Theta\}$ on the reproduction space, we want to chose the one whose limiting coding rate, call it $R(P, \theta, D)$,

$$R(P, \theta, D) \overset{\triangle}{=} \lim_{n\to\infty} -\frac{1}{n} \log Q_\theta(B(X_1^n, D)) \quad \text{bits/symbol,} \tag{2}$$

is as small as possible. If $Q^* = Q_{\theta^*}$ happens to be in the above class, then of course $R(P, \theta^*, D)$ is simply the rate-distortion function of the source. But in general we do not always require that to be the case, and we think of our target distribution Q_{θ^*} as that corresponding to $\theta^* = \arg\min_\theta R(P, \theta, D)$. Intuitively, we think of Q_{θ^*} as *the simplest distribution describing all the regularity in the data, with accuracy no worse than D.*

3 A Lossy MDL Principle

A natural way to estimate the optimal θ^* empirically is to try and minimize the idealized code-lengths (1), or equivalently to maximize the probabilities $Q_\theta(B(X_1^n, D))$. We thus define the Lossy Maximum Likelihood Estimate (LMLE) as

$$\hat{\theta}_n^{\mathrm{LML}} = \arg\min_{\theta\in\Theta}[-\log Q_\theta(B(X_1^n, D))]. \tag{3}$$

Our first main result (Theorem 1 below) states that, under very general conditions [1][4], this estimate is consistent as $n \to \infty$. But as with the classical (lossless) maximum likelihood estimate (MLE), $\hat{\theta}_n^{\mathrm{LML}}$ also has some undesirable properties. First, the infimum in the definition of $\hat{\theta}_n^{\mathrm{LML}}$ is not really a code-length; if we choose one of the θ's based on the data, we should also describe the chosen θ itself. Indeed, there are examples [1] where $\hat{\theta}_n^{\mathrm{LML}}$ is *not* consistent but it becomes consistent when appropriately modified to correspond to an actual two-part code. Second, maximum-likelihood-type estimates tend to "overfit" the data: For example, if in the classical (lossless) setting we try to estimate the distribution of a binary Markov chain, then, even if the data turns out to be i.i.d., the Maximum Likelihood Estimate (MLE) will typically be a Markov (non-i.i.d.) distribution.

To rectify these problems, we consider "penalized" versions of the lossy MLE, similar to those considered in the lossless case: We define the Lossy Minimum Description Length Estimate (LMDLE) as

$$\hat{\theta}_n^{\mathrm{LMDL}} = \arg\min_{\theta\in\Theta}[-\log Q_\theta(B(X_1^n, D)) + \ell_n(\theta)], \tag{4}$$

where $\ell_n(\theta)$ is a given "penalty function." There are two natural families of of penalties to consider – general penalties satisfying Kraft's inequality for a lossless code on a countable parameter space Θ (so that the MDL estimate then indeed corresponds to a two-part code), and dimension-based penalties. For simplicity, here we only consider penalties of the form

$$\ell_n(\theta) = \frac{1}{2}k(\theta)\log n, \tag{5}$$

where $k(\theta)$ is a non-negative integer that can be interpreted as "dimension" in a sense we make precise later.

An important point to note is that both the LMLE and the LMDLE are difficult to compute, either analytically or computationally, even in the simplest examples. For this reason, we introduce two more estimators that are useful when the data and the reproduction distributions $\{Q_\theta\}$ are i.i.d. We define the pseudo-LMLE and the pseudo-LMDLE by:

$$\tilde{\theta}_n^{\mathrm{LML}} \equiv \arg\min_{\theta\in\Theta} R(\hat{P}_n, \theta, D) \tag{6}$$

$$\tilde{\theta}_n^{\mathrm{LMDL}} \equiv \arg\min_{\theta\in\Theta}[nR(\hat{P}_n, \theta, D) + \ell_n(\theta)], \tag{7}$$

where $\hat{P}_n$ denotes the empirical distribution of X_1^n. The motivation for these definitions comes from the expansion [6][5],

$$-\log Q_\theta^n(B(x_1^n, D)) = nR(\hat{P}_{x_1^n}, \theta, D) + \frac{1}{2}\log n + O(1),$$

which suggests that, to first order, we may replace the logarithm of the Q-probability of a distortion ball by the rate function $R(\cdot, \theta, D)$ for the empirical distribution of the data.

Theorem 1 (Consistency) Under general conditions, the LMLE and LMDLE are strongly consistent estimators, in that they converge to θ^* with probability one. In the case of i.i.d. source and reproduction distributions, the pseudo-LMLE and pseudo-LMDLE are also strongly consistent.

The proof of Theorem 1 [1][4] is rather long and technical, and uses ideas and techniques related to the notion of epi-convergence in convex analysis.

In the following section we present our second main result, which shows that the LMDLE family of estimators avoids the common problems of the corresponding maximum-likelihood-based estimators. For the sake of simplicity, this is first illustrated via two representative examples, and a more general result is given in Section 5.

4 Examples: MDL vs. Maximum Likelihood

Example 1. Gaussian Codes. Suppose that the source $\{X_n\}$ is a real-valued, stationary and ergodic, with zero mean and finite variance, and suppose that each Q_θ is an i.i.d. $N(0,\theta)$ distribution with $\theta \in \Theta = [0,\infty)$. If we take ρ_n to be mean-squared error, then $\tilde{\theta}_n^{\mathrm{LML}}$ can be explicitly computed as a function of the data using the explicit form of the rate function $R(P,\theta,D)$. Furthermore, the the following dichotomy [5] [1] highlights the important difference between the pseudo-LMLE and the pseudo-LMDLE: Although they both converge to the optimal θ^*, as $n \to \infty$ we actually have:

$$\begin{aligned} \tilde{\theta}_n^{\mathrm{LML}} &\neq \theta^* \ \text{ infinitely often, w.p.1} \\ \tilde{\theta}_n^{\mathrm{LMDL}} &= \theta^* \ \text{ eventually, w.p.1.} \end{aligned} \tag{8}$$

Here we have taken $k(\theta)$ in the penalty function to be equal to 1 for all $\theta \neq \theta^*$, and zero otherwise. Figure 1 shows an explicit numerical example illustrating the above behavior.

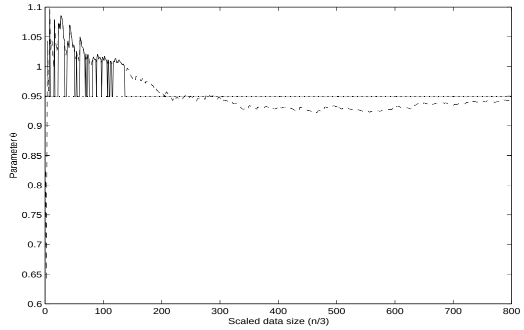

Figure 1: The dashed line denotes the pseudo-LML estimate and the solid line is the pseudo-LMDL estimate. In this example, source variance=1, $D = 0.1$ and $\theta^* = 0.949$.

This rather artificial example shows clearly that the MDLE not only converges to the correct value, but when there is a subset of the parameter space containing θ^* which we consider particularly "simple," we can make sure that the MDLE actually "finds" this simple subset eventually with probability one.

Similar conclusions are obtained for the case when we take the coding distributions Q_θ to be i.i.d. $N(\theta,1)$ with $\theta \in \mathbb{R}$.

Example 2. I.I.D. Bernoulli Codebooks. Consider an i.i.d. Bernoulli source with parameter $p = \Pr(X = 1)$. Figure 2 illustrates the behavior of the MLE and MDLE, when the $k(\theta) = 0$ for $\theta = \theta^* = (p - D)/(1 - 2D)$, and zero otherwise, where θ^* obviously corresponds to the $R(D)$-achieving output distribution. Once again, we get the same dichotomy as in (8) between the behavior of the pseudo-LMDLE and the pseudo-LMLE estimators. In fact, repeated simulations indicate that the pseudo-LMDLE not only "hits and stays at" θ^* (unlike the pseudo-LMLE which bounces around forever), but that it actually does so pretty fast. An example is shown in Figure 2 below.

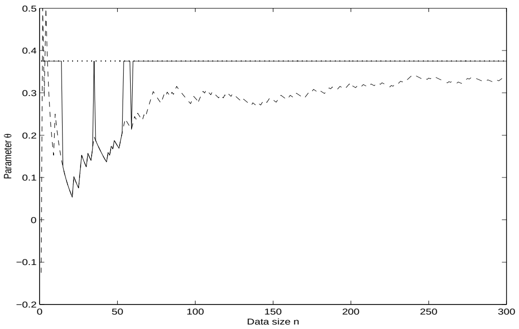

Figure 2: The dashed line represents the pseudo-LML estimate, and the solid line is the pseudo-LMDL estimate. Here $p = .4$, $D = .1$ and $\theta^* = .375$.

5 A More General Dichotomy Theorem

Suppose that the memoryless source $\{X_n\}$ takes values in a finite alphabet A of size m. Let Θ parametrize the simplex of all i.i.d. probability distributions on $A = \hat{A}$, and take Θ to be the class of reproduction distributions from which we want to pick the best code. Also take $\{\rho_n\}$ to be a family of single-letter distortion measures.

Now suppose that $L_1 \subset L_2 \subset ... \subset L_s \subset \Theta$ is a nested sequence of sets that parametrize increasingly "complicated" subsets of the simplex. We express our preference for "simpler" subsets L by penalizing θ more when it *only* belongs to more complicated sets; so $k(\theta) \equiv \min\{1 \leq i \leq s : \theta \in L_i\}$ denotes the index of the simplest subset containing θ. This defines LMDL and pseudo-LMDL estimates as in (4) and (6). For convenience, we also set $s^* = k(\theta^*)$. Then our second main result [1][5] is a more general statement of the dichotomy established in the above two examples.

Theorem 2 (Dichotomy) Suppose that:

i. The source distribution P belong to the class

$$\mathbb{S}(D) = \{P : D < D_{\max}(P), Q^* \text{is unique, support}(P) = A, \text{and support}(Q^*) = A\}$$

where $D_{\max}(P) = \min_{y \in \hat{A}} E_P[\rho(X, y)]$;

ii. The dimension of L_{s^*}, where $s^* \equiv k(\theta^*)$, is strictly less than $m - 1$.

Then, under a simple technical condition, the following hold:

1. $\tilde{\theta}_n^{\text{LML}} \notin L_{s^*}$ i.o. w.p.1

2. $\tilde{\theta}_n^{\text{LMDL}} \in L_{s^*}$ eventually w.p.1

3. $\hat{\theta}_n^{\mathrm{LMDL}} \in L_{s^*}$ eventually w.p.1.

Thus, the pseudo-LMDL estimate approaches Q^* eventually through codes in L_{s^*}. In other words, if there is a "nice" subset L_{s^*} of Θ and we want to know if θ^* is in L_{s^*}, then the (appropriately defined) pseudo-LMDL estimate *finds* this set in finite time whenever θ^* does indeed belong to it. The pseudo-LML estimate, on the other hand, *cannot* display this behavior – even though it will converge to the correct value of θ^*, it will infinitely often take excursions outside of L_{s^*}.

Perhaps the most natural choice of the nested sequence of sets would be a sequence of truncated affine sets of varying dimension, where each L_i has dimension i. This is the situation which motivated the posing of the dichotomy problem, and it is in this sense that $k(\theta)$ represents the "dimension" of θ. We thus use the word dimension, keeping in mind that θ in itself is just a vector in an $(m-1)$-dimensional space and that the "dimension" really only invokes an implicit hierarchy of sets in which we would like θ^* to be as far up as possible.

The proof [5] of Theorem 2 is based on a detailed asymptotic analysis, building on the results of [6][2] and the references therein. The main technical tools are delicate large deviation bounds and a vector form of the law of the iterated logarithm.

In addition to the clear dichotomy for the pseudo-estimates, Theorem 2 strongly suggests a dichotomy for the LMLE and LMDLE. Though we have not proved that the LMLE has to fluctuate outside of L_{s^*} forever as it approaches Q^*, it *typically* does so, while the LMDLE *always* eventually lies within L_{s^*}.

The above result also has another implication. In [3] it was shown that "lossy mixture codes" (as defined there based on the codes-distributions correspondence) are universal over the class $\mathbb{S}(D)$, with a redundancy of order $O(\frac{\log n}{n})$ bits/symbol. One consequence of Theorem 2 is that codes based on any one of the four "lossy" estimates defined in this work are universal over a similar class of sources, also with a redundancy of $O(\frac{\log n}{n})$ bits/symbol.

Further results, more general statements, and connections with vector-quantization are developed in [1].

References

[1] M. Harrison, I. Kontoyiannis, and M. Madiman. A Minimum Description Length principle in lossy data compression. *Preprint*, 2003.

[2] A. Dembo and I. Kontoyiannis. Source coding, large deviations, and approximate pattern matching. *IEEE Trans. Inform. Theory*, 48:1590–1615, June 2002.

[3] I. Kontoyiannis and J. Zhang, "Arbitrary source models and Bayesian codebooks in rate-distortion theory", *IEEE Trans. Inform. Theory*, 48:2276–2290, 2002.

[4] M. Harrison, "The convergence of lossy Maximum Likelihood Estimators", *Brown University APPTS Report # 03-5*, April 2003.

[5] M. Madiman and I. Kontoyiannis, "Second order properties of lossy likelihoods and the MLE-MDL Dichotomy in Lossy Compression", *Brown University APPTS Report # 03-7*, November 2003.

[6] E-h. Yang and Z. Zhang, "On the redundancy of lossy source coding with abstract alphabets", *IEEE Trans. Inform. Theory*, vol. 45, pp.1092-1110, 1999.