Minimum Description Length vs. Maximum Likelihood in Lossy Data Compression

M. Madiman, M. Harrison, and I. Kontoyiannis*
Division of Applied Mathematics, Brown University
Providence, RI 02912, USA
{mokshay, mth, yiannis}@dam.brown.edu

I. Introduction

Consider a source $\{X_n\}$ with values in A, to be compressed with distortion no more than D with respect to an arbitrary sequence of distortion measures $\rho_n : A^n \times \hat{A}^n \to [0,\infty)$. Let $B(x_1^n, D) = \{y_1^n \in \hat{A}^n : \rho_n(x_1^n, y_1^n) \le D\}$ denote the distortion-ball of radius D around the source string x_1^n. We consider codes $C_n : A^n \to \{0,1\}^*$ that operate at distortion level D. Any such code is the composition of a quantizer ϕ_n that maps A^n to a discrete 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 $\rho_n(x_1^n, \phi_n(x_1^n)) \le D$ for all $x_1^n \in A^n$. The figure of merit is the length function L_n of the code C_n:

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

Our starting point is the following precise correspondence between compression algorithms and probability distributions Q on $\hat{A}^n$: Similarly to the lossless case, this correspondence is expressed in terms of the *idealized lossy Shannon code-lengths*

$$L_n(X_1^n) = -\log Q(B(X_1^n, D)) \ \text{ bits.} \tag{1}$$

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.

In the lossless case the problem of optimal compression is theoretically equivalent to finding a distribution Q that in some sense minimizes the code-lengths $-\log Q(x_1^n)$. In the *lossy* case, given a family of probability distributions $\{Q_\theta; \theta \in \Theta\}$ on the reproduction space, we want to chose the one whose limiting rate $R(\theta, D)$ is as small as possible,

$$R(\theta, D) = \lim_{n\to\infty} -\frac{1}{n} \log Q_\theta(B(X_1^n, D)).$$

If the above class is large enough, then $R(\theta^*, D)$ is just the rate-distortion function, but in general our target distribution Q_{θ^*} is that corresponding to $\theta^* = \arg\min_\theta R(\theta, D)$. Thus, our goal here is to do statistical inference to estimate this distribution Q^*, and *not* the true source distribution.

II. A Lossy MDL Principle

A natural way to estimate the optimal θ^* empirically is to minimize the idealized code-lengths (1). We thus define the Lossy Maximum Likelihood Estimate (LMLE) as

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

Our first main result [1] is that: **Under very general conditions the LMLE is consistent, $\hat{\theta}_n^{\text{LML}} \to \theta^*$ w.p.1.**

But as with the classical (lossless) MLE, this $\hat{\theta}_n^{\text{LML}}$ also has some undesirable properties – it tends to "overfit" the data. To rectify this, we considered "penalized" versions of the LMLE, and define the **Lossy Minimum Description Length Estimate** (LMDLE) as

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

where $\ell_n(\theta)$ is a given "penalty function." For simplicity, we only consider penalties of the form $\frac{1}{2}k(\theta)\log n$, where $k(\theta)$ is an integer (intuitively, $k(\theta)$ is the dimension of the smallest subspace θ belongs to in a hierarchy of nested subspaces).

Our second main result [1] is that: **Under general conditions, the LMDLE is: (i) consistent; (ii) it finds the smallest-dimensional subspace that θ^* belongs to, with probability 1.**

Example. Lossy MDL vs. Lossy MLE. Suppose that the i.i.d. source $\{X_n\}$ takes values in a finite alphabet A, let Θ parametrize the simplex of all i.i.d. probability distributions on $A = \hat{A}$, and let $L_0 \subset L_1 \subset \ldots \subset L_s \subset \Theta$ be nested parameter sets. We express our preference for "simpler" subsets L by penalizing θ more when it belongs to more complicated sets: $k(\theta) = \min\{0 \le i \le s : \theta \in L_i\}$. Suppose the dimension of L_{s^*}, where $s^* = k(\theta^*)$, is strictly less than $|A| - 1$. Here:

Theorem. *Under simple technical conditions, with probability 1: The LMLE will forever fluctuate outside L_{s^*} as it approaches Q^*, whereas the LMDLE will approach Q^* eventually through codes in L_{s^*}.*

For a Bernoulli(p) source, Fig. 1 illustrates the behavior of (very good approximations to) the LMLE and LMDLE when the "preferred" set L_0 is the singleton $\{\theta^*\}$ containing the $R(D)$-achieving distribution $\theta^* = (p-D)/(1-2D)$. Repeated simulations show that the LMDLE "hits and stays at" θ^* quite fast, unlike the LMLE which bounces around forever.

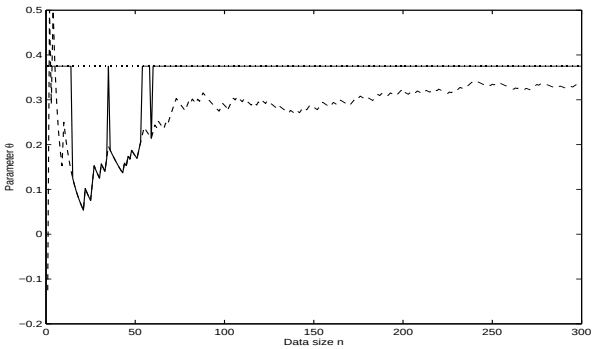

Fig. 1: The dashed and solid lines denote the LMLE and LMDLE respectively. Here $p = .4$, $D = .1$ and $\theta^* = .375$.

References

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

*Supported in part by NSF grant #0073378-CCR and USDA-IFAFS grant #00-52100-9615.