A Minimum Description Length Proposal for Lossy Data Compression

by Mokshay Madiman

Joint work with M. Harrison and I. Kontoyiannis
Outline

• The Problem: Lossy Data Compression

• Codes as Probability Distributions

• Fundamental limits and a generalized AEP

• Selecting good codes: Inference

• Consistency

• MLE/MDL Dichotomy + Examples

• Comments and conclusions
The Problem: Data Compression

Setting

Data: \( x_1^n \in A^n \), arbitrary alphabet \( A \)
Quantizer: \( q_n : A^n \to C^n \), discrete \( C \subset \hat{A} \)
Encoder: \( e_n : C^n \to \{0, 1\}^* \)
Code-length: \( L_n(x_1^n) = \text{length of } e_n(q_n(x_1^n)) \text{ bits} \)

Distortion

Distortion function: \( \rho_n : A^n \times \hat{A}^n \to [0, \infty) \) is “nice”
Distortion ball: \( B(x_1^n, D) := \{ y_1^n \in \hat{A}^n : \rho_n(x_1^n, y_1^n) \leq D \} \)
Code operates at dist’n level \( D \): \( q_n(x_1^n) \in B(x_1^n, D) \) for all \( x_1^n \in A^n \)
Codes as Probability Distributions

For lossless codes, \( L_n(x_1^n) \approx -\log Q_n(x_1^n) \)

For lossy codes, \( L_n(X_1^n) \approx -\log Q_n(B(X_1^n, D)) \)

More specifically, we have a Lossy Kraft Inequality (K&Z, 2002):

(\(\Leftarrow\)) For any code with code-lengths \( L_n \) and distortion level \( D \), there is a probability distribution \( Q_n \) on \( \hat{A}^n \) with

\[ L_n(x_1^n) \geq -\log Q_n(B(x_1^n, D)) \text{ bits, for all } x_1^n \]

(\(\Rightarrow\)) For any source \( \{X_n\} \) and any reasonable sequence of probability distributions \( Q_n \) on \( \hat{A}^n \), \( n = 1, 2, \ldots \), there is a sequence of codes with distortion levels \( D \) and code-lengths such that

\[ L_n(X_1^n) \leq -\log Q_n(B(X_1^n, D)) + 2\log n \text{ bits, eventually, w.p.1} \]

\[ EL_n(X_1^n) \leq E[-\log Q_n(B(X_1^n, D))] + 2\log n \text{ bits, eventually} \]
Fundamental limits and a generalized AEP

Asymptotic Equipartition Property (AEP)

If the source \( \{X_n\} \sim P \) is stationary and ergodic, the (lossless) compression performance w.r.t any sequence of “nice” distributions \( \{Q_n\} = Q \) is given by

\[
- \frac{1}{n} \log Q_n(X^n_1) \rightarrow H(P) + D(P||Q)
\]

bits/symbol, as \( n \rightarrow \infty \), w.p.1

A Generalized AEP (Kieffer’91, L&S’97, Y&K’98, Y&Z’99, Chi’01, D&K’02)

If the source \( \{X_n\} \sim P \) is stationary and ergodic, and \( \rho_n \) is a single-letter distortion measure, the compression performance w.r.t any sequence of “nice” distributions \( \{Q_n\} = Q \) is given by

\[
- \frac{1}{n} \log Q_n(B(X^n_1, D)) \rightarrow R(P, D) + \Delta(P, Q, D) := R(P, Q, D)
\]

bits/symbol, as \( n \rightarrow \infty \), w.p.1
How to select good codes?

The IID Example

<table>
<thead>
<tr>
<th>Lossless coding</th>
<th>Lossy coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Want a code based on the $Q^*$ that minimizes $H(P) + D(P∥Q)$</td>
<td>Want a code based on “the” $Q^*$ that minimizes $R(P, Q, D)$</td>
</tr>
<tr>
<td>The optimal $Q^*$ is true source distribution $P$</td>
<td>For $D &gt; 0$, optimal $Q^*(\neq P)$ achieves Shannon’s r.d.f. $R(P, D)$</td>
</tr>
<tr>
<td>Selecting a good code is like estimating</td>
<td>Selecting a good code is an indirect estimation problem</td>
</tr>
<tr>
<td>a source distribution from data</td>
<td></td>
</tr>
</tbody>
</table>
How to select good codes?

The IID Example

<table>
<thead>
<tr>
<th>Lossless coding</th>
<th>Lossy coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Want code based on the $Q^*$ that minimizes $H(P) + D(P | Q)$</td>
<td>Want code based on “the” $Q^*$ that minimizes $R(P, Q, D)$</td>
</tr>
<tr>
<td>Optimal $Q^*$ is true source distribution $P$</td>
<td>For $D &gt; 0$, optimal $Q^*(\neq P)$ achieves Shannon’s r.d.f. $R(P, D)$</td>
</tr>
<tr>
<td>Selecting a good code is like estimating a source distribution from data</td>
<td>Selecting a good code is an indirect estimation problem</td>
</tr>
</tbody>
</table>

which motivates...

<table>
<thead>
<tr>
<th>Information theory</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code ($L_n$)</td>
<td>Probability distribution ($Q_n$)</td>
</tr>
<tr>
<td>Class of codes</td>
<td>Statistical model ${Q_\theta : \theta \in \Theta}$</td>
</tr>
<tr>
<td>Code selection</td>
<td>Estimation: find optimal $\theta^* \in \Theta$ (i.e., one which minimizes $R(P, Q_\theta, D)$)</td>
</tr>
</tbody>
</table>
Choose a parametric family of probability distributions \( \{Q_\theta : \theta \in \Theta\} \) corresponding to a convenient class of codes.

**Lossy MLE**

The Lossy Maximum Likelihood Estimator (LMLE) is

\[
\hat{\theta}^{\text{LML}}_n = \arg \min_{\theta \in \Theta} \left[ - \log Q_\theta(B(X_1^n, D)) \right]
\]
Statistical Inference - I

Choose a parametric family of probability distributions \( \{ \mathbb{Q}_\theta : \theta \in \Theta \} \) corresponding to a convenient class of codes.

Lossy MLE

The Lossy Maximum Likelihood Estimator (LMLE) is

\[
\hat{\theta}_n^{\text{LML}} = \arg \min_{\theta \in \Theta} \left[ -\log \mathbb{Q}_\theta (B(X_1^n, D)) \right]
\]

The LMLE is nice...

The LMLE is consistent in great generality: As \( n \to \infty \), \( \hat{\theta}_n^{\text{LML}} \to \theta^* \) w.p.1 under weak conditions
Statistical Inference - 1

Choose a parametric family of probability distributions \( \{Q_\theta : \theta \in \Theta\} \) corresponding to a convenient class of codes.

Lossy MLE

The Lossy Maximum Likelihood Estimator (LMLE) is

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

The LMLE is nice...

The LMLE is consistent in great generality: As \( n \to \infty \), \( \hat{\theta}_n^{\text{LML}} \to \theta^* \) w.p.1 under weak conditions

But Problems with LMLE

- Overfitting
- Not a real code
Lossy MDL

The Lossy Minimum Description Length Estimator (LMDLE) is

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

where $\ell_n(\theta)$ is a given “penalty function”

The LMDLE is nice...

The LMDLE is consistent in great generality: As $n \to \infty$, $\hat{\theta}_{\text{LMDL}} \to \theta^*$ w.p.1 under weak conditions

Does the LMDLE solve the problems of the LMLE?
Let the source be IID $P \sim N(0, 1)$ and consider IID coding distributions $Q_\theta \sim N(0, \theta)$, $\theta \in (0, \infty)$. We use the penalty function

$$
\ell_n(\theta) = \begin{cases} 
0 & \text{if } \theta = \theta^* = 1 - D \\
\frac{1}{2} \log n & \text{if } \theta \neq \theta^* 
\end{cases}
$$

where the lower-dimensional set $\{\theta^*\} \subset (0, \infty)$ is declared to be our “preferred” set.

The dashed line denotes the pseudo-LMLE and the solid line is the pseudo-LMDLE.
Inference for IID source & coding distributions

An approximate codelength (Y&Z'98)

$$- \log Q^n_\theta (B(X^n_1, D)) \approx nR(\hat{P}_n, \theta, D) \quad \text{eventually w.p.1}$$

where $\hat{P}_n$ is the empirical distribution of the data $X^n_1$

Idea of Pseudo-estimators

Replace $[- \log Q^n_\theta (B(X^n_1, D))]$ in definition of lossy estimators by $[nR(\hat{P}_n, \theta, D)]$

Definitions

The pseudo-LMLE and pseudo-LMDLE are

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

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

where $\ell_n(\theta)$ is a given penalty function

Consistency

The pseudo-estimators are consistent
Gaussian example: Details

<table>
<thead>
<tr>
<th>Source</th>
<th>( P \sim N(0, V) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class of codes</td>
<td>( { Q_\theta \sim N(0, \theta) : \theta \in (0, \infty) } )</td>
</tr>
<tr>
<td>Distortion</td>
<td>Single-letter with ( \rho(x, y) = (x - y)^2 ); assume ( D \in (0, V) )</td>
</tr>
<tr>
<td>Optimal code</td>
<td>( \theta^* = V - D )</td>
</tr>
</tbody>
</table>
| Penalty function | \[
\ell_n(\theta) = \begin{cases} 
0 & \text{if } \theta = \theta^* \\
\frac{1}{2} \log n & \text{if } \theta \neq \theta^*
\end{cases}
\] |
| Pseudo-LMLE | \( \theta_n^{LML} = \mu_n^2 + V_n - D \) where \( \mu_n \) and \( V_n \) are the mean and variance of \( \hat{P}_n \) |
| Pseudo-LMDLE | \[
\hat{\theta}_n^{LMDL} = \begin{cases} 
\theta^* & \text{if } R_2 \leq R_1 \\
\hat{\theta}_n^{LML} & \text{otherwise}
\end{cases}
\] where \( R_1 = R(\hat{P}_n, D) + \frac{\log n}{2n} \) and \( R_2 = R(\hat{P}_n, \theta^*, D) \) |

Apply the Law of the Iterated Logarithm

- Detailed computation yields \( R_2 - R(\hat{P}_n, D) = O(V_n - V)^2 \)
- \( R_2 - R(\hat{P}_n, D) \) is \( O\left(\frac{\log \log n}{n}\right) \) and in particular, \( o\left(\frac{\log n}{n}\right) \)
- Thus \( R_2 < R_1 \Rightarrow \hat{\theta}_n^{LMDL} = \theta^* \) eventually w.p.1
- On the other hand, \( \hat{\theta}_n^{LML} - \theta^* = \frac{1}{n} \sum_{i=1}^{n}(X_i^2 - EX^2) \neq 0 \) i.o. w.p.1
The IID finite alphabet case

Setting

• Source distribution $P$ takes values in a finite alphabet $A$
• $\Theta$ parametrizes the simplex of all IID probability distributions on $\hat{A} = A$
• Single-letter distortion measures

Complexity

• Suppose $L_1 \subset L_2 \subset \ldots \subset L_s \subset \Theta$ parametrize increasingly “complicated” subsets of the simplex (or “model classes”)
• Preference for “simpler models” is expressed by using the penalty

$$\ell_n(\theta) = \frac{k(\theta)}{2} \log n$$

where

$$k(\theta) \equiv \min\{1 \leq i \leq s : \theta \in L_i\}$$

denotes the index of the simplest $L_i$ containing $\theta$
Under reasonable restrictions on $P$ and $\theta^*$ and a simple technical condition, we have

1. $\tilde{\theta}_n^{LML} \notin L_k(\theta^*)$ i.o. w.p.1
2. $\tilde{\theta}_n^{LMDL} \in L_k(\theta^*)$ eventually w.p.1
3. $\hat{\theta}_n^{LMDL} \in L_k(\theta^*)$ eventually w.p.1
Conclusions

The message

• We proposed maximum likelihood and MDL-type estimators for the purpose of finding good lossy source codes
• These estimators are consistent (i.e., they eventually yield optimal codes)
• Lossy MDL is efficient at model selection (unlike the lossy MLE)

Comments and Directions

• Penalty term of order $O(\log n)$ suffices (as in the lossless case) for the lossy MDL estimator to “find” the appropriate model class in finite time
• Practical Applications to VQ design need to be explored
• Suggests a theoretical framework for looking at lossy source coding through its statistical interpretation, and throws up many directions for future work

○ — ○ — ○