MDL Ideas in Lossy Data Compression

Ioannis Kontoyiannis
Brown University

joint work with
Junshan Zhang, Matt Harrison, Amir Dembo

MSRI Workshop on Information Theory
Berkeley, CA

February 25, 2002

Outline

Motivation and Background: Lossless Data Compression

Lossy Data Compression: MDL Point of View

1. *Ideal Compression:* Kolmogorov Distortion-Complexity
2. *Codes as Probability Distributions:* A Lossy Kraft Inequality
3. *Coding Theorems:* Asymptotics, Finite Block-lengths
4. *Code Performance:* Generalized AEP
5. *Solidarity with Shannon Theory:* Stationary Ergodic Sources
6. *Choosing a Code:* The Lossy MLE & A Lossy MDL Proposal
7. *Toward Practicality:* Pre-processing in VQ Design

$\Big[$ 8. *Computational Issues* $\Big]$

Emphasis

Concentrate On:

$\rightsquigarrow$ *General* sources, *general* distortion measures

$\rightsquigarrow$ Nonasymptotic, "pointwise" results

$\rightsquigarrow$ Precise performance bounds

$\rightsquigarrow$ Systematic develompent of MDL point of view, parallel to lossless case

$\rightsquigarrow$ Connections with VQ Applications . . .

Background Questions:

$*$ Why is lossy compression so much *harder*?

$*$ What's so *different* (mathematically) between them?

$*$ How much do the *"right" models* matter for real data?

Some Related Work

- J. Muramatsu (1994, PhD Thesis 1998)
- Chou-Effros-Gray (1996)
- Steinberg-Verdú (1996), Han (1997, 1998)
- Yang-Zhang (1998, 2000)
- A. Najmi (PhD Thesis)
- Bin Yu *et al* (1999-2001)
- R. Zamir and students (2001)
- R. Gray (DMI, Gauss mixture VQ)
- D. Donoho (2002)

$\rightsquigarrow$ Cover, Barron, Rissanen, . . .

Lossy Compression: The Basic Problem

Consider

Data string $x_1^n = (x_1, x_2, \ldots, x_n)$ to be compressed

Each x_i taking values in the *source alphabet* A
e.g., $A = \{0, 1\}$, $A = \mathbb{R}$, $A = \mathbb{R}^k$, $\ldots$

Problem

Find efficient approximate representation $y_1^n = (y_1, y_2, \ldots, y_n)$ for x_1^n
with y_i taking values in the *reproduction alphabet* $\hat{A}$

Efficient means "simple" or "compressible"

Approximate means that the distortion $d_n(x_1^n, y_1^n)$ is $\leq$ some level D
where $d_n : A^n \times \hat{A}^n$ is an "arbitrary" distortion measure

Step 1. Ideal Compression: Kolmogorov Distortion Complexity

For computable data and distortion:

Define (Muramatsu-Kanaya 1994)

The *(Kolmogorov) distortion-complexity at distortion level D*:

$$K_D(x_1^n) = \min\{\ell(p) \ : \ p \text{ s.t. } U(p) \in B(x_1^n, D)\} \quad \text{bits}$$

where $U(\cdot)$ = universal Turing machine

$B(x_1^n, D)$ = distortion-ball of radius D around x_1^n:

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

Properties

$K_D(x_1^n)$ is: (a) “machine-independent” (b) *not* computable

(c) $\approx nR(D)$ for stationary ergodic data

$\rightsquigarrow$ THE fundamental limit of compression

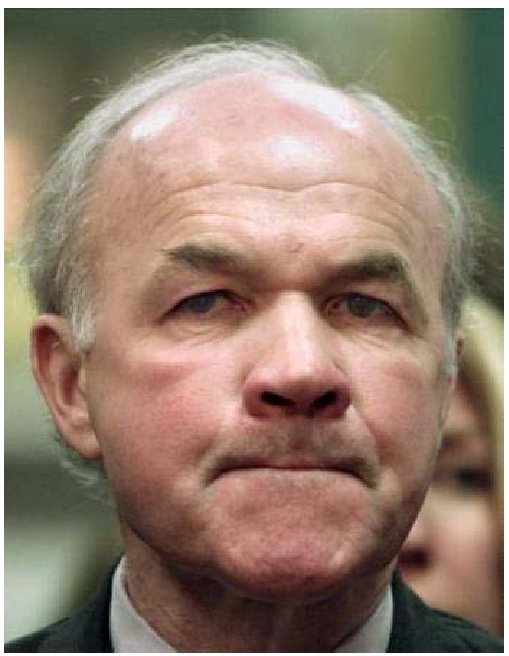
Lossy Compression in More Detail

Data: $X_1^n = X_1, X_2, \ldots, X_n$ with distribution P_n on A^n

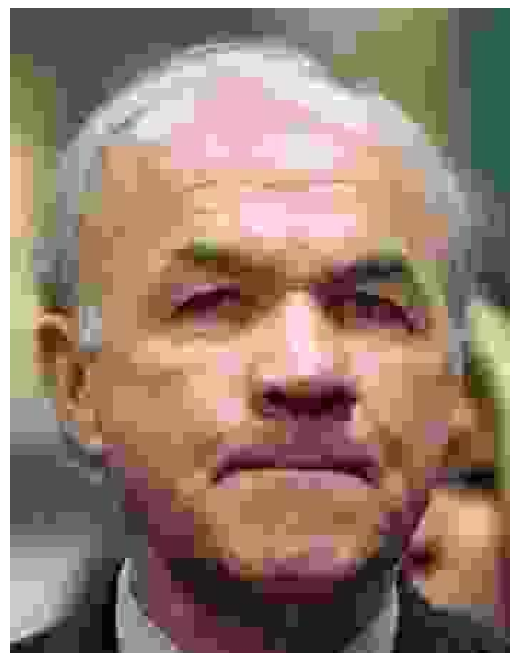
Quantizer: $q_n : A^n \to$ codebook $B_n \subset \hat{A}^n$

Encoder: $e_n : B_n \to \{0,1\}^*$ (prefix-free)

Code-length: $L_n(X_1^n) = L_n(q_n(X_1^n)) =$ length of $e_n(q_n(X_1^n))$ bits



$\xrightarrow{q_n}$



$\xrightarrow{e_n}$ 0010111010110
101101000 . . .

$\xleftarrow{e_n^{-1}}$

Step 2. Codes as Probability Distributions

The code $(C_n, L_n) = (B_n, q_n, e_n, L_n)$ *operates at distortion level* D, if

$$d_n(x_1^n, q_n(x_1^n)) \le D \quad \text{for every} \quad x_1^n \in A^n$$

Kraft Inequality

($\Leftarrow$) For every lossless code (C_n, L_n) there is a prob measure Q_n on A^n s.t.

$$L_n(x_1^n) \ge -\log Q_n(x_1^n) \text{ bits}$$

for *all* x_1^n

($\Rightarrow$) For any prob measure Q_n on A^n there is a code (C_n, L_n) s.t.

$$L_n(x_1^n) \le -\log Q_n(x_1^n) + 1 \text{ bits}$$

for *all* x_1^n

Theorem: Lossy Kraft Inequality

($\Leftarrow$) For every code (C_n, L_n) operating at distortion level D there is a prob meas. Q_n on $\hat{A}^n$ s.t.

$$L_n(x_1^n) \ge -\log Q_n(B(x_1^n, D)) \text{ bits}$$

for *all* x_1^n

($\Rightarrow$) For any "admissible" sequence of measures $\{Q_n\}$ on $\hat{A}^n$ there are codes $\{C_n, L_n\}$ at dist'n level D s.t.

$$L_n(X_1^n) \le -\log Q_n(B(X_1^n, D)) + \log n$$

bits, eventually, w.p.1

Remarks on the Codes-Measures Correspondence

- The converse part is a finite-n result as in the lossless case
- The direct part is asymptotic (random coding)
  but with (near) *optimal convergence rate*
- Both results are valid without ANY (...) assumptions on the source
  or the distortion measure
- Similar results hold in expectation with a $\frac{1}{2}\log n$ rate
- Admissibility $\Leftrightarrow$ the $\{Q_n\}$ yield codes with finite rate:

$$\limsup_{n\to\infty} -\frac{1}{n}\log Q_n(B(X_1^n, D)) \le \text{ some finite } R \text{ bits/symbol, w.p.1}$$

- This suggests a natural lossy analog for the Shannon code-lengths:

$$\boxed{L_n(X_1^n) \;=\; -\log Q_n\Big(B(X_1^n, D)\Big) \quad \text{"bits"}}$$

"*All codes are random codes*"

Proof Outline

($\Leftarrow$) Given a code (C_n, L_n), let $Q_n(y_1^n) \propto \begin{cases} 2^{-L_n(y_1^n)} & \text{if } y_1^n \in B_n \\ 0 & \text{otherwise} \end{cases}$

Then for *all* x_1^n :

$$L_n(x_1^n) = L_n(q_n(x_1^n)) \geq -\log Q_n(q_n(x_1^n)) \geq -\log Q_n(B(x_1^n, D)) \quad \text{bits}$$

($\Rightarrow$) Given Q_n, generate IID codewords $Y_1^n(i) \sim Q_n$:

$$Y_1^n(1) \quad Y_1^n(2) \quad Y_1^n(3) \quad \cdots$$

Encode X_1^n as the position of the first D-close match:

$$W_n = \min\{i \;:\; d_n(X_1^n, Y_1^n(i)) \leq D\}$$

This takes
$$\begin{aligned} L_n(X_1^n) &\approx \log W_n \quad \text{bits} \\ &\approx \log[\text{waiting time for a match}] \\ &\approx \log[1/\text{prob of a match}] \\ &\approx -\log Q_n(B(X_1^n, D)) \quad \text{bits} \end{aligned}$$

□

Step 3. Coding Theorems: Best Achievable Performance

Let Q_n^* achieve:

$$K_n(D) \stackrel{\triangle}{=} \inf_{Q_n} E[-\log Q_n(B(X_1^n, D))]$$

Theorem: Finite-n Bounds

i. For any code (C_n, L_n) operating at distortion level D :

$$E[L_n(X_1^n)] \geq K_n(D) \geq R_n(D) \quad \text{bits}$$

ii. For any (other) prob measure Q_n on A^n and any K:

$$\Pr\Big\{ -\log Q_n(B(X_1^n, D)) \leq -\log Q_n^*(B(X_1^n, D)) - K \ \text{ bits}\Big\} \ \leq \ 2^{-K}$$

Proof. Selection in convex families:
Bell-Cover version of the Kuhn-Tucker conditions $\square$

Coding Theorems Continued

Theorem: Asymptotic Bounds

i. For any seq of codes $\{C_n, L_n\}$ operating at distortion level D :

$$L_n(X_1^n) \geq -\log Q_n^*(B(X_1^n, D)) - \log n \quad \text{bits, eventually, w.p.1}$$

with Q_n^* as before

ii. There is a seq of codes $\{C_n^*, L_n^*\}$ operating at distortion level D s.t.

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

Proof.

i. Finite-n bound + Markov inequality + Borel-Cantelli + extra care

ii. Already proved □

Interpretation

Target

Approximate the performance of the optimal Shannon code:

Find $\{Q_n\}$ that yield code-lengths

$$L_n(X_1^n) \;=\; -\log Q_n\Big(B(X_1^n, D)\Big) \quad \text{bits}$$

close to those of the optimal "Shannon code":

$$L_n^*(X_1^n) \;=\; -\log Q_n^*\Big(B(X_1^n, D)\Big) \quad \text{bits}$$

Performance?

Step 4. Code Performance: Generalized AEP

Suppose

The source $\{X_n\}$ is stationary ergodic with distribution $\mathbb{P}$

$\{Q_n\}$ are the marginals of a stationary ergodic $\mathbb{Q}$

$d_n(x_1^n, y_1^n) = \frac{1}{n}\sum_{i=1}^n d(x_i, y_i)$ is a single-letter distortion measure

Theorem: Generalized AEP [L. & Szpan.], [Dembo & K], [Chi], [...]

If $\mathbb{Q}$ is mixing enough and $d(x, y)$ is not wild:

$$-\frac{1}{n}\log Q_n(B(X_1^n, D)) \to R(\mathbb{P}, \mathbb{Q}, D) \quad \text{bits/symbol, w.p.1}$$

where

$$R(\mathbb{P}, \mathbb{Q}, D) = \lim_{n\to\infty} \frac{1}{n} \inf_{P_{X_1^n} = P_n,\ E[d_n(X_1^n, Y_1^n)] \le D} H(P_{X_1^n, Y_1^n} \| P_n \times Q_n)$$

Proof. Based on very technical large deviations $\square$

Step 5. Sanity Check: Stationary Ergodic Souces

Suppose

The source $\{X_n\}$ is stationary ergodic with distribution $\mathbb{P}$

$d_n(x_1^n, y_1^n) = \frac{1}{n}\sum_{i=1}^n d(x_i, y_i)$ is a single-letter distortion measure

As before, Q_n^* achieves $K_n(D) = \inf_{Q_n} E[-\log Q_n(B(X_1^n, D))]$

Theorem ([Kieffer], [K & Zhang])

i. $K(D) = \lim_{n\to\infty} \frac{1}{n} K_n(D) = \lim_{n\to\infty} \frac{1}{n} R_n(D) = R(D)$

ii. $-\frac{1}{n}\log Q_n^*(B(X_1^n, D)) \to R(D)$ bits/symbol, w.p.1

Outline

Recall our program:

1. *Ideal Compression:* Kolmogorov Distortion-Complexity
2. *Codes as Probability Distributions:* A Lossy Kraft Inequality
3. *Coding Theorems:* Asymptotics, Finite Block-lengths
4. *Code Performance:* Generalized AEP
5. *Solidarity with Shannon Theory:* Stationary Ergodic Sources

6. *Choosing a Code:* The Lossy MLE & A Lossy MDL Proposal

7. *Toward Practicality:* Pre-processing in VQ Design

[**8.** *Computational Issues*]

So far

Setting

We identified codes with prob measures

$$L_n(X_1^n) \;=\; -\log Q_n\Big(B(X_1^n, D)\Big) \quad \text{bits}$$

Design

– What are good codes like?

– How de we find them?

How can we empirically design/choose a good code?

Given a parametric family $\{\mathbb{Q}_\theta \;;\; \theta \in \Theta\}$ of codes how do we choose θ?

Step 6. Choosing a Code: The Lossy MLE & A Lossy MDL Principle

Greedy empirical code selection

Given a parametric family $\{\mathbb{Q}_\theta \,;\, \theta \in \Theta\}$ of codes, define:

$$\hat{\theta}_{\rm MLE} \stackrel{\triangle}{=} \arg\inf_{\theta\in\Theta} \Big[-\log \mathbb{Q}_\theta(B(X_1^n, D)) \Big]$$

Problems With the MLE: As with classical MLE

Encouraged overfitting
Does *not* lead to real codes

Solution: Follow coding intuition

For the MLE to be useful, *it needs to be described as well*

$\Rightarrow$ Consider two-part codes with code-lengths

$$L_n(X_1^n) = -\log \mathbb{Q}_\theta(B(X_1^n, D)) + \underbrace{\ell_n(\theta)}_{\text{“description” of } \theta}$$

where: either (c_n, ℓ_n) is a prefix-free code on Θ
or $\ell_n(\theta)$ is an appropriate penalization term

Two Lossy MDL Proposals

Two Specific MDL-like proposals: Define:

$$\hat{\theta}_{\mathrm{MDL}} \stackrel{\triangle}{=} \arg\inf_{\theta \in \Theta} \Big[-\log \mathbb{Q}_\theta(B(X_1^n, D)) + \ell_n(\theta) \Big]$$

with either: (a) $\ell_n(\theta) = \frac{\dim(\mathbb{Q}_\theta)}{2} \log n$

(b) $\ell_n(\theta) = \begin{cases} \ell(\theta) & \text{for } \theta \text{ in some countable } \Gamma \subset \Theta; \\ \infty & \text{otherwise} \end{cases}$

Consistency?

Does $\hat{\theta}_{\mathrm{MLE}}$ /$\hat{\theta}_{\mathrm{MDL}}$ asymptotically lead to optimal compression?

What is the optimal θ^*?

What if θ^* is not unique? (the typical case)

As in the classical case: Often hard to prove

Proof is often example-specific

Motivation For The Lossy MDL Estimate

1. *Leads to realistic code selection*
2. *An example*:

Theorem

Let $\{X_n\}$ be real-valued, stationary, ergodic, $E(X_n) = 0$, $\mathrm{Var}(X_n) = 1$

Take $d_n(x_1^n, y_1^n) =$ MSE, let $D \in (0,1)$ fixed

Θ: $\mathbb{Q}_\theta \sim$ IID $\frac{1}{2}N(0, 1-D) + \frac{1}{2}N(0, \theta)$, $\theta \in [0,1]$

With $\ell_n(\theta) = \frac{\dim(\mathbb{Q}_\theta)}{2} \log n$ we have:

(a) $\hat{\theta}_{\mathrm{MLE}}$ and $\hat{\theta}_{\mathrm{MDL}}$ both $\to \theta^* = 1 - D$ w.p.1

(b) $\hat{\theta}_{\mathrm{MLE}} \neq (1 - D)$ i.o., w.p.1

(c) $\hat{\theta}_{\mathrm{MDL}} = (1 - D)$ ev., w.p.1

Example Interpretation and Details

Although artificial, the above example illustrates a general phenomenon:

"$\hat{\theta}_{\mathrm{MLE}}$ overfits whereas $\hat{\theta}_{\mathrm{MDL}}$ doesn't"

Proof. To compare the $\hat{\theta}_{\mathrm{MLE}}$ with $\hat{\theta}_{\mathrm{MDL}}$, need estimates of the "log-likelihood"

$$\log \mathbb{Q}_\theta(B(X_1^n, D))$$

with accuracy better than $O(\log n)$, uniformly in θ. This involves *very intricate* large deviations: STEPS 1 & 2:

$$\begin{aligned} -\log \mathbb{Q}_\theta(B(X_1^n, D)) &= -\log \Pr\left\{ \frac{1}{n}\sum_{i=1}^n d(X_i, Y_i) \leq D \,\middle|\, X_1^n \right\} \\ &\approx nR(\hat{P}_{X_1^n}, \mathbb{Q}_\theta, D) + \tfrac{1}{2}\log n + O(1) \quad \text{w.p.1} \\ &\approx \sum_{i=1}^n g_\theta(X_i) + \tfrac{1}{2}\log n + O(\log\log n) \quad \text{w.p.1} \end{aligned}$$

STEP 3: Implicitly identify $g_\theta(x)$ as the "derivative" a convex dual

STEP 4: Expand $g_\theta(x)$ in Taylor series around θ^*

STEP 5: Use the LIL to compare $\hat{\theta}_{\mathrm{MLE}}$ with $\hat{\theta}_{\mathrm{MDL}}$

STEP 6: *Justify* a.s.-uniform approximation $\square$

Another Example of Consistency: Gaussian Mixtures

Let: The source $\{X_n\}$ be $\mathbb{R}^k$-valued, stationary, ergodic
finite mean and covariance, arbitrary distr $\mathbb{P}$

Θ be Gaussian mixtures: $\mathbb{Q}_\theta \sim$ IID $\sum_{i=1}^{L} p_i N(\boldsymbol{\mu}_i, \boldsymbol{K}_i)$
where $\theta = \Big((p_1, \ldots, p_L), (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_L), (\boldsymbol{K}_1, \ldots, \boldsymbol{K}_L)\Big)$
k, L fixed, $\boldsymbol{\mu}_i \in [-M, M]^k$, $\boldsymbol{K}_i$ has eigenvalues in some $[\lambda, \Lambda]$

$d_n(x_1^n, y_1^n)$ = MSE, $D > 0$ fixed

Motivation: Practical quantization/clustering schemes
e.g., Gray's Gauss mixture VQ and MDI selection

Theorem: There is a "unique" optimal θ^* characterized by
a k-dimensional variational problem,

$$\inf_{\theta \in \Theta} R(P_k, Q_{\theta,k}, D) = R(P_k, Q_{\theta^*,k}, D),$$

and $\hat{\theta}_{\text{MLE}}$, $\hat{\theta}_{\text{MDL}}$ both $\to \theta^*$ w.p.1 and in L^1

Weak Consistency: A General Theorem

Given our present setup: $\{X_n\}$, Θ, and a $d_n(x_1^n, y_1^n)$

Suppose

(a) The generalized AEP holds for all $\mathbb{Q}_\theta$ in Θ

(b) The rate function $R(\mathbb{P}, \mathbb{Q}_\theta, D)$ of the AEP is lower semicont's in θ

(c) The lower bound of the AEP holds uniformly on compacts in Θ

(d) $\hat{\theta}$ stays in a compact set, eventually, w.p.1

Then

$$\mathrm{dist}(\hat{\theta}_{\mathrm{MLE}}, \{\theta^*\}) \to 0 \quad \text{w.p.1}$$
$$\mathrm{dist}(\hat{\theta}_{\mathrm{MDL}}, \{\theta^*\}) \to 0 \quad \text{w.p.1}$$

Proof. Based on epiconvergence (or Γ-convergence); quite technical □

Conditions.

(a) we saw; (c) often checked by Chernov bound-like arguments;

(b) and (d) need to be checked case-by-case

$\rightsquigarrow$ These sufficient conditions can be *substantially* weakened

Strong Consistency: A Conjecture

Under conditions (a)–(d) above:

Conjecture

For "smooth enough" parametric families, under regularity conditions:

$$\textit{always}: \qquad \dim(\hat{\theta}_{\mathrm{MDL}}) = \dim(\theta^*) \quad \text{ev., w.p.1}$$

$$\textit{typically}: \qquad \dim(\hat{\theta}_{\mathrm{MLE}}) \neq \dim(\theta^*) \quad \text{i.o., w.p.1}$$

Proof ?! Saw the "brutal" technicalities in simple Gaussian case;
general result is still open □

Step 7. Applications: Preprocessing in VQ Design

Remarks

So far, everything based on "measures $\leftrightarrow$ (random) codes" correspondence

Practical implications?!

Candidate Application #1: Gaussian-mixture VQ

stationary, ergodic, $\mathbb{R}^k$-valued source $\{X_n\}$

finite mean and covariance, arbitrary distr $\mathbb{P}$

Θ are Gaussian mixtures: $\mathbb{Q}_\theta \sim$ IID $\sum_{i=1}^{L} p_i N(\boldsymbol{\mu}_i, \boldsymbol{K}_i)$

$\boldsymbol{\mu}_i \in [-M, M]^k$, $\boldsymbol{K}_i$ has eigenvalues in some $[\lambda, \Lambda]$

$d_n(x_1^n, y_1^n) =$ MSE, $D > 0$ fixed

Problem: Choose L

MDL Estimate: $\hat{L} = [\ \#$ of components in $\hat{\theta}_{\mathrm{MDL}}]$

Candidate Application #2: Codebook Support

Let: The source $\{X_n\}$ be arbitrary stationary, ergodic

The reproduction alphabet $\hat{A}$ be finite

Θ : all IID measures on $\hat{A}$

$d_n(x_1^n, y_1^n) =$ "arbitrary" single-letter dist measure

Motivation:

– Covers classical (Shannon) case

– Except for IID assumption, covers "all" cases

– Since all good VQ codebooks look like they come from Q_{θ^*}, important to know the support $S \subset \hat{A}$ of Q_{θ^*} before designing VQ

$\rightsquigarrow$ *Hard* problem!

MDL Estimate:

$$\hat{S} = \text{support}(\hat{\theta}_{\text{MDL}})$$

Step 8. Implementation

Recall:

$$\hat{\theta}_{\mathrm{MDL}} \triangleq \arg\inf_{\theta \in \Theta} \Big[-\log \mathbb{Q}_\theta(B(X_1^n, D)) + \ell_n(\theta) \Big]$$

Questions

Is this calculable?

The ball $B(x_1^n, D)$ typically has exponentially many elements –
Is $\mathbb{Q}_\theta(B(x_1^n, D))$ calculable even *for one* θ?

A Quick Answer

In special cases YES, in $O(n^3)$ time
with a dynamical-programming-like algorithm

Final Remarks: The MDL Point of View

Guideline: Kolmogorov Distortion-Complexity: Not computable

Codes-Measures Correspondence: "*All codes are random codes*"

$$L_n(X_1^n) \;=\; -\log Q_n\Big(B(X_1^n, D)\Big) \quad \text{bits}$$

Optimal code: $\quad Q^* \;=\; \arg\inf_{Q_n} E[-\log(Q_n(B(X_1^n, D))]$

Generalized AEP(s):

$$-\frac{1}{n}\log Q_n(B(X_1^n, D)) \to R(\mathbb{P}, \mathbb{Q}, D)$$

Lossy MLE: consistent but overfits

$$\hat{\theta}_{\text{MLE}} \stackrel{\triangle}{=} \arg\inf_{\theta\in\Theta}\Big[-\log \mathbb{Q}_\theta(B(X_1^n, D))\Big]$$

Lossy MDL: consistent and does NOT overfit

$$\hat{\theta}_{\text{MDL}} \stackrel{\triangle}{=} \arg\inf_{\theta\in\Theta}\Big[-\log \mathbb{Q}_\theta(B(X_1^n, D)) + \ell_n(\theta)\Big]$$

VQ Design: Preprocessing with Lossy MDL reduces problem dimensionality

References

The results above can be found at:

[K&Zhang] "General source models and Bayesian codebooks in rate-distortion theory," *IEEE IT Trans*, 2002

[Dembo&K] "Source coding, large deviations, and approximate pattern matching," *IEEE IT Trans*, 2002

[Harrison&K] "An MDL proposal for lossy data compression," in preparation

available on:

`www.dam.brown.edu/people/yiannis`