Statistical Data Compression with Distortion

Mokshay Madiman

Department of Statistics, Yale University

Joint work with M. Harrison and I. Kontoyiannis

2nd EPFL-UMLV Workshop on Entropy, Lausanne

8 September 2008

Outline

- The Problem: Lossy Data Compression
- Codes as Probability Distributions
- Selecting good codes as an estimation problem
- Proposing new estimators based on “lossy likelihood”
- Consistency of proposed estimators
- MLE/MDL Dichotomy + Examples
- Comments and conclusions

The Problem: Data Compression

Data $X^n = (X_1, X_2, \ldots, X_n)$ in A^n

Quantized version $\hat{X}^n = (\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n)$ in discrete $C_n \subset \hat{A}^n$

Binary codeword for $\hat{X}^n$ is a binary string $e_n(\hat{X}^n)$ (e.g., 010010)

Goal

Find an efficient and approximate representation

$$\hat{X}^n = q_n(X^n)$$

for X^n

$\xrightarrow{q_n}$

$\xrightarrow{e_n}$

0010111010110
101101000 ...

“Efficient” and “Approximate”

Efficient

Codelength $L_n(X^n)$ is the # of bits in $e_n(\hat{X}^n)$

We wish to minimize the codelength per symbol

Approximate

Distortion function $d_n(x^n, y^n) = \frac{1}{n}\sum_{i=1}^{n} d_1(x_i, y_i)$

Examples: $A = \hat{A} = \{0, 1\}$ $d_1(x, y) = \mathbf{1}_{\{x=y\}}$

$A = \hat{A} = \mathbb{R}$ $d_1(x, y) = (x - y)^2$

We wish to keep the distortion small

“Efficient” and “Approximate”

Efficient

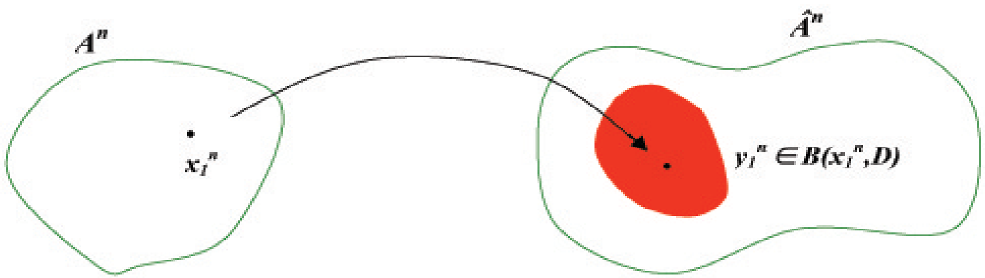
Codelength $L_n(X^n)$ is the # of bits in $e_n(\hat{X}^n)$

We wish to minimize the codelength per symbol

Approximate

Distortion function $d_n(x^n, y^n) = \frac{1}{n} \sum_{i=1}^{n} d_1(x_i, y_i)$

Distortion ball $B(x^n, D) := \left\{ y^n \in \hat{A}^n : d_n(x^n, y^n) \leq D \right\}$



A code operates at distortion level D if

$$\hat{x}^n = q_n(x^n) \in B(x^n, D) \quad \text{for } \mathbf{all} \quad x^n \in A^n$$

Reminder: Classical Estimation and Data Compression

1. *Probability Distributions correspond to (Lossless) Codes*

$$L_n(x^n) \approx -\log Q_n(x^n)$$

$\rightsquigarrow$ Maximum Likelihood is Minimum Codelength

2. *Log likelihood ratios per symbol converge to a relative entropy*

$\rightsquigarrow$ Consistency of the MLE

3. *Model too big or too small creates major problems*

$\rightsquigarrow$ Do not know which class of codes to pick

4. *Minimizing description length*

$\rightsquigarrow$ Total description requires description of the selected code

$\rightsquigarrow$ Penalized MLE also controls overfitting

Lossless data compression suggests a way to think about estimation and model selection

Lossy Codes as Probability Distributions

Recall $L_n(x^n)$ is the codelength in bits used to represent x^n

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

Why? (K&Z'02)

Let

$$Q_n(y^n) \;\propto\; \begin{cases} 2^{-L_n(y^n)} & \text{if } y^n \text{ is a codeword} \\ 0 & \text{otherwise} \end{cases}$$

Then for **all** x^n :

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

with equality if the codewords are D-separated

Random Code Construction

Construction

Given Q_n,

1. Generate a random codebook by drawing independent strings using Q_n:

$$Y^n(1) \qquad Y^n(2) \qquad Y^n(3) \qquad \dots$$

2. The quantizer maps the data X^n to the first D-close match $\hat{X}^n = Y^n(W_n)$, where

$$W_n = \min\{i \ : \ d_n(X^n, Y^n(i)) \le D\}$$

3. The encoder represents X^n by W_n written in binary

Performance

For **any** process $\{X_n\}$ and any reasonable sequence of probability distributions Q_n on $\hat{A}^n$, the code constructed in this way operates at distortion level D, and its codelength satisfies (K&Z'02)

$$L_n(X^n) \le -\log Q_n(B(X^n, D)) + 2\log n \text{ bits, eventually, w.p.1}$$

Fundamental limits and a generalized AEP

Asymptotic Equipartition Property (AEP)

If the process $\{X_n\} \sim P$ is IID, the (lossless) compression performance w.r.t any IID sequence of distributions $\{Q^n\}$ is given by

$$-\frac{1}{n}\log Q^n(X^n) \to H(P) + D(P\|Q) \quad \text{bits/symbol, as } n \to \infty\text{, w.p.1}$$

where H is entropy, and D is relative entropy or Kullback-Leibler distance

A Generalized AEP (L&S'97, Y&K'98, Y&Z'99, D&K'98)

If the process $\{X_n\} \sim \mathbb{P}$ is stationary and ergodic, and d_n is a single-letter distortion function, the compression performance w.r.t **any** sequence of "nice" distributions $\{Q_n\} = \mathbb{Q}$ is given by

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

bits/symbol, as $n \to \infty$, w.p.1

where the rate function $R(\mathbb{P}, \mathbb{Q}, D)$ is well-defined

Representations of the rate function

Information-theoretic representation

When a code based on Q is used to encode process based on P is

$$R(P,Q,D) = \inf_{W} D(W \| P \times Q),$$

where the inf is taken over all W such that $(X,Y) \sim W$ satisfies $X \sim P$ and $E\rho(X,Y) \leq D$.

Large deviations representation

$R(P,Q,D)$ is the convex dual in the last argument of

$$\Lambda(P,Q,\lambda) := E_P \left[\log E_Q e^{\lambda\rho(X,Y)} \right],$$

i.e., $R(P,Q,D) = \sup_{\lambda \leq 0} [\lambda D - \Lambda(P,Q,\lambda)]$.

What is a good code?

The IID Case

Lossless coding	Lossy coding
Want a code based on the Q_* that minimizes $H(P) + D(P\|Q)$	Want a code based on "the" Q_* that minimizes $R(P, Q, D)$
The optimal Q_* is true process distribution P	For $D > 0$, optimal Q_* achieves Shannon's r.d.f. $R(P, D) = \inf_Q R(P, Q, D)$
Selecting a good code is like estimating a process distribution from data	Selecting a good code is an indirect estimation problem

Goal: Restated

Approximate the performance of the optimal coding distribution Q_*,

i.e., find Q that yields code-lengths

$$L_n(X^n) = -\log Q^n\big(B(X^n, D)\big) \quad \text{bits}$$

close to those of the optimal "lossy Shannon code":

$$L_n^*(X^n) = -\log Q_*^n\big(B(X^n, D)\big) \quad \text{bits}$$

Coding with P known

Suppose the data X_1^n is IID, and its distribution P is known.

Let $\tilde{Q}_n$ achieve

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

Then a code based on $\tilde{Q}_n$ (K&Z'02)

- is competitively optimal
- asymptotically achieves the rate $R(P, D) \stackrel{\triangle}{=} \inf_Q R(P, Q, D)$
- no other code can have a better rate
- asymptotically behaves like a code based on Q_*^n , where

$$R(P, D) = R(P, Q_*, D)$$

Compression and Statistics

Our problem is code selection, not estimating a true distribution

Yet we observe:

Code (L_n)	Probability distribution (Q_n)
Classes of codes	Statistical models $\{\mathbb{Q}_\theta : \theta \in \Theta\}$
Code selection	Estimation : find optimal $\theta^* \in \Theta$ (i.e., one which minimizes $R(P, Q_\theta, D)$)
Code class selection	Model selection

Coding with Unknown P

Definition

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

The lossy likelihood is $Q_\theta^n(B(X^n, D))$ (NOT like a traditional likelihood!)

The lossy version of the negative log likelihood function is

$$LL(\theta; X^n) = -\log Q_\theta^n(B(X^n, D))$$

An equivalent notion

The codelength can be approximated using the empirical distribution $\hat{P}_{X^n}$ of the data (D&K'98, Y&Z'98, M&K'04) :

$$-\log Q_\theta^n(B(X^n, D)) = nR(\hat{P}_{X^n}, Q_\theta, D) + \frac{1}{2}\log n + O(1) \quad \text{eventually w.p.1}$$

This suggests that the empirical rate function

$$\hat{R}(\theta; X^n) = nR(\hat{P}_{X^n}, Q_\theta, D)$$

can be used in place of $LL(\theta; X^n)$

mile-marker

What we have:

$\rightsquigarrow$ A characterization of the optimal coding distribution Q_{θ^*} as that achieving

$$\inf_{\theta \in \Theta} R(P, Q_\theta, D)$$

$\rightsquigarrow$ A notion (in fact, two) of lossy likelihood for parametric families of codes / distributions

What we want:

$\rightsquigarrow$ Ways to estimate θ^*

What can we learn from classical theory?

$\rightsquigarrow$ Maximum likelihood and related ideas

The MALL and SMALL Estimators

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

Definitions

The MAximum Lossy Likelihood (MALL) and pSeudo-MALL (SMALL) estimators are

$$\hat{\theta}_n^{\text{MALL}} \equiv \arg\min_{\theta \in \Theta}[-\log Q_\theta(B(X^n, D))]$$

$$\tilde{\theta}_n^{\text{SMALL}} \equiv \arg\min_{\theta \in \Theta} R(\hat{P}_{X^n}, Q_\theta, D)$$

The MALL and SMALL Estimators

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

Definitions

The MAximum Lossy Likelihood (MALL) and pSeudo-MALL (SMALL) estimators are

$$\hat{\theta}_n^{\mathsf{MALL}} \equiv \arg\min_{\theta \in \Theta} [-\log Q_\theta(B(X^n, D))]$$

$$\tilde{\theta}_n^{\mathsf{SMALL}} \equiv \arg\min_{\theta \in \Theta} R(\hat{P}_{X^n}, Q_\theta, D)$$

The MALL/SMALL estimators are nice...

The MALL and SMALL estimators are consistent in great generality:

Theorem 1: Under weak conditions, as $n \to \infty$,

$$\hat{\theta}_n^{\mathsf{MALL}} \longrightarrow \theta^* \quad \text{w.p.1}$$

$$\tilde{\theta}_n^{\mathsf{SMALL}} \longrightarrow \theta^* \quad \text{w.p.1}$$

Consistency: Comments on Proof

Key Idea

A uniform, second-order expansion of the empirical rate function:

$$nR(\hat{P}_{X^n}, Q_\theta, D) = nR(P, Q_\theta, D) + \sum_{i=1}^{n} g(X_i) + O(\log\log n)$$

eventually w.p.1, uniformly in θ

Comments

– Very fine large deviation estimates

– Uses a uniform LIL (A&T'78), based on VC theory

– Technically very hard

– This approach works for IID case; an even more abstract approach yields even more general results

The MALL and SMALL Estimators

The MALL/SMALL estimators are nice...

The MALL and SMALL estimators are consistent in great generality

But Problems with MALL/SMALL

- Overfitting
- Not real codes

Lossy MDL Estimators

Definitions

The Lossy Minimum Description Length (LMDL) and the pSeudo Lossy Minimum Description Length (SLMDL) Estimators are

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

$$\tilde{\theta}_n^{\mathsf{SLMDL}} \equiv \arg\min_{\theta\in\Theta}[nR(\hat{P}_{X^n}, Q_\theta, D) + \ell_n(\theta)]$$

where $\ell_n(\theta) = o(n)$ is a given "penalty function"

Lossy MDL Estimators

Definitions

The Lossy Minimum Description Length (LMDL) and the pSeudo Lossy Minimum Description Length (SLMDL) Estimators are

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

$$\tilde{\theta}_n^{\text{SLMDL}} \equiv \arg\min_{\theta\in\Theta}[nR(\hat{P}_{X^n}, Q_\theta, D) + \ell_n(\theta)]$$

where $\ell_n(\theta) = o(n)$ is a given "penalty function"

LMDL/SLMDL are nice...

The LMDL and SLMDL estimators are consistent in great generality:

Theorem 2: Under weak conditions, as $n \to \infty$,

$$\hat{\theta}_n^{\text{LMDL}} \to \theta^* \quad \text{w.p.1}$$

$$\tilde{\theta}_n^{\text{SLMDL}} \to \theta^* \quad \text{w.p.1}$$

Do LMDL/SLMDL solve the problems of MALL/SMALL?

Illustration: Gaussian example

Consider IID coding distributions $Q_\theta \sim N(0,\theta)$, $\theta \in (0,\infty)$, and the penalty function

$$\ell_n(\theta) = \begin{cases} 0 & \text{if } \theta = \theta_0 \\ \frac{1}{2}\log n & \text{if } \theta \neq \theta_0 \end{cases}$$

where the lower-dimensional set $\{\theta_0\} \subset (0,\infty)$ is declared to be our "preferred" set

If $P \sim N(0,\sigma^2)$ and $d_1(x,y) = (x-y)^2$ then optimal $Q_* \sim N(0,\theta^*)$, with

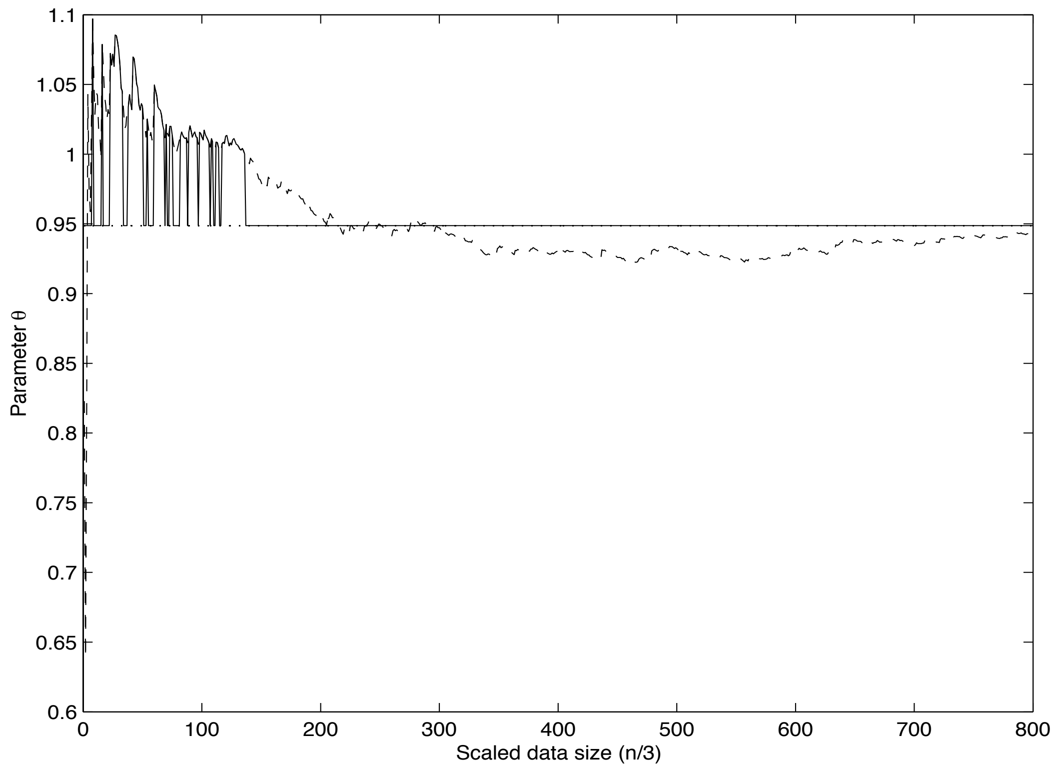
$$\theta^* = \sigma^2 - D$$

If θ^* is indeed in our preferred set (i.e., $\theta^* = \theta_0$), we wish to know it in finite time

Illustration: Gaussian example (contd.)

E.g. $\sigma^2 = 1\,,\ D = 0.05$

Under the null hypothesis that $\theta^* = \theta_0\,,$



Dotted $= \{\theta = \theta^*\}$, Dashed = SMALL estimator, Solid = SLMDL estimator

Nested Discrete Parametric Families

Setting

- Source distribution P takes values in a finite alphabet A
- Θ parametrizes the simplex of all IID probability distributions on $\hat{A} = A$
- Arbitrary single-letter distortion function

Complexity

- Suppose $L_1 \subset L_2 \subset ... \subset L_s \subset \Theta$ are increasingly complicated "models", and $k_1 < k_2 < \ldots < k_s = k_{\max}$ are the corresponding complexity coefficients
- Preference for simpler models is expressed by using the penalty

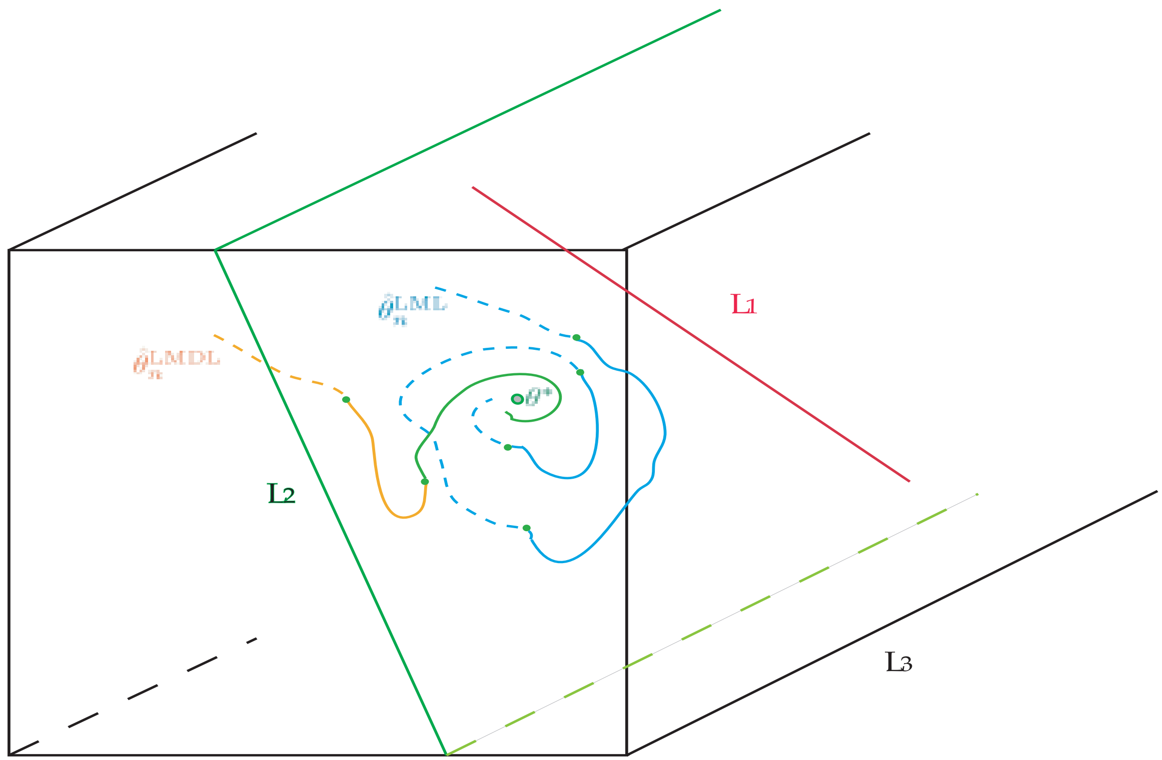
$$\ell_n(\theta) = k(\theta) \log n$$

where

$$k(\theta) \equiv \min\{k_i : \theta \in L_i\}$$

is the index of the simplest L_i containing θ

Lossy MDL works



Theorem 3: Under reasonable restrictions on P and if $k(\theta^*) < k_{\max}$,

1. $\tilde{\theta}_n^{\mathsf{SMALL}} \notin L_{k(\theta^*)}$ i.o. w.p.1
2. $\tilde{\theta}_n^{\mathsf{SLMDL}} \in L_{k(\theta^*)}$ eventually w.p.1
3. $\hat{\theta}_n^{\mathsf{LMDL}} \in L_{k(\theta^*)}$ eventually w.p.1

Model Identification: Outline of Proof

Step 1. Let $Q_{\theta^*(\beta)}$ be the optimal coding distribution for P_β

Then $\tilde{\theta}_n^{\text{SMALL}} = \theta^*(\hat{\beta})$

Step 2. $\theta^*(\hat{\beta}) - \theta^*(\beta)$ is Taylor expanded, justified by repeated uses of Implicit Function Theorem

Step 3. Multivariate LIL is applied to obtain

$$[\tilde{\theta}_n^{\text{SMALL}} - \theta^*]_j \approx \sqrt{\frac{\log\log n}{n}} \quad \text{for each coordinate } j$$

This gives Part 1: "SMALL fluctuates forever"

Step 4. A Taylor expansion of $\hat{R}(\theta) = nR(\hat{P}_{X^n}, Q_\theta, D)$ gives

$$\hat{R}(\theta^*) - \hat{R}(\tilde{\theta}_n^{\text{SMALL}}) \approx \log\log n \text{ eventually w.p.1}$$

Step 5. A sample path argument yields Part 2; approximation yields Part 3

Remarks

- Our estimator "finds" the optimal model class in finite time with any penalty function of form $k(\theta)c(n)$, as long as

$$c(n) = o(1) \quad \text{and} \quad \frac{\log\log n}{c(n)} = o(1)$$

- Penalty of form $\frac{k(\theta)}{2}\log n$ has total description length motivation
- Analogous to the findings of Hannan–Quinn '79 and Rissanen in classical estimation / lossless coding context
- State-of-the-art algorithms for compression (such as Gray's Gaussian mixture vector quantizers) have associated model selection problems
- The idea of lossy MDL has been used for clustering by MDHW '07 and YWMS '08
- The plug-in estimator for Shannon's r.d.f. $R(P, D)$ is seen to be accurate
- These results are initial illustrations; the ideas are very general

Conclusions

- We proposed maximum likelihood and MDL-type estimators for the purpose of finding good lossy codes
- These estimators are consistent (i.e., they eventually yield optimal codes)
- Lossy MDL has better code selection properties than lossy MLE
- Theoretical framework for lossy coding via its statistical interpretation

○ — ○ — ○

Extras

Lossy MDL Proof (details)

Step 5. The sample path argument:

Let

$$l(\theta) = \hat{R}(\theta) + k(\theta) \log n$$

be the "description length" that is minimized to obtain SLMDL estimator

For n such that $k(\tilde{\theta}_n^{\text{SMALL}}) \leq k(\theta^*)$,

$$k(\tilde{\theta}_n^{\text{SLMDL}}) \leq k(\tilde{\theta}_n^{\text{SMALL}}) \leq k(\theta^*)$$

For n such that $k(\tilde{\theta}_n^{\text{SMALL}}) > k(\theta^*)$,

$$\begin{aligned} l(\tilde{\theta}_n^{\text{SLMDL}}) &\leq l(\theta^*) \\ &< \hat{R}(\tilde{\theta}_n^{\text{SMALL}}) + \delta \log n + k(\theta^*) \log n \\ &\leq \hat{R}(\tilde{\theta}_n^{\text{SLMDL}}) + [k(\theta^*) + \delta] \log n \end{aligned} \tag{1}$$

so that $\quad k(\tilde{\theta}_n^{\text{SLMDL}}) < k(\theta^*) + \delta \quad$ eventually w.p.1

Additional Comments

Why not estimate P first and then use Q^* for that P?

- Goal is to finding good code from available family, Q^* may not be in family
- Optimal coding distribution may not be a continuous function of P
- $R(P, D)$ very hard to compute, let alone $Q^*(P, D)$