Second-order properties of lossy likelihoods and the MLE/MDL dichotomy in lossy compression

Mokshay Madiman and Ioannis Kontoyiannis
Division of Applied Mathematics
Brown University.

May 2004
Revised September 2005

Abstract

This paper develops a theoretical framework for lossy source coding that treats it as a statistical problem, in analogy to the approach to universal lossless coding suggested by Rissanen's Minimum Description Length (MDL) principle. Two methods for selecting efficient compression algorithms are proposed, based on lossy variants of the Maximum Likelihood and MDL principles. Their theoretical performance is analyzed, and it is shown under appropriate assumptions that the MDL approach to universal lossy coding identifies the optimal model class of lossy codes.

[1]Mokshay Madiman is with the Department of Statistics, Yale University, 24 Hillhouse Avenue, New Haven, CT 06511, USA. Email: mokshay.madiman@yale.edu

[2]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

Contents

1 Introduction **3**
- 1.1 The Lossy Compression Problem and Its Formulation . . . 3
- 1.2 A Solution Paradigm: Codes and Measures . . . 5
- 1.3 Outline . . . 9

2 Principles and Main Results **9**
- 2.1 Conventions and Notation . . . 9
- 2.2 Likelihood-based lossy coding principles . . . 11
- 2.3 Main Results . . . 15

3 Second-order properties of the lossy likelihood **18**
- 3.1 Uniform Approximations of the Lossy Likelihood . . . 18
- 3.2 Implications . . . 19

4 Three Examples: Lossy MDL vs. Lossy Maximum Likelihood **20**
- 4.1 Gaussian codes . . . 20
- 4.2 Bernoulli case . . . 24

5 The LML/LMDL Dichotomy for i.i.d. finite-alphabet codebooks **26**
- 5.1 The admissible class of sources . . . 26
- 5.2 Behavior of the pseudo-LML estimator . . . 29
  - 5.2.1 Parameters . . . 29
  - 5.2.2 Fluctuations . . . 31
  - 5.2.3 Rates . . . 32
- 5.3 Behavior of the pseudo-LMDL estimator . . . 34
- 5.4 The LMDL estimator . . . 35

6 Conclusion **37**

A Proof of Theorem 3 **37**
- A.1 Part 1 . . . 38
- A.2 Part 2 . . . 39
- A.3 Part 3 . . . 40

B Connections to the Method of Types **42**
- B.1 Background . . . 42
- B.2 The second-order generalized AEP using types . . . 43

C Remarks on Asymptotic Normality **44**

Bibliography **46**

1 Introduction

1.1 The Lossy Compression Problem and Its Formulation

Information theory, classically, has had two primary goals: source coding (efficient data compression) and channel coding (communicating reliably over a noisy channel). In the approximately half century since Shannon's fundamental work on both these subjects, a tremendous amount of progress has been made in both areas, in terms of theory ("Shannon theory") as well as practice ("coding theory"). In particular, the fundamental theory for source coding with a fidelity criterion (alternatively, "lossy" data compression) is well-developed and pleasing, and there exist sophisticated algorithms to perform lossy compression of various kinds of data (audio formats such as MP3, image formats such as JPEG, and so on). However, the bond between the theoretical and practical work has not been as strong as one might expect. In particular, the algorithms available today are based more on engineering intuition and experimentation than on fundamental theoretical principles; they are extremely ingenious and useful, but are still typically far from the optimal performance expected theoretically.

The objective of lossy source coding is to encode the data in such a way as to be maximally compressed (occupy the least amount of "space") and yet enable recovery of the data to within an allowable distortion level D. We will follow the traditional[1] practice of modelling a source by a stochastic process $\{X_n\}_{n\geq 1}$ (whose realization is the "data"). The optimal performance that can be achieved was described by Shannon in 1959 through the rate-distortion function $R(D)$, which characterizes the optimal coding rate at a given distortion level. The Blahut-Arimoto algorithm and its generalizations[21] allow fairly efficient computation of rate-distortion functions for specific problems, but there is little indication today of any principles that can be used to construct real codes that achieve it.

Shannon's approach as well as subsequent work on the problem until the 1980's, however, was based on a key premise: that the probability distribution of the stochastic source was known. In most real-life situations, such a premise would not hold. This suggests the practically important problem of *universal* data compression- where the objective is to select a coding scheme in order to obtain good compression performance *when the source distribution is not completely known*. The answer to this question is still very unclear. In this work, we propose and develop a new theoretical framework for the problem of universal lossy data compression.

More precisely, consider a source[2] $\{X_n\}$ with values in the alphabet A, which is to be compressed with distortion no more than $D \geq 0$ with respect to an arbitrary sequence of distortion functions[3] $\rho_n : A^n \times \hat{A}^n \to [0, \infty)$, where $\hat{A}$ is the reproduction alphabet. Let $B(x_1^n, D)$ denote the distortion-ball of radius D around the source string $x_1^n \in A^n$:

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

Recall from the theory of lossless coding that a prefix-free encoder is a lossless code whose output can be uniquely decoded because no codeword is a prefix of any other codeword.

Definition 1. *A D-semifaithful code or lossy code operating at distortion level D (or simply, a lossy code) is a sequence of maps $C_n : A^n \to \{0,1\}^*$ satisfying the following conditions:*

[1] Other modelling frameworks have been suggested, e.g., the Kolmogorov complexity approach, the "individual sequence" approach pioneered by Ziv, and the grammar-based approach of Yang and Kieffer.

[2] A source is just any discrete-time A-valued stochastic process; alternatively, a probability measure on the sequence space of the alphabet A.

[3] In the information theory literature, a distortion function is often called a "distortion measure" or a "fidelity criterion".

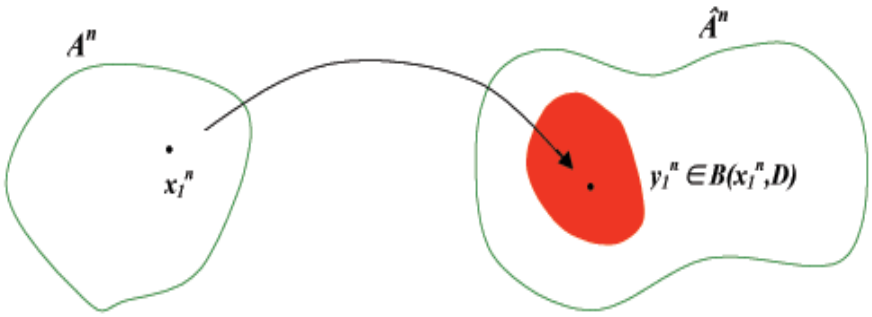

Figure 1: A D-semifaithful code quantizes the data to a point in the distortion ball around the data, and then represents this point by a binary string.

(i) C_n *is the composition* $\psi_n \circ \phi_n$ *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* $\psi_n : B_n \to \{0,1\}^*$.

(ii) $\rho_n(x_1^n, \phi_n(x_1^n)) \leq D$ *for all* $x_1^n \in A^n$.

We make some comments on the choice of the distortion functions ρ_n. Firstly, our results share with most previous work in this area the feature that the distortion functions are assumed to be given, somehow fixed *a priori* by the nature of the specific application. Since we do not assume a particular form of the distortion functions, this makes the framework flexible and general.[4]Secondly, we assume for ease of analysis that we are dealing with single-letter distortion functions. This means that the ρ_n simply measure the average bitwise distortion according to $\rho_1 = \rho$:

$$\rho_n(x_1^n, y_1^n) = \frac{1}{n} \sum_{i=1}^{n} \rho(x_i, y_i).$$

Thirdly, when $A = \hat{A}$ (which is a common situation), any "nice" distortion function has the property that $\rho(x, y) = 0$ iff $x = y$. Thus, we should expect to recover results from the theory of universal lossless compression when we consider the case $D = 0$.

Definition 2. *The codelength function is the length of the code word used to encode a data string:*

$$len_n(x_1^n) = length\ of C_n(x_1^n), \quad in\ bits.$$

Given the source distribution $\mathbb{P}$ *on the sequence space of* A*, the rate of the code is*

$$R' = \operatorname*{ess\,sup}_{\omega} \limsup_{n \to \infty} \frac{1}{n} len_n(X_1^n).$$

The operational rate-distortion function of the stationary, ergodic source $\mathbb{P}$ *is the smallest rate that can be achieved by a* D*-semifaithful code in compressing the source:*

$$R(\mathbb{P}, D) = \inf \left\{ R' : \begin{array}{c} \exists\ lossy\ code\ \{C_n\}\ operating\ at\ level\ D, \\ which\ compresses\ source\ P\ at\ rate\ R' \end{array} \right\}.$$

[4]One may hope that the structure in the data itself could somehow suggest what the natural choice of distortion function should be, but that is an open problem still.

When $\mathbb{P}$ is i.i.d. with marginal distribution P, we write $R(P,D) = R(\mathbb{P},D)$.

Remark 1. Lossy codes can be defined in two ways that are dual to each other in some ways: as distortion-constrained codes or rate-constrained codes. Our framework is based on distortion-constrained codes. Rate-constrained codes require the sequence of maps $\{C_n\}$ to have a rate $\leq R$, and the goal is to minimize the distortion (typically, the expected distortion); they are not treated in this work.

The definition of a lossy code here differs from the definition used by Shannon[27] and in texts such as [7]. The difference lies in the fidelity requirement: whereas the classical approach is to ask for $E\rho_n(X_1^n, Y_1^n) \leq D$, we demand the more stringent requirement that the distortion between *any* string and its quantized version is not more than D. It is now well-known, based on the work of Kieffer[14], that this does not change the first-order asymptotics of the problem; in particular, Shannon's rate-distortion function characterizes the fundamental achievable limit for either of these fidelity constraints as long as the source is stationary and ergodic. For simplicity, we only state the Rate-Distortion Theorem for i.i.d. sources.

Fact 1. [Rate-Distortion Theorem] *The operational rate distortion function for an i.i.d. source with marginal P is given by the solution of the nonlinear optimization problem:*

$$R(P,D) = \inf_{W \in \mathcal{W}_{P,D}} I(X;Y), \tag{1}$$

where $\mathcal{W}_{P,D} = \{W \in \mathcal{P}(A \times \hat{A}) : \text{ first marginal } W_1 = P, \text{ and } E\rho(X,Y) \leq D\}$. This function is known as the rate-distortion function.

Let us make precise the notion of a universal lossy code. Since we want this to be optimal irrespective of the source, we have the definition below.

Definition 3. *Let $\mathcal{C}$ be a class of stationary, ergodic sources. A lossy code C_n is said to be universal over the class $\mathcal{C}$ if*

$$\limsup_{n\to\infty} \frac{1}{n} len_n(X_1^n) \leq R(\mathbb{P},D) \quad w.p.1 \qquad \forall \mathbb{P} \in \mathcal{C},$$

when the true distribution of $\{X_n\}$ is $\mathbb{P} \in \mathcal{C}$.

1.2 A Solution Paradigm: Codes and Measures

The fact that $D = 0$ corresponds to lossless compression suggests that one may try to take inspiration from the well-developed theory of universal lossless data compression. The key idea underlying this theory is the correspondence between codes and measures that was already implicit in [26], and put on a firm foundation by Kraft[18] for prefix-free codes and McMillan[22] for uniquely decodable codes. This is the fact that any uniquely decodable lossless code (when, say, coding with blocks of length n) has codelength function bounded below by $-\log Q_n(x_1^n)$ for some probability distribution Q_n on $\hat{A}^n$; conversely, given any Q_n, one can find a prefix-free lossless code whose codelength function is bounded above by $-\log Q_n(x_1^n) + 1$. See, e.g., [7] for details. Since the integer constraint is irrelevant when coding with large blocks, the Kraft-McMillan inequality can be paraphrased as: "There is a correspondence between lossless codes of block length n and probability distributions on A^n, given by $\text{len}_n(x_1^n) = -\log Q_n(x_1^n)$."

Kontoyiannis and Zhang [17] showed that this idea can be generalized to lossy compression by identifying lossy compression algorithms with probability distributions on the reproduction space.

Fact 2. [Codes–Measures Correspondence] *Suppose for given $D \geq 0$, the Weak Quantization Condition of [17] holds (i.e., there exists a sequence of measurable, D-semifaithful quantizers with countable range). For any code C_n operating at distortion level D, there is a probability measure Q_n on $\hat{A}^n$ such that*

$$len_n(x_1^n) \geq -\log Q_n(B(x_1^n, D)) \text{ bits, for all } x_1^n \in A^n. \tag{2}$$

Further, if $\{Q_n\}$ is an admissible sequence of probability measures in the sense that

$$\limsup_{n\to\infty} -\frac{1}{n}\log Q_n(B(X_1^n, D)) \leq R < \infty \text{ w.p.1}, \tag{3}$$

then there is a sequence of codes $\{C_n\}$ operating at distortion level D whose length functions satisfy

$$len_n(X_1^n) \leq -\log Q_n(B(X_1^n, D)) + \log n + O(\log\log n) \text{ bits, eventually w.p.1.} \tag{4}$$

A similar result holds in expectation.

Remark 2. Fact 2 is very general, and holds for *any* sequence of distortion functions. However, our study crucially depends on the single-letter assumption.

Remark 3. Note that when A is finite and $A = \hat{A}$, the Weak Quantization Condition of [17] is trivially satisfied for any $D \geq 0$ since the identity quantizer is measurable and has zero distortion.

Fact 2 outlines precisely the nature of the correspondence between lossy compression algorithms using a block length of n and probability distributions Q_n on $\hat{A}^n$. The first part is proved using Kraft's inequality. The second direct coding part was proved by [17] using a random coding argument– one estimates the waiting time for a match of X_1^n within distortion D, looking through a codebook $\{Y_1^n(i)\}_{i\in\mathbb{N}}$ whose code words are generated independently from the probability distribution Q_n. Note that this random coding procedure is not practically constructive– since the waiting times in order to identify the code word corresponding to the data is exponential in the data size.

In the lossless case, the codes–measures correspondence suggested a correspondence of codelengths with $-\log Q_n(x_1^n)$. In the lossy case, codelengths correspond to quantities of the form

$$L_n(Q_n, x_1^n) = -\log Q_n(B(x_1^n, D)) \text{ bits.} \tag{5}$$

Unlike in the lossless case, the correspondence between lossy codes of block length n and probability measures on the n-th order product of the reproduction space is only valid when coding with large blocks.

Given a lossy code, how does one evaluate how good it is? The figure of merit is naturally the codelength per symbol, and thanks to the codes–measures correspondence, this is asymptotically equivalent to the rate of exponential growth of the probabilities $Q_n(B(X_1^n, D))$. For ease of analysis, we will only consider lossy codes corresponding to product distributions on $\hat{A}^n$; thus $Q_n = Q^n$. The asymptotic performance of such codes is described by the Generalized or Lossy AEP (so called because it is a generalization of the Asymptotic Equipartition Property, or AEP, known to information theorists and statistical physicists).

Fact 3. [Generalized or Lossy AEP] *Let $\{X_n\}$ be an i.i.d. source. For any $D \geq 0$,*

$$\liminf_{n\to\infty} -\frac{1}{n}\log Q^n(B(X_1^n, D)) = R(P, Q, D) \text{ w.p.1}, \tag{6}$$

where $R(P, Q, D)$ is a function, lower semicontinuous in Q and non-increasing in D, that is defined in the next section. Furthermore, the limit exists for all $D \neq D_{\min}(P, Q)$, where $D_{\min}(P, Q)$ is defined in the next section.

Remark 4. The Lossy AEP holds in great generality. In fact, necessary and sufficient conditions for the limit to exist are derived in [12], for single-letter distortion functions on abstract (Borel space) alphabets, provided the source is stationary and ergodic, and the coding distributions $\{Q_n\}$ form a stationary process that is sufficiently strongly mixing. Only for $D = D_{\min} > 0$ is there sometimes a problem. These considerations will not bother us because all the cases we deal with have $D_{\min} = 0$.

Note that the lossy AEP proves the admissibility of all reasonable probability distributions on sequence space (including all i.i.d. distributions, and all stationary, ergodic distributions satisfying certain mixing conditions), so that the codes–measures correspondence is true in wide generality.

For convenience of discussion, we henceforth *restrict our consideration to i.i.d. source and reproduction distributions*. The principles proposed in Section 2 can be stated in the much more general setting of stationary and ergodic distributions, but the task of verifying that the principles work in general is formidable and not attempted here.

The lossy AEP identifies $R(P, Q, D)$ as the figure of merit when doing lossy coding with large blocks, suggesting that the best codes may correspond to minimizers of this rate function. We write

$$Q^* = Q^*_{P,D} = \arg\min_Q R(P, Q, D),$$

when a minimizer exists and is unique, and call it the *optimal reproduction or coding distribution*. As expected from the rate-distortion theorem,

$$R(P, D) = R(P, Q^*, D) = \min_Q R(P, Q, D)$$

is the best achievable lossy compression rate.

The connection of Q^* with optimal codes for a *finite* but large block length n was made precise in [16] and [17]. Recall that for $D = 0$, the problem of optimal lossless compression using block length n is, theoretically at least, equivalent to finding a probability distribution Q_n that minimizes the average codelength $E[-\log Q_n(X_1^n)] = H(P_n) + D(P_n\|Q_n)$. Thus the optimal choice is simply to take Q_n to be the n-th order marginal P_n of the true source distribution $\mathbb{P}$. When the source is i.i.d., this choice of codelength– $-\log P(x)$ for each symbol– has a special name, the *idealized Shannon codelength*. In [3], Barron proved the celebrated lemma on its competitive optimality properties. Thus, the problem of finding the optimal lossless codelength function (and hence the optimal code[5]) is identical to the problem of finding the distribution of a data source, which is where statistics comes in. How does this development carry over to the lossy case, when $D > 0$?

[16] shows that the sequence of probability distributions $\tilde{Q}_n$ which achieve the infima in the definitions of $K_n(D) = \inf_{Q_n} E[-\log Q_n(B(X_1^n, D))]$ corresponds to an optimal lossy

[5]When A is finite, there is a simple constructive procedure to build a lossless code whose codelength is within 1 bit of the idealized Shannon codelength. See, e.g., [7][pg.123].

Lossless coding	Lossy coding
Any code has codelength close to $-\log Q_n(x_1^n)$ for some Q_n	Any code has codelength close to $-\log Q_n(B(x_1^n, D))$ for some Q_n
AEP: $-\frac{1}{n}\log Q^n(X_1^n) \to H(P) + D(P\|Q)$	Lossy AEP: $-\frac{1}{n}\log Q^n(B(X_1^n, D)) \to R(P, Q, D)$
Want code based on the Q^* that minimizes $H(P) + D(P\|Q)$	Want code based on "the" Q^* that minimizes $R(P, Q, D)$
Optimal Q^* is true source distribution P	For $D > 0$, optimal $Q^*(\neq P)$ achieves Shannon's r.d.f. $R(P, D)$
Selecting a good code is like estimating a source distribution from data	Selecting a good code is an indirect estimation problem

Table 1: How to select good codes?

code. In particular, the codelength $-\log \tilde{Q}_n(B(X_1^n, D))$ has competitive optimality properties as for lossless compression. Furthermore, when the source is i.i.d., it suffices to consider the lossy code corresponding to powers of the optimal reproduction distribution Q^* on $\hat{A}$, i.e., the optimal output distribution that achieves the infimum in the definition of the rate-distortion function. More precisely, the difference in performance between $\tilde{Q}_n$ and $(Q^*)^n$ is $O(\log n)$ and the per-symbol difference asymptotically vanishes. Thus, when the source is memoryless with marginal P, our goal is to do statistical inference with the hope that we can somehow estimate the distribution $Q^* = Q^*_{P,D}$, and *not* the true source distribution P, from the data. In analogy with the lossless terminology, we call $-\log(Q^*)^n(B(X_1^n, D))$ the *idealized lossy Shannon codelength*.[6]

In statistical estimation, the key conceptual simplification of considering parametric models (families of probability distributions, typically parametrized by a subset of Euclidean space) made early advances in the field possible. The consideration of parametric models is justified in two ways: firstly, it is often reasonable to assume that our distributions have some structure (e.g., the observed data is generated by a deterministic process perturbed by Gaussian noise of unknown mean and variance), and secondly, the non-parametric infinite-dimensional problem can be nearly intractable to solve and parametric families may give practically useful if not completely ideal results. For the same reasons, we also choose to focus on *parametric models of coding distributions*. In order to make this more realistic, we also allow for a nested sequence of parametric models of increasing "complexity".

Motivated by the codes–measures correspondence and the above remarks, 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[7]. Specifically, we want to choose the one whose limiting coding rate $R(P, Q_\theta, D)$ is as small as possible.

If $Q^* = Q_{\theta^*}$ happens to be in the above class, then of course $R(P, Q^*, D)$ is simply

[6] As a matter of fact, Shannon's proof of the rate-distortion theorem used a particular random codebook– the best possible random codebook generated by $(Q^*)^n$. The implications of using sub-optimal random codebooks Q^n for some $Q \neq Q^*$ remained unexplored till various authors (see, e.g., [20][30][15]) began exploring the issue in the 1990's. The primary motivation for these works was the analysis of the universal lossy coding algorithms (e.g., [28][15]) inspired by the success of the Lempel-Ziv algorithms in lossless compression.

[7] Θ itself may contain a hierarchy of subsets, that we identify as models of decreasing complexity.

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 $\theta^* = \arg\min_\theta R(P, Q_\theta, D)$. Intuitively, we think of Q_{θ^*} as *the simplest measure in the class of measures parametrized by* Θ *that can describe all the regularity in the data with accuracy no worse than* D.

1.3 Outline

The search for a good universal lossy code can, based on the above discussion, be viewed as the search for a good estimator for the optimal reproduction distribution, since the latter will yield the former at least for large data sizes. Recall that a similar connection held in the lossless case, and in that case, two of the prime methods used are the mixture method and the MDL (Minimum Description Length) method. Mixture coding for lossy compression was investigated in [17]. We investigate an MDL approach in this work.

The principles underlying our approach are described in Section 2. In Section 3, second-order properties of the lossy likelihood are studied for the setting of universal coding. Sections 4 and 5 explore the dichotomy in the behavior of lossy maximum likelihood and MDL estimators for i.i.d. sources and coding distributions. While Section 4 motivates the study using examples of Gaussian and Bernoulli codes, Section 5 proves a general result in the case of finite alphabets. Section 6 offers some suggestions for future work.

2 Principles and Main Results

2.1 Conventions and Notation

The data $\{X_n\}_{n\in\mathbb{N}}$ are drawn from an i.i.d. source with marginal distribution P on an alphabet A. Although the distribution P is not known in general, we will typically assume that it lies in some well-behaved subset of $\mathcal{P}(A)$, so that the existence and uniqueness of the optimal reproduction distribution $Q^*_{P,D}$ is assured for some range of D. Most of our main results and analysis will assume that A is finite; however since we consider two examples with $A = \mathbb{R}$ in Section 4, we allow for this more general situation in our discussion of notation below. We always assume that the reproduction alphabet $\hat{A}$ is the same as the source alphabet A, retaining the two different symbols for conceptual clarity.

The single-letter distortion function $\rho : A \times \hat{A} \to [0, \infty)$ is arbitrary. By Pinkston's lemma (see, e.g., [7][Chapter 13]), we can assume without loss of generality that $\min_{y\in\hat{A}} \rho(a, y) = 0$ for each $a \in A$.

We only consider lossy codes corresponding to probability measures on $\hat{A}^n$ that are i.i.d. Indeed, the class of marginal distributions on $\hat{A}$ ("random coding distributions" or "reproduction distributions") that we allow is a parametric model with parameter space Θ, where Θ is a compact subset of Euclidean space. A generic probability distribution from this model is denoted Q_θ. The optimal coding distribution is denoted Q^* or Q_{θ^*}, since Q^* lies in our paremetric model. Call the parametrization *sensible*, if one of the following conditionshold:

1. $\theta_m \to \theta$ implies that $Q_{\theta_m} \to Q_\theta$ in the τ-topology[8], or
2. A and $\hat{A}$ are Polish spaces with the Borel σ-algebra, $\rho(\cdot, \cdot)$ is continuous, and $\theta_m \to \theta$ implies weak convergence.

The first condition is satisfied when $\hat{A}$ is finite and the canonical parametrization or any homeomorphic reparametrization of it is used, whereas many reasonable distortion functions

for continuous alphabets (such as squared error for the real line) would satisfy the second condition.

We always denote a A-valued random variable by X, and a $\hat{A}$-valued random variable by Y, indicating which distributions they came from by subscripts on expectations when necessary. The value of D above which $R(P, D)$ is 0 is given by

$$D_{\max}(P) = \min_{y \in \hat{A}} E_P[\rho(X, y)] \tag{7}$$

The argument P may be dropped when obvious from context. Clearly, if $D > D_{\max}(P)$, the data can simply be represented by a string consisting only of the minimizing $y \in \hat{A}$ while staying within mean distortion D.

It is well-known (see, e.g., [9],[11]) that the rate function in (6) is the convex dual of the averaged cumulant generating function of the distortion:

$$R(P, Q, D) = \Lambda^*(P, Q, D) = \sup_{\lambda < 0} [\lambda D - \Lambda(P, Q, \lambda)] \tag{8}$$

where

$$\Lambda(P, Q, \lambda) = E_P\big[\log E_Q[e^{\lambda \rho(X,Y)}]\big]. \tag{9}$$

Further, as shown by these authors, the rate function can alternatively be characterized as

$$R(P, Q, D) = \inf_{W \in \mathcal{W}_{P,D}} [I(X; Y) + D(Q_Y \| Q)] \tag{10}$$

where the infimum is taken over the same class $\mathcal{W}_{P,D} = \{W \in \mathcal{P}(A \times \hat{A}) : \text{ first marginal } W_1 = P, \text{ and } E\rho(X, Y) \leq D\}$ of joint distributions that appears in the rate-distortion theorem, and Q_Y denotes the second marginal of W.

The quantity $D_{\min}$, which represents the infimum of distortion levels D at which the rate function is finite, is given by

$$D_{\min}(P, Q) = E_P[\operatorname*{ess\,inf}_{\omega} \rho(\cdot, Y(\omega))]. \tag{11}$$

When P is understood to be fixed, we abuse notation and write $D_{\min}(\theta) = D_{\min}(P, Q_\theta)$ and $D^{(n)}_{\min}(\theta) = D_{\min}(\hat{P}_{X_1^n}, Q_\theta)$, where $\hat{P}_{X_1^n}$ is the empirical distribution of the data.

Let $\lambda^*(P, Q, D)$ denote the unique achieving λ in the definition (8) of the rate function $R(P, Q, D)$. Then, if Λ' and Λ'' denote the first and second derivatives of Λ with respect to λ, the following hold:

$$\begin{aligned} R(P, Q, D) &= \lambda^* D - \Lambda(P, Q, \lambda^*) \\ \Lambda'(P, Q, \lambda^*) &= D \\ \Lambda''(P, Q, \lambda^*) &> 0 \end{aligned} \tag{12}$$

Note that this is meaningful only when $D_{\min}(P, Q) < D$. When P is fixed, we define

$$\lambda_\theta = \lambda^*(P, Q_\theta, D) \tag{13}$$

[8]The τ-topology is the topology corresponding to convergence of expectations of all bounded, measurable functions; thus it is stronger than the topology of weak convergence.

and

$$g_\theta(a) \equiv \Lambda(P, Q_\theta, \lambda_\theta) - \Lambda_a(Q_\theta, \lambda_\theta), \tag{14}$$

where $\Lambda_a(Q, \lambda) = \log E_Q[e^{\lambda\rho(a,Y)}]$.

The following sequence of constants (when finite) provide constraints on the source and coding distributions that are used in the literature:

$$d_k(P, Q) \equiv E[\rho^k(X, Y)], \tag{15}$$

where we take $X \sim P$ and $Y \sim Q$ to be independent. Note that $d_1(P, Q)$ has variously appeared in the literature as $D_{\max}(P, Q)$ (eg: [16],[17]) and as $D_{\text{av}}(P, Q)$ (eg: [9]).

For most of this chapter, $A = \hat{A}$ is finite. In this case, we make use of the canonical parametrization, which parametrizes a probability distribution on $A = \{1, \ldots, m\}$ by the masses of the first $m-1$ symbols. Clearly, this is a sensible parametrization in the sense elucidated above. Let $\Sigma \subset \mathbb{R}^{m-1}$ denote the parameter space for the canonical parametrization of the simplex $\mathcal{P}(A)$. A generic probability distribution from the simplex is denoted P_σ, and the true source distribution by $P = P_{\sigma^*}$. The empirical distribution $\hat{P}_{X_1^n} = \hat{P}_{X_1^n}$ of the data X_1^n also belongs to the simplex, and is parametrized by $\hat{\sigma_n}$. For each $D \geq 0$, the collection of source distributions

$$\mathbb{S}(D) = \{P : D < D_{\max}(P), Q^* \text{is unique, } \text{supp}(P) = A, \text{and } \text{supp}(Q^*) = \hat{A}\} \tag{16}$$

is important, and we will call it the *admissible class of sources*. We denote by $\Sigma_0 \subset \Sigma$ the set parametrizing the $\mathbb{S}(D) \subset \mathcal{P}(A)$. Since the distortion function is bounded when $A = \hat{A}$ is finite, $d_k < \infty$ for all $k \in \mathbb{N}$ and $D_{\min}(P, Q) = 0$ as long as Q has full support.

2.2 Likelihood-based lossy coding principles

In [17], Kontoyiannis and Zhang proved the universality of i.i.d. mixtures (i.e., of Bayesian codebooks).

Fact 4. [Lossy Mixture Codes are Universal] *Let $\{X_1^n\}$ be an i.i.d. source with distribution P on A. For $D \in (0, D_{\max})$, let Q^* denote the optimal reproduction distribution of P at distortion D. If a prior π has a density with respect to Lebesgue measure on the simplex that is strictly positive in a neighborhood of Q^*, and if we define the mixture distribution*

$$M_n(y_1^n) = \int_{\mathcal{P}(\hat{A})} Q^n(y_1^n) d\pi(Q), \tag{17}$$

then:

$$-\log M_n(B(X_1^n, D)) \leq -\log(Q^*)^n(B(X_1^n, D)) + o(n) \quad \text{w.p.1, as } n \to \infty \tag{18}$$

In this chapter, it is shown that not only are "lossy MDL codes" universal, but they have a remarkable model selection property that is not shared by the codes corresponding to either mixtures or to lossy maximum likelihood estimates. We expand on this statement below.

A natural way to estimate the optimal θ^* empirically is to try and minimize the idealized codelengths (5), or equivalently to maximize the probabilities $Q_\theta^n(B(X_1^n, D))$.

Definition 4. *The lossy likelihood function (or simply, the lossy likelihood) is $Q_\theta^n(B(X_1^n, D))$, viewed as a function of θ. The lossy log likelihood is $L_n(Q_\theta^n, X_1^n) = -\log Q_\theta^n(B(X_1^n, D))$, viewed as a function of θ.*

Lossy Compression	Statistical Interpretation
Code (L_n)	Probability distribution (Q_n)
Class of codes	Statistical model $\{\mathbb{Q}_\theta : \theta \in \Theta\}$
Code selection	Estimation : find optimal $\theta^* \in \Theta$ (i.e., one which minimizes $R(P, Q_\theta, D)$)
Minimizing the codelength per symbol	Lossy analog of Maximum Likelihood Estimation
Minimizing codelength of a 2-part code	Lossy analog of MDL

Table 2: Developing the statistical approach to lossy compression.

We define the Lossy Maximum Likelihood Estimate as the parameter corresponding to the reproduction distribution which maximizes the lossy likelihood.

Definition 5. *The Lossy Maximum Likelihood (LML) Estimate is defined as*

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

when the minimizer exists and is unique.

In [10], it is shown that under very general conditions this estimate is consistent as $n \to \infty$, in that it converges to θ^* with probability one.

Fact 5. [Consistency of LML Estimator] *Suppose the parametrization of the class of coding distributions is sensible (in the sense defined earlier). If $\hat{\theta_n}$ is a sequence of possibly non-unique maximizers of the lossy likelihood which is relatively compact in Θ w.p.1, then $\hat{\theta_n} \to \Theta^*$, the set of minimizers of $R(P, Q_\theta, D)$.*

Remark 5. This also holds under the general conditions mentioned in Remark 4.

But as with the classical (lossless) MLE, this $\hat{\theta}_n^{\mathrm{LML}}$ also has several undesirable properties. First, the infimum in the definition of $\hat{\theta}_n^{\mathrm{LML}}$ is not really a codelength; if we choose one of the θ's based on the data, we should also describe the chosen θ itself. Indeed, there are examples [10] where the $\hat{\theta}_n^{\mathrm{LML}}$ is *not* consistent, but it becomes consistent when appropriately modified to correspond to an actual two-part code.

Second, the MLE estimate tends 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 MLE will be a Markov (non-i.i.d.) distribution for most n.

To rectify these problems, we consider "penalized" versions of the MLE, similar to those considered in the lossless case. This is an instance of the Minimum Description Length (MDL) principle proposed and developed initially by Rissanen (see, e.g., [24][25]). For a comprehensive recent review of the applications of the MDL principle (in particular, for lossless coding), see [4].

Definition 6. *Let $\ell_n(\theta)$ be a given "penalty function" such that $\ell_n(\theta) = o(1)$. The Lossy Minimum Description Length (LMDL) Estimate is defined as*

$$\hat{\theta}_n^{LMDL} = \arg\min_{\theta\in\Theta}\left[-\frac{1}{n}\log Q_\theta^n(B(X_1^n, D)) + \ell_n(\theta)\right],$$

when the minimizer exists and is unique.

By [10][13], the LMDL estimator is also consistent. Moreover, in Section 5 we present some simple examples illustrating how the LMDL estimator avoids the common problems of the LML estimator mentioned above.

However, both the LML estimator and the LMDL estimator share a severe disadvantage– they are very hard to determine in any specific situation. This is because both involve the minimization of a functional– the probability of a ball– a complicated integral that becomes exponentially harder to compute as the dimension grows. This motivates the usage of approximations of this integral that are easier to compute: we call these pseudo-estimators. The pseudo-estimators are only valid when the class of coding distributions being considered is i.i.d; furthermore they too are not easy to calculate in general, but can be computed when the form of the rate function is known.

The approximation that suggests pseudo-estimators for the i.i.d. case is one originally suggested by Yang and Kieffer in [30], and subsequently refined and generalized in [31], [8], etc. In fact, Theorem 2 proved in this chapter is a further refinement of this result. However, for the purposes of motivating our pseudo-estimators, we only need the following fact: that for abstract alphabets and arbitrary distortion functions,

$$\log Q^n(B(X_1^n, D)) = O(1) - \tfrac{1}{2}\log n - nR(\hat{P}_{X_1^n}, Q, D) \text{ eventually w.p.1.} \tag{19}$$

This suggests that for large n, we can replace the idealized lossy Shannon codelengths $L_n(Q^n, X_1^n) = -\log Q^n(B(X_1^n, D))$ by

$$\tilde{L}_n(Q, X_1^n) = nR(\hat{P}_{X_1^n}, Q, D). \tag{20}$$

In the case of memoryless sources and coding distributions, *this length function is completely functionally equivalent to the idealized lossy Shannon codelength.* This fact is the content of Theorem 1, which is a simple observation based on [31] and [16].

Theorem 1. *For any code C_n operating at distortion level D, there is a probability measure Q on $\hat{A}$ such that*

$$len_n(X_1^n) \geq nR(\hat{P}_{X_1^n}, Q, D) - \frac{3}{2}\log n + O(\log\log n) \text{ bits, eventually w.p.1} \tag{21}$$

Conversely, suppose the Weak Quantization Condition of [17] holds at a distortion level D, and the probability measure Q on $\hat{A}$ satisfies $R(P, Q, D) < \infty$. Then there is a code $\{C_n\}$ operating at distortion level D whose length functions satisfy

$$len_n(X_1^n) \leq nR(\hat{P}_{X_1^n}, Q, D) + \frac{3}{2}\log n + O(\log\log n) \text{ bits, eventually w.p.1} \tag{22}$$

Proof. By [16][Corollary 1], for any code C_n operating at distortion level D, $\text{len}_n(x_1^n) \geq -\log(Q^*)^n(B(X_1^n, D)) - 2\log n$ eventually w.p.1, where Q^* is an optimal reproduction distribution. Combining with (19) gives the first part.

For the second part, note that the sequence of product measures $\{Q^n\}$ is admissible because $\limsup_{n\to\infty} -\frac{1}{n}\log Q^n(B(X_1^n, D)) = R(P, Q, D) < \infty$. Thus Fact 2 implies the existence of a lossy code $\{C_n\}$ operating at distortion level D with length functions satisfying

$$\begin{aligned} \text{len}_n(X_1^n) &\leq -\log Q^n(B(X_1^n, D)) + \log n + O(\log\log n) \\ &\leq nR(\hat{P}_{X_1^n}, Q, D) + \frac{3}{2}\log n + O(\log\log n) \text{ bits, eventually w.p.1} \end{aligned} \tag{23}$$

The second inequality follows from (19). □

In other words, for i.i.d. sources, just as we have the asymptotic equivalence $\text{len}_n(X_1^n) \approx L_n(Q^n, X_1^n)$ where Q^n is the distribution on $\hat{A}^n$ corresponding to C_n, so also we have $\text{len}_n(X_1^n) \approx \tilde{L}_n(Q, X_1^n)$. This asymptotic equivalence suggests the following definition.

Definition 7. *The pseudo-lossy log likelihood function (or simply, the pseudo-lossy log likelihood) is $\tilde{L}_n(Q_\theta, X_1^n) = nR(\hat{P}_{X_1^n}, Q_\theta, D)$, viewed as a function of θ.*

We can now define "pseudo-estimators" that maximize the pseudo-lossy likelihood in order to estimate θ^*.

Definition 8. *The pseudo-Lossy Maximum Likelihood (pseudo-LML) Estimate is defined as*

$$\tilde{\theta}_n^{LML} = \arg\min_{\theta\in\Theta} \tilde{L}_n(Q_\theta, X_1^n) = \arg\min_{\theta\in\Theta} R(\hat{P}_{X_1^n}, Q_\theta, D), \tag{24}$$

when the minimizer exists and is unique.

Note that the lower semicontinuity of $R(P, Q, D)$ in Q implies that the existence of a minimizer is assured whenever Θ is be compact.

Definition 9. *Let $\ell_n(\theta)$ be a given "penalty function" such that $\ell_n(\theta) = o(1)$. The pseudo-Lossy Minimum Description Length (pseudo-LMDL) Estimate is defined as*

$$\tilde{\theta}_n^{LMDL} = \arg\min_{\theta\in\Theta}[R(\hat{P}_{X_1^n}, Q_\theta, D) + \ell_n(\theta)], \tag{25}$$

when the minimizer exists and is unique.

We note that in the literature on lossless data compression, two kinds of penalties have been considered– general penalties satisfying Kraft's inequality for a lossless code on a countable parameter space Θ (so that the MDL estimate then corresponds to a "real" two-part code), and dimension-based penalties. The latter is often motivated via the former using appropriate discretizations and limiting procedures. Barron [3] obtained path-breaking results of the former style, followed by further seminal results motivated by density estimation in [5]. Unfortunately we could not find an easy generalization of these elegant results to the case of lossy compression. One way to interpret the difficulty in generalizing the universality of two-part codes involving a code on the parameter space lies in the fact that the dichotomy theorem for likelihood ratios involving stationary ergodic distributions does not carry over to a dichotomy theorem for "lossy likelihood ratios" because the distortion balls do not just contain "typical strings" for a particular ergodic probability measure.

In this work, for simplicity, we will only consider penalties of the form

$$\ell_n(\theta) = k(\theta)c(n) \tag{26}$$

where $k : \Theta \to \mathbb{Z}_+$. Thus the pseudo-LMDL estimator is

$$\tilde{\theta}_n^{\text{LMDL}} = \arg\min_{\theta\in\Theta} \left[R(\hat{P}_{X_1^n}, Q_\theta, D) + k(\theta)c(n)\right], \tag{27}$$

with the LMDL estimator given analogously. Both the complexity coefficient $k(\theta)$ and the penalty decay rate $c(n)$ can be chosen in a variety of ways; the canonical choice for $k(\theta)$ is the "dimension" of θ, in a sense that will be clarified later, and the canonical choice of $c(n)$ is $\frac{\log n}{n}$.

2.3 Main Results

The key to analysis of the lossy estimators is the lossy likelihood function. The first-order behavior of the lossy likelihood is captured by the "lossy AEP" (Fact 3), so-called since it is the lossy analogue of the traditional asymptotic equipartition property. It says that the probability of a distortion ball around the data is, w.p.1, approximately equal to $e^{-nR(P,Q,D)}$. This result is refined in [9] by computing the nature of the second order term, which is of course a polynomial factor. However, that analysis is for the situation when the source P is known. In order to adapt the power of the second-order lossy AEP for universal coding, a generalization is necessary. Our first two theorems proceed in this direction.

The following assumption is key.

> ($\star$) Let the data $\{X_n\}$ come from an i.i.d. source with marginal P on a finite alphabet A, and let $P \in \mathbb{S}(D)$ (the admissible class of sources defined earlier).

We call $\mathbb{S}(D)$ the admissible class of sources because $P \in \mathbb{S}(D)$ ensures that there is a unique Q^* in the interior of the simplex $\mathcal{P}(\hat{A})$. The restriction to the admissible class of sources is also used in [17], which also contains comments about how strong a restriction this is.

Theorem 2. [Lossy Likelihood and Empirical Rate] *Suppose Assumption ($\star$) holds. Let Θ_0 be a compact subset of the interior of Θ, where Θ provides the canonical parametrization of the set of reproduction distributions on $\hat{A} = A$. Set $d_1(P, \Theta_0) \equiv \inf_{\theta \in \Theta_0} d_1(P, Q_\theta)$. Then, for $0 < D < d_1(P, \Theta_0)$,*

$$\sup_{\theta \in \Theta_0} \left| -\log Q_\theta^n(B(X_1^n, D)) - nR(\hat{P}_{X_1^n}, Q_\theta, D) - \tfrac{1}{2}\log n \right| = O(1) \quad \text{eventually w.p.1.} \tag{28}$$

The next result connects this with the rate function evaluated at the source distribution by an appropriate expansion of $R(\hat{P}_{X_1^n}, Q_\theta, D)$.

Theorem 3. [True and Empirical Rates] *Suppose Assumption ($\star$) holds. Consider the class of i.i.d. reproduction distributions on the reproduction alphabet $\hat{A} = A$, with Θ providing the canonical parametrization. Then for each $D > 0$, there exists $\delta' > 0$ such that*

$$\sup_{\theta \in B(\theta^*, \delta')} \left| R(\hat{P}_{X_1^n}, Q_\theta, D) - R(P, Q_\theta, D) - \frac{1}{n}\sum_{i=1}^{n} g_\theta(X_i) \right| = O\left(\frac{\log\log n}{n}\right) \quad \text{eventually w.p.1.} \tag{29}$$

These uniform approximations of the lossy likelihood and pseudo-lossy likelihood are proved in Section 3. Some consequences are also discussed there.

The next three theorems analyze the behavior of the various lossy and pseudo-lossy estimators in the i.i.d., finite-alphabet context. The assumptions below are labelled for convenience.

> ($\star\star$) Let $L_1 \subset L_2 \subset \ldots \subset L_s \subset \Theta$ be any nested sequence of sets in the parameter space Θ for the simplex $\mathcal{P}(A)$ according to the canonical parametrization. Define the complexity coefficient $k(\cdot)$ by
>
> $$k(\theta) = \min\{1 \le i \le s : \theta \in L_i\}. \tag{30}$$
>
> ($\star\star\star$) Suppose the (true) dimension of L_{s^*}, where $s^* \equiv k(\theta^*)$, is strictly less than $|\hat{A}| - 1$.

Assumption ($\star\star$) is our way of formulating the problem of model selection in the context of the lossy compression problem. The nested sequence $L_1 \subset L_2 \subset \ldots \subset L_s \subset \Theta$ is to be thought of as a sequence of increasingly complex parametric models for the optimal reproduction distribution Q^* that we are trying to estimate. The preference for simpler models (arising from the fact that codes based on distributions from simpler models are more easily described) is expressed by making the penalty coefficient $k(\theta)$ strictly increasing in the order of the nesting. Thus what k essentially does is to partition the parameter space Θ– each set of the partition being the pre-image of a value in the discrete range of k. The specific values of k are unimportant but the ordering is crucial. However, for convenience and without loss of generality, we define $k(\theta)$ as in (30), where the values in the range of k are successive integers. If a form for the decay rate $c(n)$ is also given, that specification completes the definition of the penalty function $k(\theta)c(n)$, as well as of the LMDL and pseudo-LMDL estimators (as per (27)).

Our first result on lossy estimators is a negative result for the pseudo-LML estimator.

Theorem 4. [Behavior of Pseudo-LML Estimator] *Suppose Assumptions ($\star$), ($\star\star$) and ($\star\star\star$) hold, and in addition, $\eta : \Sigma_0 \to \Theta$ that yields the optimal reproduction distribution is such that the derivative matrix $D\eta(\sigma^*)$ is non-singular. Then $\tilde{\theta}_n^{LML} \notin L_{s^*}$ i.o. w.p.1.*

Recall that consistency of the LML and LMDL estimators is already guaranteed under these assumptions due to [10]. Theorem 4 is therefore saying something about the behavior of $\tilde{\theta}_n^{\mathrm{LML}}$ as it approaches θ^*, namely that it never stops overestimating the complexity of θ^*.

Theorem 5. [Behavior of Pseudo-LMDL Estimator] *Suppose Assumptions ($\star$) and ($\star\star$) hold, and assume that the decay rate $c(n)$ in the penalty function $\ell_n(\theta) = k(\theta)c(n)$ is such that $\log\log n = o(nc(n))$ and $c(n) = o(1)$. Then $\tilde{\theta}_n^{LMDL} \in L_{s^*}$ eventually w.p.1.*

Thus, the pseudo-LMDL estimator approaches Q^* eventually through codes in L_{s^*}. In other words, if there is a "nice" subset of Θ and we express our desire to know if θ^* is in the "nice" subset by choosing it as a model in the model sequence that is used to define the penalty function, then the pseudo-LMDL estimator *finds* the "nice" subset in finite time whenever θ^* does indeed belong to it. The pseudo-LML estimator, on the other hand, *cannot* display this behavior– it must make excursions outside of L_{s^*} infinitely often.

The most natural choice of the nested sequence of sets would be a sequence of hypersurfaces L_i (for instance, truncated affine subspaces) of varying dimension, where L_i has dimension i. This is the picture we had in mind when formulating the model selection problem for lossy compression, and it is in this sense that the complexity coefficient $k(\theta)$ can represent the "dimension" of θ. This picture and the associated behaviors mandated by our results are illustrated in Figure 2.

To reiterate, there may be a difference between the true Euclidean dimension of a model and its complexity coefficient assigned by k. However, the canonical example we have in mind for a class of models is a class of hypersurfaces as depicted in Figure 2, and in this case the two measures of model complexity are the same. Even in that case, θ in itself is just a vector in an $(|\hat{A} - 1|)$-dimensional space and the complexity coefficient really invokes an implicit hierarchy of sets in which we would like θ^* to be as far up as possible (because we can perform the lossy coding in an increasingly efficient manner along the hierarchy for whatever reason). The theorems stated only depend on this idea, and therefore hold for *any* nested sequence of models.

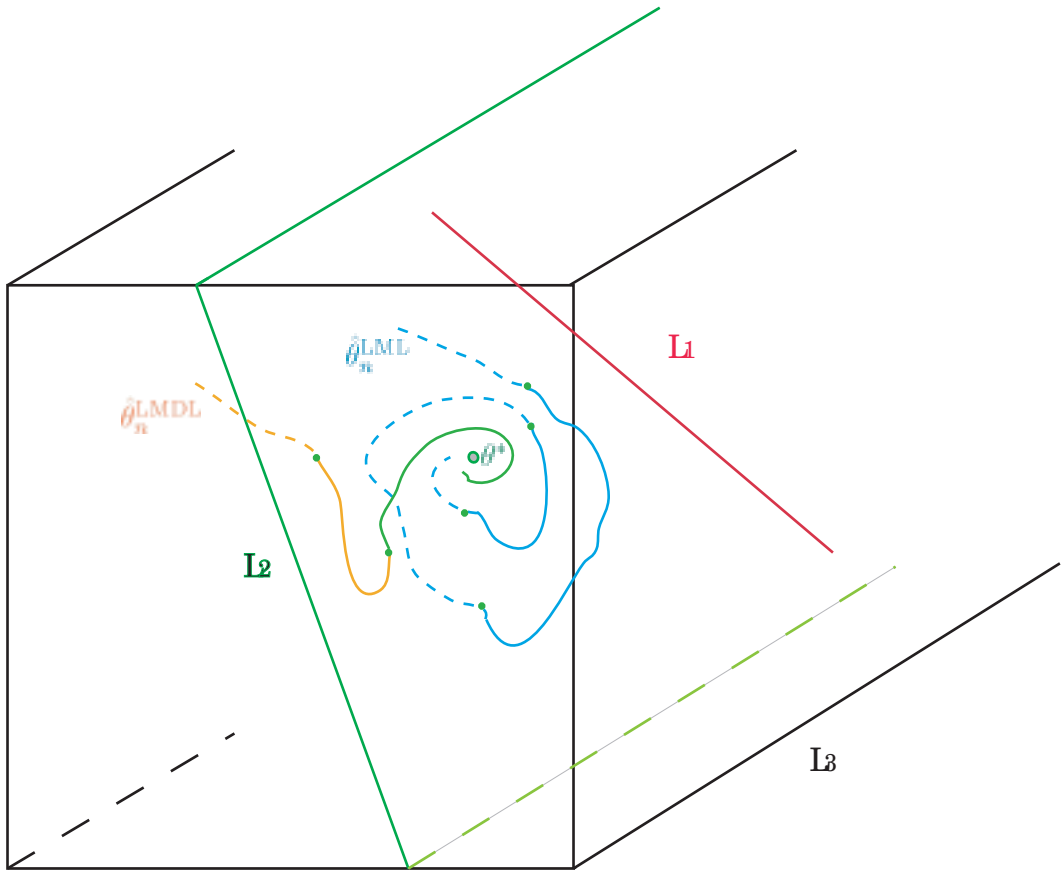

Figure 2: In this schematic figure, $\theta^* \in L_2$, the blue line represents the pseudo-LML estimator, the orange line represents the pseudo-LMDL estimator, and all points where either trajectory intersects with L_2 are marked green.

Next, the above analysis of the simpler pseudo-lossy estimators is used to study the lossy estimators themselves. It is proved in Section 5. It says that the LMDL estimator approaches Q^* eventually through codes in L_{s^*}.

Theorem 6. [Behavior of LMDL Estimator] *Suppose Assumptions ($\star$) and ($\star\star$) hold, and assume that the decay rate $c(n)$ in the penalty function $\ell_n(\theta) = k(\theta)c(n)$ is such that $\log\log n = o(nc(n))$ and $c(n) = o(1)$. Then $\hat{\theta}_n^{LMDL} \in L_{s^*}$ eventually w.p.1.*

Remark 6. The freedom in the choice of the decay rate $c(n)$ in the penalty function is remarkable. It tells us, for example, that the exact values of the complexity coefficient $k(\theta)$ are quite irrelevant, as long as they are strictly increasing in the order of the nesting of the $\{L_i\}$. To be precise, we only need that there exists $\epsilon > 0$ such that for each $2 \leq i \leq s-1$,

$$\sup_{\theta \in L_i \setminus L_{i-1}} k(\theta) + \epsilon < \sup_{\theta \in L_{i+1} \setminus L_i} k(\theta). \tag{31}$$

Furthermore, while penalties of order $\frac{\log n}{n}$ work fine, so do– for instance– penalties of order $\frac{(\log\log n)^2}{n}$ or $\frac{1}{\sqrt{n}}$. This may be useful for potential practical applications, since it allows for tuning the estimator in a particular situation depending on the relative importance of

overfitting and underfitting. See, e.g., [2], for a discussion of such aspects in the context of statistical inference.

Remark 7. We have not addressed the problem of how to choose an appropriate sequence of models here. That is a question that would naturally follow a study of how the theoretical framework developed by [17] and this work can be applied to build constructive (non-random) codes.

3 Second-order properties of the lossy likelihood

3.1 Uniform Approximations of the Lossy Likelihood

The second-order lossy AEP emerges as a consequence of [31][Theorem 3] and [9][Theorem 16]. In order to prove the uniform version, we adopt a brute-force approach and prove stronger, uniform versions of these two results in Theorems 2 and 3. There is another possible proof approach using the method of types, and a sketch of this approach is outlined in Appendix B.

Proof of Theorem 2. We wish to apply the expansion of the distortion ball probabilities proved by Yang and Zhang, [31][Theorem 3], to the random string X_1^n. Since $d_3(\hat{P}_{X_1^n}, Q_\theta) < \infty$ and $D_{\min}(\hat{P}_{X_1^n}, Q_\theta) = 0$ for all $\theta \in \Theta_0$ (owing to the finite alphabet assumption and the fact that Q_θ is in the interior of the simplex respectively), the conditions of that theorem are satisfied and we have for any $\theta \in \Theta_0$ and any $c > 0$,

$$L_n(c,\theta) \leq \frac{Q_\theta^n(B(X_1^n, D))}{\exp(-nR(\hat{P}_{X_1^n}, Q_\theta, D) - \frac{1}{2}\log n)} \leq U_n(\theta), \tag{32}$$

for $D \in (0, d_1(\hat{P}_{X_1^n}, Q_\theta))$. Here,

$$\begin{aligned} L_n(c,\theta) &= e^{\lambda_\theta^{(n)} c}\left[\frac{cS_{n,\theta}^{1/2}}{\sqrt{2\pi}} e^{-\frac{c^2 S_{n,\theta}}{2n}} - 16 S_{n,\theta}^{3/2} d_3^{(n)}\right], \\ U_n(\theta) &= \frac{1}{1 - e^{\lambda_\theta^{(n)}}}\left[16 S_{n,\theta}^{3/2} d_3^{(n)} + \frac{S_{n,\theta}^{1/2}}{\sqrt{2\pi}}\right], \\ d_3^{(n)} &= d_3(\hat{P}_{X_1^n}, Q_\theta) \\ \text{and } S_{n,\theta} &= \frac{\partial^2}{\partial D^2} R(\hat{P}_{X_1^n}, Q_\theta, D) \end{aligned} \tag{33}$$

are all random variables depending on n and θ.

The first step is to note that a uniform strong law of large numbers guarantees that as $n \to \infty$, $d_1(\hat{P}_{X_1^n}, Q_\theta) \to d_1(P, Q_\theta)$ and $d_3(\hat{P}_{X_1^n}, Q_\theta) \to d_3(P, Q_\theta)$ uniformly over Θ_0. This tells us that for n large enough and any $\theta \in \Theta_0$,

$$d_1(\hat{P}_{X_1^n}, Q_\theta) \geq \inf_{\theta \in \Theta_0} d_1(P, Q_\theta) - \epsilon \tag{34}$$

for arbitrarily small $\epsilon > 0$, so that (32) holds for all $\theta \in \Theta_0$ and all $D \in d_1(P, \Theta_0)$ (which is the range of D specified in the statement of our theorem). Of the various criteria that can be used to obtain a uniform law of large numbers (such as the method of Vapnik and Chervonenkis[29], or the econometric approaches of [1] and [23]), the criterion of Pötscher and Prucha [23] seems most convenient to verify in this case. In particular, the compactness

of Θ_0 and the joint continuity of $E_{Q_\theta}[\rho(x,Y)]$ in x and θ imply that the criterion is satisfied. Note that the continuity of $E_{Q_\theta}[\rho(x,Y)]$ in θ is a consequence of the assumption that the parametrization by Θ is sensible: when $\theta_m \to \theta$, Q_{θ_m} converges weakly to Q_θ, and since ρ is automatically bounded and continuous on a finite alphabet, the expectations converge and continuity is verified.

It remains to show that, eventually w.p.1, $\sup_{\theta\in\Theta_0} U_n(\theta) < \infty$ and $\inf_{\theta\in\Theta_0} L_n(c,\theta) > 0$ for some $c > 0$. This would imply that

$$\log\left[\frac{Q_\theta^n(B(X_1^n,D))}{\exp(-nR(\hat{P}_{X_1^n},Q_\theta,D)-\frac{1}{2}\log n)}\right] = O(1) \text{ eventually w.p.1,} \tag{35}$$

which is the conclusion of the theorem. First note that $U_n(\theta)$ and $L_n(c,\theta)$ are continuous in θ; this follows from the smoothness of λ_θ and $R(P,Q_\theta,D)$ implied by [17] and from repeating the continuity argument in the previous paragraph for d_3. Now $U_n(\theta)$ is a continuous function over the compact set Θ_0 and hence achieves a maximum that is finite. For the lower bound, note that by similar arguments as before, [23] can be used to show that $L_n(c,\theta)$ converges uniformly to $L(c,\theta)$ over Θ_0, where

$$L(c,\theta) \quad = e^{\lambda_\theta c}\left[\frac{cS_\theta^{1/2}}{\sqrt{2\pi}} - 16S_\theta^{3/2}d_3\right], \tag{36}$$

where

$$S_\theta = \frac{\partial^2}{\partial D^2}R(P,Q_\theta,D) \tag{37}$$

and $d_3 = d_3(P,Q_\theta)$. Since $L(c,\theta)$ can be made arbitrarily large by choosing c large enough, and since it is a continuous function over a compact set, the minimum of $L_n(c,\theta)$ over Θ_0 is bounded away from 0 for large enough n and we are done.

□

Is the lossy likelihood expressible not merely in terms of the rate function at the empirical source distribution but also in terms of the rate function at the true source distribution? The answer to this question is provided by Theorem 3. The proof is lengthy and involved, and requires the use of ideas from the Vapnik-Chervonenkis theory for uniform limit laws. It is given in Appendix A.

3.2 Implications

Corollary 1. [Uniform 2nd-order lossy AEP] *Suppose $\{X_n\}$ is an i.i.d. process with marginal distribution P on a finite alphabet A, and $\{Q_\theta : \theta \in \Theta\}$ is a family of i.i.d. probability measures on the finite reproduction alphabet $\hat{A}$. For an arbitrary measurable distortion function ρ, let $R(P,Q_\theta,D)$ and $g_\theta(\cdot)$ be defined as in* (10) *and* (14)*. Suppose the optimal θ^* lies in the interior of the simplex of probability measures on $\hat{A}$. Then there exists a neighborhood of θ^* in Θ such that for any $D > 0$:*

$$-\log Q_\theta^n(B(X_1^n,D)) = nR(P,Q_\theta,D) + \sum_{i=1}^n g_\theta(X_i) + \tfrac{1}{2}\log n + \alpha_n(P,\theta,D) \tag{38}$$

holds for a.s.-ω in the probability space underlying X_1^∞, and $|\alpha_n(P,\theta,D)| \le C_{P,D}\log\log n$ for every $n > N(\omega)$, where $N(\omega)$ is finite and independent of θ.

Proof. Combine Theorems 2 and 3. □

A "central limit theorem" for the lossy likelihood L_n follows from the pointwise second-order generalized AEP that was proved in [9]. These pointwise results are extended to the locally uniform (in θ) case in Corollary 2.

Corollary 2. *Under the assumptions of Corollary 1, we have:*

1. *Locally Uniform "CLT": For $\epsilon > 0$ sufficiently small,*

$$\sup_{\theta \in B(\theta^*, \epsilon)} \frac{-\log Q_\theta^n(B(X_1^n, D)) - nR(P, Q_\theta, D)}{Var[g_{\theta'}(X)]\sqrt{n}} \Rightarrow N(0,1) \tag{39}$$

   for some θ' in the closure of the ball $B(\theta^*, \epsilon)$. In fact, θ' is the maximizer of $\sum_{i=1}^n g_\theta(X_i)$ viewed as a function of θ over the closed ball.

2. *Locally Uniform "LIL": For $\epsilon > 0$ sufficiently small,*

$$\limsup_{n\to\infty} \sup_{\theta \in B(\theta^*, \epsilon)} \frac{-\log Q_\theta^n(B(X_1^n, D)) - nR(P, Q_\theta, D)}{\sqrt{2\, Var[g_{\theta'}(X)] n \log\log n}} = 1 \quad w.p.1 \tag{40}$$

   for some θ' in the closure of the ball $B(\theta^*, \epsilon)$. In fact, θ' is the maximizer of $\sum_{i=1}^n g_\theta(X_i)$ viewed as a function of θ over the closed ball.

Proof. We only need to note that since $\sum_{i=1}^n g_\theta(X_i)$ is a continuous function of θ over the closed ball, at least one maximizer θ' exists, at which the supremum of the function over the open ball is achieved. □

4 Three Examples: Lossy MDL vs. Lossy Maximum Likelihood

The examples in this section are contrived and somewhat artificial, since they all involve choosing between only 2 models, with the simpler model being a singleton. However, not only do they provide a sanity check and some concrete simulation-based illustrations, but they also contain the basic proof ingredients that are utilized in the full-fledged result for finite alphabets in the next section.

4.1 Gaussian codes

Let us denote the normal distribution with mean μ and variance σ^2 by $N(\mu, \sigma^2)$. Suppose the source distribution P is $N(\mu, \sigma^2)$, whereas the coding distribution is $N(\nu, \tau^2)$. If X is drawn from P and Y from Q, and for the squared-error distortion function $\rho(x, y) = (x-y)^2$,

$$E_Q[e^{\lambda\rho(x,Y)}] = \sqrt{\frac{\pi}{-\lambda}}\phi_{\nu,\tau^2} * \phi_{0,-(2\lambda)^{-1}}(x) = \sqrt{\frac{\pi}{-\lambda}}\phi_{\nu,\tau^2-(2\lambda)^{-1}}(x),$$

where ϕ is the density function of the normal with the subscripted mean and variance. Thus,

$$\Lambda(P, Q, \lambda) = -\tfrac{1}{2}\log(1 - 2\lambda\tau^2) - \frac{\sigma^2 + (\mu-\nu)^2}{2\tau^2 - \frac{1}{\lambda}}.$$

Setting $\Lambda'(P, Q, \lambda) = D$ yields a quadratic equation for λ, solving which gives

$$\lambda^* = -\frac{(v - D)}{2D\tau^2}, \quad \text{where} \quad v = \tfrac{1}{2}[\tau^2 + \sqrt{\tau^4 + 4D\{\sigma^2 + (\mu-\nu)^2\}}\,]. \tag{41}$$

Recalling that $R(P,Q,D) = \Lambda^*(P,Q,D) = \lambda^* D - \Lambda(P,Q,\lambda^*)$, we have

$$R(P,Q,D) = \tfrac{1}{2}\log(\frac{v}{D}) - \frac{(v-D)(v-V)}{2\tau^2 v} \quad \text{where} \quad V = \sigma^2 + (\mu-\nu)^2. \tag{42}$$

Such an explicit calculation of the rate function, though very difficult to obtain for more general source and coding distributions, will turn out to be extremely useful in specific cases as we shall see. In particular, since the parameter space (for fixed P) is just two-dimensional here (for the mean and variance of Q), we can use simple calculus to minimize the rate function with respect to Q.

This yields the optimal rate

$$R(P,D) = \tfrac{1}{2}\log\left(\frac{V}{D}\right) \tag{43}$$

with the optimal distribution Q^* of the encoded data being parametrized by

$$\nu^*(\mu,\sigma,D) = \mu \quad \text{and} \quad \tau^*(\mu,\sigma,D) = \sqrt{V-D} \tag{44}$$

Example 1. Dichotomy for class of coding distributions with varying means.
Suppose that the source $\{X_n\}$ is a real-valued, stationary and ergodic, with zero mean and finite variance. Consider a parametric family of i.i.d. coding distributions $\{Q_\theta : \theta \in \Theta = \mathbb{R}\}$ where Q_θ is $N(\theta,1)$, and let the single-letter distortion function ρ_n be mean-squared error (that is, $\rho_n(x_1^n, y_1^n) = \frac{1}{n}\sum_{i=1}^n (x_i - y_i)^2$). According to (44), $\theta^* = 0$. We show below that $\hat{\theta}_n^{\mathrm{LML}}$ is simply the empirical average of the data points X_1^n.

The LML estimator is the maximizer of $Q_\theta^n(B(X_1^n, D))$, which depends only on the Euclidean distance between the point $X_1^n \in \mathbb{R}^n$ and the mean of the distribution Q_θ^n, due to the spherical symmetry of the multivariate normal and the fact that the distortion function is simply Euclidean distance. The mean of Q_θ^n is the point $(\theta, \theta, \ldots, \theta)$ on the main diagonal in $\mathbb{R}^n$, and its distance to $(X_1, X_2, \ldots, X_n)$ is minimized (hence the probability of the ball around X_1^n maximized) by simple calculus:

$$\frac{\partial}{\partial\theta}\left[\sum_{i=1}^n (X_i - \theta)^2\right] = 0 \iff \sum_{i=1}^n (2\theta - 2X_i) = 0 \iff \sum_{i=1}^n X_i = n\theta$$

so that the LML estimator as a function of the data is the same as the (classical, lossless) maximum likelihood estimator! That is,

$$\hat{\theta}_n^{\mathrm{LML}} = \frac{1}{n}\sum_{i=1}^n X_i. \tag{45}$$

Thus the consistency and asymptotic normality of this estimator trivially follow from the corresponding results for the lossless maximum likelihood estimator. (Note that $\theta^* = 0$ for all $D \geq 0$ by (44).) Furthermore, as in the lossless case, this estimator will forever fluctuate around θ^*, by an application of the Law of the Iterated Logarithm. In other words, $\hat{\theta}_n^{\mathrm{LML}} \neq \theta^* = 0$ infinitely often, w.p.1.

Now consider a penalty function such that the $k(\theta)$ is equal to 1 for all $\theta \neq 0$ and zero otherwise, and $c(n) = \frac{\log n}{n}$. This means that we are expressing a preference for the simpler 0-dimensional set $\{0\}$ within the real line, and we would like it to be selected when θ^* is

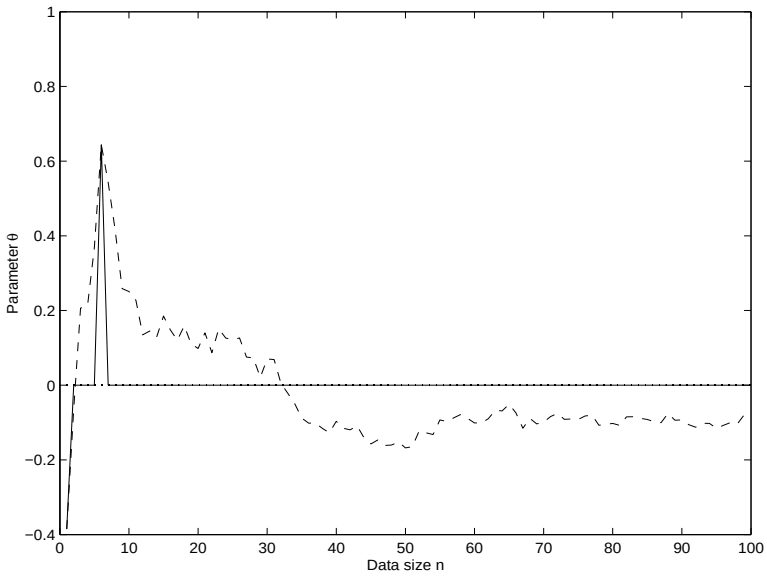

Figure 3: The dashed line denotes the pseudo-LML estimator and the solid line is the pseudo-LMDL estimator. Here $\theta^* = 0$.

in deed 0. With the fixing of a penalty function, the pseudo-LMDL and LMDL estimators are well-defined. The fact that $\hat{\theta}_n^{\mathrm{LMDL}} = 0$ eventually w.p.1 is implied by the classical model selection theory in statistics, since this is just the classical MDL estimator.

Figure 3 shows an explicit numerical example illustrating the behavior of the two pseudo-estimators, which suggests that the pseudo-LMDL estimator not only converges to the correct value, but also "finds" θ^* in finite time and then stays there.

Example 2. Dichotomy for class of coding distributions with varying variances, through analysis of pseudo-estimators.
Similar conclusions hold for the case when we take the coding distributions Q_θ to be i.i.d. $N(0, \theta)$ with $\theta \in [0, \infty)$.

To see this analytically, we use a very different approach from that used for Example 1. This is because when the variance is the parameter, it is very difficult to analytically (or even through simulation) determine the LML and LMDL estimators. The root of the problem is the difficulty of computing many instances of Q-integrals (distortion ball probabilities), and then following this difficult computation, to maximize the result over the parameter space. In Example 1, the geometric symmetry of the problem caused the result to be *independent of the distortion level* D, and enabled the direct computation of the LML estimator. This is obviously an exceptional circumstance.

Thus, the approach we use here is to focus on the pseudo-estimators. Since the pseudo-LML estimator is just the minimizer of $R(\hat{P}_{X_1^n}, Q_\theta, D)$, (44) implies

$$\tilde{\theta}_n^{\mathrm{LML}} = \tau^*(\mu_n, \sigma_n, D) = \sqrt{V_n - D} \tag{46}$$

where $\mu_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2$ are the mean and the variance of the empirical distribution, and $V_n = \sigma_n^2 + \mu_n^2$. Further, using (43) and (42),

$$\begin{aligned} R(\hat{P}_{X_1^n}, D) &= R(\hat{P}_{X_1^n}, \tilde{\theta}_n^{\mathrm{LML}}, D) = \tfrac{1}{2}\log\frac{V_n}{D} \\ R(\hat{P}_{X_1^n}, \theta^*, D) &= \tfrac{1}{2}\log\frac{v}{D} - \frac{(v-D)(v-V_n)}{2\tau^2 v} \end{aligned} \tag{47}$$

Now suppose we choose a penalty function that simply adds a penalty of $\log n$ for every $\theta \neq \theta^*$, so that the lower-dimensional subset L containing θ^* that we are considering here is just the singleton. Then,

$$\tilde{\theta}_n^{\mathrm{LMDL}} = \arg\min_{\theta\in\Theta}\left[R(\hat{P}_{X_1^n}, Q_\theta, D) + 1_{\theta^*}^c \frac{\log n}{n}\right] = \arg\min\{R_1, R_2\} \tag{48}$$

$$\text{where} \quad R_1 = \arg\min_{\theta\in\Theta-\{\theta^*\}}\left[R(\hat{P}_{X_1^n}, Q_\theta, D) + \frac{\log n}{2n}\right] = R(\hat{P}_{X_1^n}, D) + \frac{\log n}{n} \tag{49}$$

$$\text{and} \quad R_2 = R(\hat{P}_{X_1^n}, \theta^*, D) \tag{50}$$

Noting that the argument for R_1 is exactly $\tilde{\theta}_n^{\mathrm{LML}}$ and the minimizing argument in (48) is exactly $\tilde{\theta}_n^{\mathrm{LMDL}}$, we see that *the behavioral differences between the pseudo-LML and pseudo-LMDL estimators must be completely captured by the relationship between R_1 and R_2* (or equivalently, by the relationship between the two rates in (47) above). This is the key insight which allows us to unravel the dichotomy in this example.

From (41) and (44), observe that

$$\begin{aligned} & v = Va + Db \\ & \text{where} \quad a = \tfrac{1}{2}\left[1 + \left\{1 + \frac{4D}{(V+D)^2}(V_n - V)\right\}^{\frac{1}{2}}\right] \\ & \text{and} \quad\ b = \tfrac{1}{2}\left[\left\{1 + \frac{4D}{(V+D)^2}(V_n - V)\right\}^{\frac{1}{2}} - 1\right] \end{aligned} \tag{51}$$

Expanding a and b as series in terms of $V_n - V$, we have

$$b = a - 1 = \frac{D}{(V+D)^2}(V_n - V) - \frac{D^2}{(V+D)^4}(V_n - V)^2 + O(V_n - V)^3 \tag{52}$$

which yields

$$\begin{aligned} v - V_n &= V - V_n + b(V+D) \\ &= (V - V_n)\left[\frac{V}{V+D}\right] - (V - V_n)^2\left[\frac{D^2}{(V+D)^3}\right] + O(V - V_n)^3 \end{aligned} \tag{53}$$

the first line following from (51) and the fact that $a = 1 + b$, while the second following from (52).

Using the fact that $\log(1+x) = x + O(x^2)$ for small x,

$$\begin{aligned}\log\left(\frac{v}{V_n}\right) &= \log\left[1+\frac{v-V_n}{V_n}\right] \\ &= (V-V_n)\left[\frac{V}{V_n(V+D)}\right] + O(V-V_n)^2\end{aligned} \tag{54}$$

so that

$$\begin{aligned}R(\hat{P}_{X_1^n},\theta^*,D) - R(\hat{P}_{X_1^n},\tilde{\theta}_n^{\mathrm{LML}},D) &= \tfrac{1}{2}\log\frac{v}{V_n} - \frac{(v-D)(v-V_n)}{2\tau^2 v} \\ &= (V-V_n)\chi + O(V-V_n)^2 \\ \text{where}\quad \chi &= \frac{V}{2(V+D)V_n(V-D)(Va+Db)}\,\bar{\chi} \\ \text{and}\quad \bar{\chi} &= [(V-D)(Va+Db) - V_n(Va+Db-D)]\end{aligned} \tag{55}$$

However,

$$\begin{aligned}\bar{\chi} &= (V-V_n)[Va+Db] + D(V_n - Db - Va) \\ &= (V-V_n)\left[Va+Db-D\left(\frac{V}{V+D}\right)\right] + O(V-V_n)^2 \\ &= O(V-V_n)\end{aligned} \tag{56}$$

where the second line of the display followed from (51) and (53), and plugging this back into (55) yields

$$R_2 - R(\hat{P}_{X_1^n},D) = O(V_n - V)^2 \tag{57}$$

In other words, the first order term in the expansion of the difference (55) vanishes! Since $V_n - V$ is a sum of zero-mean random variables, the Law of the Iterated Logarithm tells us that the difference of rates above is of order $O(\frac{\log\log n}{n})$ (not $O(\sqrt{\frac{\log\log n}{n}})$)! In particular, the difference is $o(\frac{\log n}{n})$. Consequently R_2 is eventually strictly less than R_1, implying that $\tilde{\theta}_n^{\mathrm{LMDL}} = \theta^*$ eventually w.p.1. Further, by using the complementary part of the Law of the Iterated Logarithm for $(V_n - V)$, which says that the sum makes excursions outside an interval of $O(\frac{\log\log n}{n})$ infinitely often with probability 1, we have that $\tilde{\theta}_n^{\mathrm{LML}}$ fluctuates forever around θ^* even as it approaches it.

Figure 4 contains a simulation comparing the behavior of the lossy pseudo-estimators.

4.2 Bernoulli case

Consider an i.i.d. Bernoulli source with parameter $p = \Pr(X=1)$. The class of reproduction distributions we consider is the class of i.i.d. Bernoulli distributions (with parameter $\theta \in [0,1]$).

Just as we did in the case of Gaussian codes, we can explicitly compute the rate function in this case using its characterization as the Legendre-Fenchel transform of the mean of the log moment generating function. This yields

$$\mu_\theta \equiv \exp\lambda_\theta = \frac{-\{(1-2D)\theta^2 - 2(p-D)\theta + (p-D)\} \pm \sqrt{\Delta}}{2\theta(1-D)(1-\theta)} \tag{58}$$

$$\text{where}\quad \Delta = [(p-D)(1-\theta)^2 + (1-p-D)\theta^2]^2 + 4D(1-D)\theta^2(1-\theta)^2 \tag{59}$$

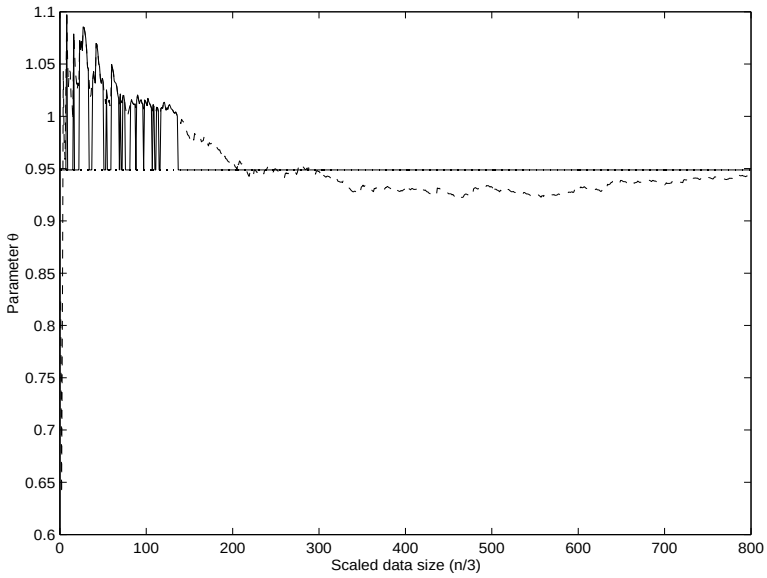

Figure 4: The dashed line denotes the pseudo-LML estimator and the solid line is the pseudo-LMDL estimator. In this example, source variance=1, $D = 0.1$ and $\theta^* = 0.949$.

and

$$R(p, \theta, D) = D \log \mu_\theta - p \log[\theta + (1-\theta)\mu_\theta] - (1-p) \log[\theta \mu_\theta + 1 - \theta] \tag{60}$$

It is messy but straightforward to verify that $\theta^* = \frac{p-D}{1-2D}$ for $D < \min p, 1-p$, and that $R(p, D) = H_B(p) - H_B(D)$, as we expect from the direct computation of the rate-distortion function (see, e.g., [7]). This rate distortion function is plotted in Figure 5 (separately for fixed p and fixed D), while Figure 6 plots θ^* versus p. Note that for fixed $D > 0$, there is a symmetric middle region where $R(D)$ is positive (and lies below the Bernoulli entropy function which represents the lossless rate), and this region shrinks as D increases.

Figure 6 reveals an interesting insight: as intuition would suggest, lossy compression typically involves producing encoded strings whose distribution is less random (has less entropy or smaller lossless compression rate). In the Bernoulli context, this means that the optimal reproduction parameter is closer to the nearer periphery (0 or 1) than the source parameter. However, the gap between the source parameter and the optimal reproduction parameter decreases as we approach $p = 0.5$ from either side, in such a way that $\theta^* = 0.5$ for $p = 0.5$. Yet this does not mean that data from a Bernoulli($\frac{1}{2}$) source cannot be compressed; indeed, Figure 5 indicates that for a distortion level of 0.1, such a source has optimal compression rate of 0.36 bits/symbol, which is significantly less than the optimal lossless compression rate of 1 bit/symbol. This fact (of the gap vanishing) is what is responsible for the fact that θ^* remains unique even for $p = 0.5$, since a gap would have implied two minimizers by symmetry.

Let us now investigate the behavior of the various lossy estimators for θ^*. As in the second part of Example 1, we penalize outside the singleton set containing θ^*. Thus (48), (49) and (50), which determine $\tilde{\theta}_n^{\mathrm{LMDL}}$ using just two real numbers, hold.

Using the formula (60) for rate obtained above, we have

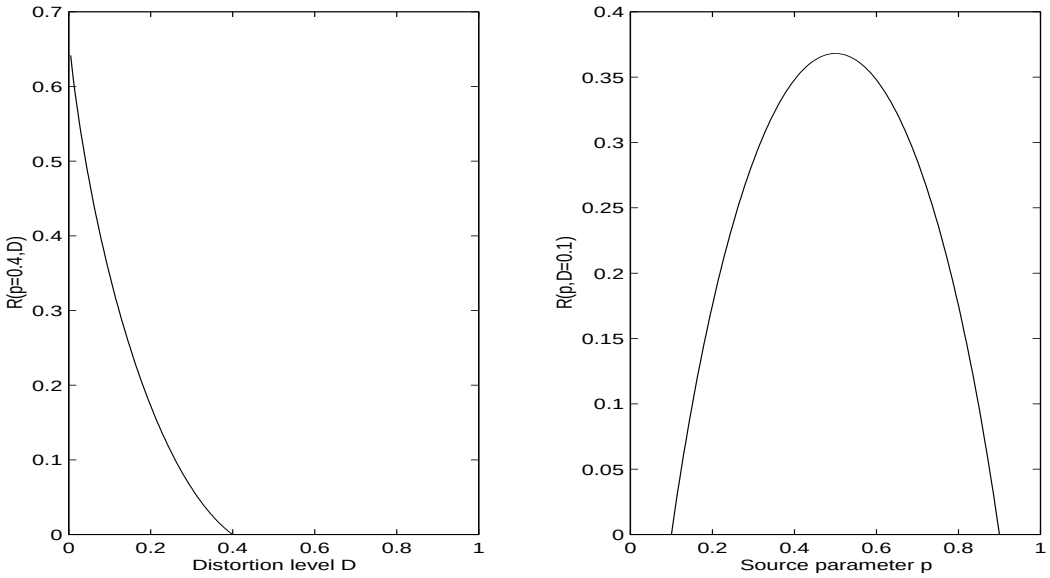

Figure 5: On the left is plotted the Bernoulli rate-distortion function for fixed source distribution $p = 0.4$ as the distortion level D varies. On the right is plotted the Bernoulli rate-distortion function for fixed distortion level $D = 0.1$ as the source distribution p varies.

$$\begin{aligned} R(\hat{P}_{X_1^n}, \theta^*, D) &= D \log\left(\frac{D}{1-D}\right) - \hat{p}_n \log\left(\frac{p}{1-D}\right) - (1-\hat{p}_n)\log\left(\frac{1-p}{1-D}\right) \\ &= H_B(\hat{p}_n) - H_B(D) + D(\hat{p}_n \| p) \end{aligned} \tag{61}$$

which implies

$$R(\hat{P}_{X_1^n}, \theta^*, D) - R(\hat{P}_{X_1^n}, D) = D(\hat{p}_n \| p) = O\left(\frac{\log\log n}{n}\right) \text{ eventually w.p.1} \tag{62}$$

As in the second part of Example 1, the law of the iterated logarithm then yields the dichotomy for the pseudo-estimators.

Figure 7 illustrates the behavior of the pseudo-LML and pseudo-LMDL estimators, when the "preferred" set L is simply the singleton $\{\theta^*\}$ containing the $R(D)$-achieving output distribution $\theta^* = (p - D)/(1 - 2D)$. It is clear from repeated simulations that the pseudo-LMDL estimator "hits and stays at" θ^* quite fast (unlike the pseudo-LML estimator which bounces around forever).

5 The LML/LMDL Dichotomy for i.i.d. finite-alphabet codebooks

5.1 The admissible class of sources

Suppose that the source data $\{X_n\}$ taking values in a finite alphabet A of size m is generated i.i.d. from a probability distribution on A. Let $\Sigma = \Theta$ parametrize the simplex of all

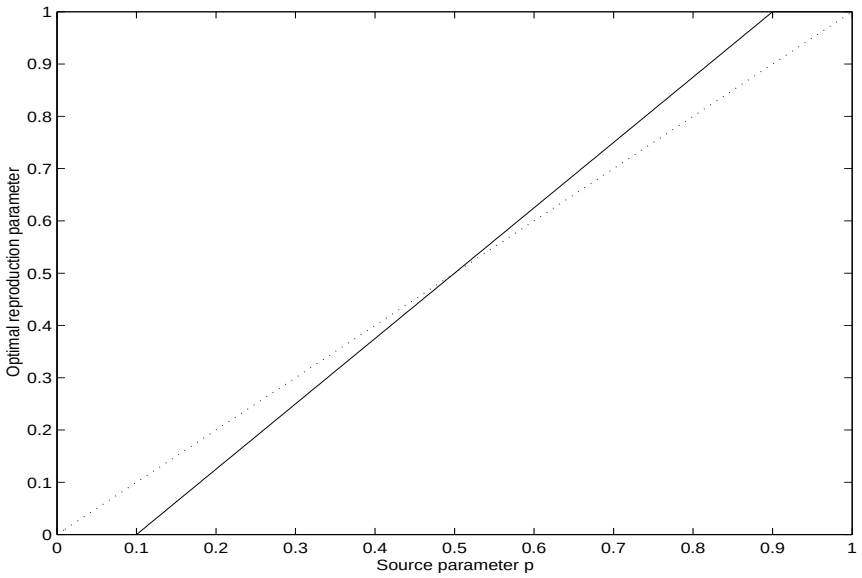

Figure 6: The solid line denotes the optimal θ^* for fixed distortion $D = 0.1$, while the dashed line denotes the optimal θ^* for fixed distortion $D = 0$ (which is just the parameter for the source distribution).

i.i.d. probability measures on $A = \hat{A}$ via the canonical parametrization that uses the first $m-1$ coordinates. Suppose the ρ_n are single-letter distortion functions, so that $\rho_n(x_1^n, y_1^n) = \frac{1}{n}\sum_{i=1}^n \rho(x_i, y_i)$. For clarity, we use Σ to denote the class of source distributions, and Θ to denote the class of reproduction distributions, though they are the same class. Without loss of generality, denote the m symbols in A by $1, 2, \ldots, m$. A nested sequence of models $L_1 \subset L_2 \subset \ldots \subset L_s \subset \Theta$ and the associated complexity coefficient $k(\cdot)$ and penalty $k(\theta)c(n)$ are set up as described in Section 2.3.

By the discussion in Example 4.3.1 in [10], we know that the LML estimator is consistent in this setup, because the simplex of i.i.d. distributions is a Polish space. Further, the discussions of Examples 4.3.5 and 4.3.6 in [10] show that the pseudo-LML estimator, the pseudo-LMDL estimator and the LMDL estimator are all consistent estimators. We wish to investigate the behaviour of these four estimators more closely and compare the LML estimator (pseudo-LML estimator) against the LMDL estimator (pseudo-LMDL estimator).

The key to comparing the various estimators is to investigate the function that takes a source distribution to its optimal reproduction distribution (assuming the latter is unique). Let η_D be a function on the class of source distributions that takes each source distribution to the optimal reproduction distribution. For convenience, first set $R(P_\sigma, Q_\theta, D) = f(\sigma, \theta)$, so that $\eta(\sigma) = \arg\min_\theta f(\sigma, \theta)$ parametrizes $\eta_D(P_\sigma)$. Note that we are using η_D to refer to a map between spaces of probability distributions, while η refers to the corresponding map between the parameter spaces. Since we are dealing with finite alphabets, Θ is compact, and the continuity of f implies that a minimizer of f exists. Thus η is non-empty, though it can be many-valued for some values of σ. The restriction of sources to the class $\mathbb{S}(D)$ in this section is precisely to eliminate undesirable possibilities like a many-valued η. To set down what restrictions on the source distribution are needed, we introduce the following Proposition.

Proposition 1. *Let $\Sigma_0 \subset \Sigma$ be the set that parametrizes the class $\mathbb{S}(D)$ of sources. The func-*

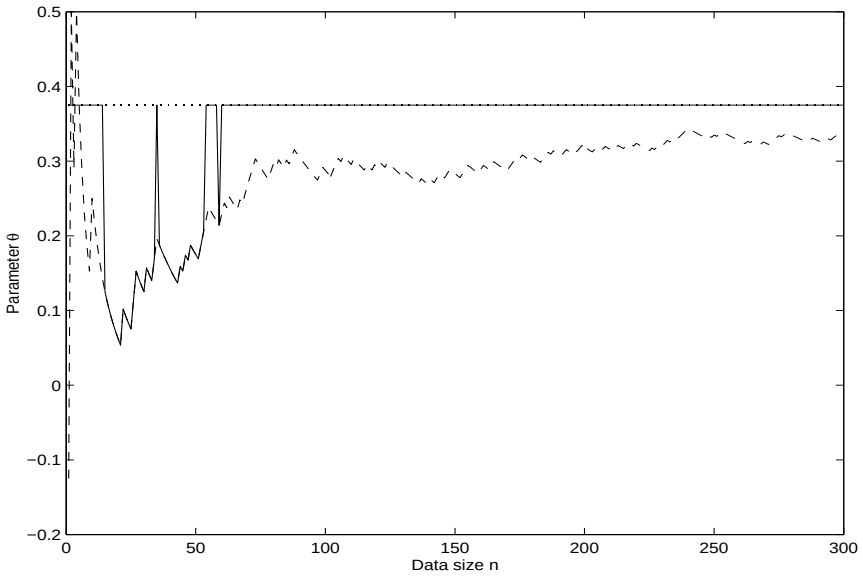

Figure 7: The dashed line represents the pseudo-LML estimator, and the solid line is the pseudo-LMDL estimator. Here $p = .4$, $D = .1$ and $\theta^* = .375$.

tion $\eta : \Sigma_0 \to \Theta$ *given by* $\eta(\sigma) = \arg\min_\theta f(\sigma, \theta)$ *is well-defined and is* C^∞ *in a neighborhood of* σ^*.

In order to prove Theorem 1, we need the following lemmas, which are an extension of [17][Lemma 4].

Lemma 1. *Let* $D > 0$ *and* $\sigma^* \in \Sigma_0$. *Then, the function* $\ell(\sigma, \theta) \equiv \lambda^*(P_\sigma, Q_\theta, D)$ *is smooth (or* C^∞*) in both of its arguments, in a neighborhood of* (σ^*, θ^*).

Proof. First, note that for finite alphabets and any $\lambda < 0$, $\Lambda(P_\sigma, Q_\theta, \lambda)$ can be written out in terms of the components of θ and σ using finite sums, and is clearly differentiable in λ and θ an arbitrary number of times. Being linear in σ, it is also smooth in σ, though second and higher derivatives are 0.

Define the function $\Psi_1 : \Sigma_0 \times \text{int}(\Theta) \times (-\infty, 0) \to \mathbb{R}$ by

$$\Psi_1(\sigma, \theta, \lambda) = \Lambda'(P_\sigma, Q_\theta, \lambda) - D \tag{63}$$

By the definition of $\ell(\sigma, \theta)$ and setting $\ell^* = \ell(\sigma^*, \theta^*)$, we have

$$\Psi_1(\sigma, \theta, \ell(\sigma, \theta)) = \Psi_1(\sigma^*, \theta^*, \ell^*) = 0 \tag{64}$$

The smoothness of Λ noted above implies that Ψ_1 is smooth in each of its arguments.

Also, Lemma 1 in [16] implies that $\Lambda''(P, Q, \lambda) > 0$ provided $D_{\min}(P, Q) < d_1(P, Q)$. We need to check whether this is true in our case. Since $\sigma^* \in \Sigma_0$ and Σ_0 is open, there exists a neighborhood of σ^* in Σ_0 on which P_σ has full support. Q_θ has full support since it lies in the interior of the simplex. This means that for σ in a neighborhood of σ^*, $D_{\min}(P_\sigma, Q_\theta) = 0$, while $d_1(P_\sigma, Q_\theta) > 0$ since $\rho \equiv 0$ is ruled out. Thus

$$\Psi_1'(\sigma, \theta, \lambda) \quad = \Lambda''(P_\sigma, Q_\theta, \lambda) > 0 \tag{65}$$

so that the Implicit Function Theorem can be invoked not only to show that $\ell(\sigma, \theta)$ is well-defined but that it is smooth in a neighborhood of (σ^*, θ^*) (by the smoothness of Ψ_1). □

Lemma 2. *Let $D > 0$ and $\sigma^* \in \Sigma_0$. Then, the function $f(\sigma, \theta) \equiv R(P_\sigma, Q_\theta, D)$ is smooth (or C^∞) in both of its arguments, in a neighborhood of (σ^*, θ^*).*

Proof. Define the function $\Psi_2 : \Sigma_0 \times \text{int}(\Theta) \times (0, \infty) \to \mathbb{R}$ by

$$\Psi_2(\sigma, \theta, R) = R - \ell(\sigma, \theta)D + \Lambda(P_\sigma, Q_\theta, \ell(\sigma, \theta)) \tag{66}$$

By the definition of the rate function $R(P, Q, D)$, we have

$$\Psi_2(\sigma, \theta, f(\sigma, \theta)) = \Psi_2(\sigma^*, \theta^*, R(P, D)) = 0 \tag{67}$$

Note that the local smoothness of $\ell(\sigma, \theta)$ implies that $\Lambda_x(Q_\theta, \ell(\sigma, \theta))$ and consequently $\Lambda(P_\sigma, Q_\theta, \ell(\sigma, \theta))$ are locally smooth in σ and θ. Thus Ψ_2 is smooth in a neighborhood of $(\sigma^*, \theta^*, R(P, D))$ and furthermore,

$$\frac{\partial \Psi_2(\sigma, \theta, R)}{\partial R} = 1 \neq 0 \tag{68}$$

Hence by the Implicit Function Theorem, $f(\sigma, \theta)$ is not just well-defined but also smooth in a neighborhood of (σ^*, θ^*) (by the smoothness of Ψ_2). □

Proof of Proposition 1. The Proposition follows from repeated applications of the Implicit Function Theorem starting with the basic fact of smoothness of Λ, and based on the observation that $\theta = \eta(\sigma)$ solves $\nabla_\theta f(\sigma, \theta) = 0$.

Let $W \subset \Sigma \times \Theta$ be a neighborhood of (σ^*, θ^*) on which $f(\sigma, \theta)$ is smooth. Consider the function $F : W \to \mathbb{R}^{|\hat{A}|-1}$ defined by

$$F(\sigma, \theta) = \nabla_\theta f(\sigma, \theta) \tag{69}$$

Clearly F is smooth on W, and by definition of η and the fact that $\sigma^* \in \Sigma_0$ ensures that W is an open subset of $\Sigma \times \Theta$ (hence does not contain any boundary points), we have

$$F(\sigma, \eta(\sigma)) = F(\sigma^*, \theta^*) = 0 \tag{70}$$

Furthermore, since $\eta(\sigma)$ minimizes $f(\sigma, \cdot)$,

$$\nabla_\theta F(\sigma^*, \theta^*) = \text{Hess}_\theta f(\sigma^*, \theta^*) > 0 \tag{71}$$

so that the Implicit Function Theorem not only assures us that $\eta(\sigma)$ is well-defined but also that is smooth in a neighborhood of σ^* (by the smoothness of F). □

In the rest of this section, we use this theorem to describe the behavior of various lossy estimators.

5.2 Behavior of the pseudo-LML estimator

5.2.1 Parameters

Let P and $\hat{P}_{X_1^n}$ be parametrized by σ^* and $\hat{\sigma_n}$ respectively. Denoting by p_k the probability of the symbol k under P, we have $p_k = \sigma_k^*$ for $k = 1, \ldots, m-1$, and $p_m = 1 - \sum_{i=1}^{m-1} p_i$.

Lemma 3. *For any probability distribution P on A, define the matrix*

$$\Xi \equiv \Xi^{(m-1)} = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 & \dots & -p_1p_{m-1} \\ -p_1p_2 & p_2(1-p_2) & -p_2p_3 & \dots & -p_2p_{m-1} \\ \vdots & \vdots & \ddots & & \vdots \\ -p_1p_{m-1} & -p_2p_{m-1} & \dots & & p_{m-1}(1-p_{m-1}) \end{pmatrix}. \tag{72}$$

If P is in the interior of the simplex $\mathcal{P}(A)$, then Ξ is non-singular.

Proof. Suppose $p_j \neq 0$ for each $j = 1, \dots, m-1$, and also that $1 - \sum_j p_j \neq 0$. Let us assume Ξ is singular, i.e., $\Xi v = 0$ for some $v \neq 0$, and then obtain a contradiction. For every i,

$$\begin{aligned} \sum_j \Xi_{ij} v_j = 0 &\iff \sum_{j \neq i} -p_i p_j v_j + (p_i - p_i^2) v_i = 0 \\ &\iff p_i v_i = p_i \sum_j p_j v_j \\ &\iff v_i = \sum_j p_j v_j \end{aligned} \tag{73}$$

where we used the fact that $p_i \neq 0$ to obtain the last statement. Since the right-hand side is a constant independent of the index, the eigenvector v must be a multiple of the constant vector $(1, \dots, 1)$. Using again the last display, this implies $\sum_j p_j = 1$, which contradicts our assumption. □

Remark 8. We conjecture that for the $(m-1)$-dimensional matrix Ξ,

$$\det(\Xi) = (1 - \sum_{j=1}^{m-1} p_j) \prod_{j=1}^{m-1} p_j = \prod_{j=1}^{m} p_j \tag{74}$$

This is hard to prove by induction, and we have not been able to prove it by softer arguments.

For n large enough so that $\hat{\sigma_n}$ is in a sufficiently small neighborhood of σ^*,

$$\tilde{\theta}_n^{\mathrm{LML}} \equiv \eta(\hat{\sigma_n}) = \eta(\sigma^*) + (\hat{\sigma_n} - \sigma^*) \cdot \Gamma + O(\|\hat{\sigma_n} - \sigma^*\|^2) \tag{75}$$

where we have denoted the matrix $D\eta(\sigma^*)$ by Γ for convenience.

Now

$$\hat{\sigma_n} - \sigma^* = \frac{1}{n} \sum_{i=1}^{n} \zeta(X_i) \tag{76}$$

where the components of the function $\zeta : A \to [-1, 1]^{m-1}$ are defined by

$$\zeta_k(x) = 1_{\{x=k\}} - p_k \qquad (k = 1, \dots, m-1) \tag{77}$$

It is easy to check the following: $E[\zeta_k(X)] = 0$ and $E[\zeta_k(X)]^2 = p_k(1-p_k)$ for each $k = 1, \dots, m-1$, and $E[\zeta_k(X)\zeta_j(X)] = -p_k p_j$ for $k \neq j$. Thus, for each $i = 1, \dots, n$, the covariance matrix of $\zeta(X_i)$ is the $(m-1) \times (m-1)$ matrix Ξ specified by (72).

To analyze the detailed almost-sure behavior of the first-order term $\hat{\sigma_n} - \sigma^*$, we use Berning's multivariate version of the Law of the Iterated Logarithm. This is restated below, with slight notational changes for convenience.

Fact 6. *[6]If $\{Z_n\}$ are independent random vectors in $\mathbb{R}^p$ with $EZ_n = 0$ and $Cov[Z_n] = \Xi_n$, if for some positive constants $\{s_n^2\}$, $s_n^2 \uparrow \infty$, $s_{n+1}^2/s_n^2 \to 1$, $\operatorname{ess\,sup}|Z_n| \leq \epsilon_n s_n (\log\log s_n^2)^{-\frac{1}{2}}$ for some sequence $\epsilon_n \to 0$, and $\frac{1}{s_n^2}\Sigma_{j=1}^n \Xi_j \to \Xi$, then the a.s. limit set D of $\{(2s_n^2 \log\log s_n^2)^{-\frac{1}{2}}\Sigma_{j=1}^n Z_j\}$ is K_Ξ, where K_Ξ is the unit ball of the space $H_\Xi = \{x\Xi : x \in \mathbb{R}^n\}$ with respect to the norm $\|\cdot\|_\Xi$ defined by $\|x\Xi\|_\Xi = (x\Xi x^t)^{\frac{1}{2}}$.*

Recall that the source distribution is in the interior of the simplex, and so Ξ is non-singular. Applying Fact 6 with $s_n^2 = n$, noting that the boundedness and other conditions are trivially satisfied, we arrive at the conclusion that the a.s. limit set of $\left\{(2n\log\log n)^{-\frac{1}{2}}\Sigma_{j=1}^n Z_j\right\}$, where $Z_j = \zeta(X_j)$, is the unit ball K_Ξ. Consequently, the a.s. limit set of

$$\{\sqrt{\frac{n}{2\log\log n}}(\hat{\sigma_n} - \sigma^*)\cdot\Gamma\}$$

is the ellipsoid $\mathcal{E} = \{u \in \mathbb{R}^{m-1} : u = v\cdot\Gamma, v \in K_\Xi\}$.

The ellipsoid $\mathcal{E}$ has dimension less that $m-1$ if and only if Γ is singular. In either case, however, the boundary of $\mathcal{E}$ intersects the j-th coordinate axis in $\mathbb{R}^{m-1}$ in exactly two points $\pm E_j$ (where E_j may equal 0 for the deficient dimensions). Equation (75) now implies

$$|[\eta(\hat{\sigma_n}) - \eta(\sigma^*)]_j| \leq \sqrt{\frac{2\log\log n}{n}}E_j + O\left(\frac{\log\log n}{n}\right) \quad \text{eventually w.p.1} \tag{78}$$

for each coordinate j.

If Γ is non-singular, we see below that the ellipsoid $\mathcal{E}$ must have full dimension. By the definition of K_Ξ, $\|v\|_\Xi^2 = v\Xi^{-1}v^t \leq 1$ since $\|v\|_\Xi = \|v\Xi^{-1}\cdot\Xi\|_\Xi = (v\Xi^{-1}\cdot\Xi\cdot(\Xi^{-1})^t v^t)^{\frac{1}{2}} = (v\Xi^{-1}v^t)^{\frac{1}{2}}$ using the fact that Ξ is non-singular and symmetric. Rewriting in terms of u with $v = u\cdot\Gamma^{-1}$ yields $u\Gamma^{-1}\Xi^{-1}(\Gamma^{-1})^t u^t \leq 1$ as the defining condition of the required limit set. Since $P = P_{\sigma^*}$ is in the interior of the simplex, the limit set of the sequence of vectors $\{\sqrt{\frac{n}{2\log\log n}}(\hat{\sigma_n} - \sigma^*)\cdot\Gamma\}$ is, w.p.1, the solid $(m-1)$-dimensional ellipsoid $\mathcal{E}$ in $\mathbb{R}^{m-1}$ given by $\{u : u\Phi u^t \leq 1\}$, where $\Phi = \Gamma^{-1}\Xi^{-1}(\Gamma^{-1})^t$. Since $\mathcal{E}$ is of full (that is, $m-1$) dimension, its boundary intersects the j-th coordinate axis in $\mathbb{R}^{m-1}$ in exactly two points $\pm E_j$ (where $E_j \neq 0$). Equation (75) now implies

$$|[\eta(\hat{\sigma_n}) - \eta(\sigma^*)]_j| \geq \sqrt{\frac{2\log\log n}{n}}(E_j - \epsilon) \quad \text{i.o. w.p.1} \tag{79}$$

for each coordinate j.

5.2.2 Fluctuations

Let $\{L_i\}$ be a nested sequence of subsets of Θ. Suppose for some fixed q, the dimension of L_q is strictly less than $(m-1)$ and L_q contains θ^*. This means that any ball around θ^* in the simplex will contain directions not in L (more precisely, if V_Θ is the tangent space of Θ at θ^* and V_L is the tangent space of L at θ^*, then $V_L^c \cap V_\Theta \neq \phi$). Then, if we change coordinates in Θ so that one of the missing directions is along the first coordinate axis, we can use (79) to get (with an obvious abuse of notation)

$$|(\tilde{\theta}_n^{\mathrm{LML}} - \theta^*)_1| \geq \sqrt{\frac{2\log\log n}{n}}(E_1 - \epsilon) \quad \text{i.o. w.p.1} \tag{80}$$

Thus the pseudo-LML estimator must forever make excursions outside of L_q. Of course, this is true for L_{s^*} in particular, which proves Theorem 4.

5.2.3 Rates

It is instructive to compare not just the parameter values of Q^* and the LML estimator but also the associated rates.

Proposition 2. *If* $l_1(\theta) = R(\hat{P}_{X_1^n}, Q_\theta, D)$, *then*

$$l_1(\theta^*) - l_1(\tilde{\theta}_n^{LML}) \leq C^* \frac{\log\log n}{n} \quad \textit{eventually w.p.1} \tag{81}$$

Remark 9. In fact, $l_1(\theta^*) - l_1(\tilde{\theta}_n^{\mathrm{LML}}) = \Omega(\frac{\log\log n}{n})$, as can be seen from the first proof below.

First Proof. We will prove that there exist positive constants C_1 and C_2 such that

$$C_1\|\tilde{\theta}_n^{\mathrm{LML}} - \theta^*\|^2 \leq R(\hat{P}_{X_1^n}, Q^*, D) - R(\hat{P}_{X_1^n}, D) \leq C_2\|\tilde{\theta}_n^{\mathrm{LML}} - \theta^*\|^2. \tag{82}$$

Then, the upper bound in (78) implies Proposition 2.

From Theorem 1, we know that f is C^2 in θ if $P_\sigma \in \mathbb{S}(D)$. Hence expanding f in a Taylor series in its second argument about $\eta(\hat{\sigma_n}) = \tilde{\theta}_n^{\mathrm{LML}}$ yields

$$\begin{aligned} f(\hat{\sigma_n}, \theta^*) =& f(\hat{\sigma_n}, \tilde{\theta}_n^{\mathrm{LML}}) + \nabla_\theta f(\hat{\sigma_n}, \tilde{\theta}_n^{\mathrm{LML}}) \cdot (\theta^* - \tilde{\theta}_n^{\mathrm{LML}}) \\ &+ \tfrac{1}{2}(\theta^* - \tilde{\theta}_n^{\mathrm{LML}})^T J_2(\hat{\sigma_n})(\theta^* - \tilde{\theta}_n^{\mathrm{LML}}) + O(\|\theta^* - \tilde{\theta}_n^{\mathrm{LML}}\|)^3 \end{aligned} \tag{83}$$

where the subscript denotes the variable with respect to which the derivatives are taken, and $J_2(\hat{\sigma_n}) = \mathrm{Hess}_\theta(f(\hat{\sigma_n}, \tilde{\theta}_n^{\mathrm{LML}}))$.

$\tilde{\theta}_n^{\mathrm{LML}}$ lies in the interior of Θ for high enough n because consistency of the pseudo-LML estimator implies that it is close to θ^*, and we know Q^* is in the interior of the simplex from the definition of $\mathbb{S}(D)$. Since $\tilde{\theta}_n^{\mathrm{LML}}$ also minimizes $f(\hat{\sigma_n}, \cdot)$, $\nabla_\theta f(\hat{\sigma_n}, \tilde{\theta}_n^{\mathrm{LML}}) = 0$. Further, since $P_\sigma \in \mathbb{S}(D)$, we know that $J_2(\hat{\sigma_n})$ is a positive-definite matrix and hence invertible. Consequently, the quadratic form $(\theta^* - \tilde{\theta}_n^{\mathrm{LML}})^T J_2(\hat{\sigma_n})(\theta^* - \tilde{\theta}_n^{\mathrm{LML}})$ which represents the $J_2(\hat{\sigma_n})$-induced Euclidean norm is equivalent to the canonical Euclidean norm (induced by the identity matrix), and

$$\begin{aligned} &4C_1\|\tilde{\theta}_n^{\mathrm{LML}} - \theta^*\|^2 \\ &\leq (\theta^* - \tilde{\theta}_n^{\mathrm{LML}})^T J_2(\hat{\sigma_n})(\theta^* - \tilde{\theta}_n^{\mathrm{LML}}) \;=\; \|\theta^* - \tilde{\theta}_n^{\mathrm{LML}}\|^2_{J_2(\hat{\sigma_n})} \\ &\leq C_2\|\tilde{\theta}_n^{\mathrm{LML}} - \theta^*\|^2 \end{aligned} \tag{84}$$

for some positive constants C_1 (the factor 4 is chosen for convenience) and C_2.

Using (83) and (84), we now have

$$\begin{aligned} C_1\|\tilde{\theta}_n^{\mathrm{LML}} - \theta^*\|^2 &\leq \frac{1}{4}\|\theta^* - \tilde{\theta}_n^{\mathrm{LML}}\|^2_{J_2(\hat{\sigma_n})} \\ &\leq f(\hat{\sigma_n}, \theta^*) - f(\hat{\sigma_n}, \tilde{\theta}_n^{\mathrm{LML}}) \\ &= R(\hat{P}_{X_1^n}, Q^*, D) - R(\hat{P}_{X_1^n}, D) \end{aligned} \tag{85}$$

and

$$\begin{aligned} R(\hat{P}_{X_1^n}, Q^*, D) - R(\hat{P}_{X_1^n}, D) &\leq \|\theta^* - \tilde{\theta}_n^{\mathrm{LML}}\|^2_{J_2(\hat{\sigma_n})} \\ &\leq C_2\|\tilde{\theta}_n^{\mathrm{LML}} - \theta^*\|^2 \end{aligned} \tag{86}$$

provided n is large enough so that the second-order term in (83) dominates over the higher-order terms. □

Second Proof. This alternative proof of Proposition 2 uses an important inequality relating $R(P,Q,D)$ and the rate-distortion function. Recall that the rate function can be written in the form

$$R(P,Q,D) = \inf_W [I(X;Y) + D(Q_Y \| Q)]$$

By choosing W to be the optimal joint distribution, we have $I(X;Y) = R(P,D)$ and $Q_Y = Q^*$. Thus, for any Q,

$$R(P,Q,D) \leq R(P,D) + D(Q^* \| Q) \tag{87}$$

This implies the following for the sequence of empirical source distributions (with Q in the above set to $Q^* = \eta_D(P)$):

$$R(\hat{P}_{X_1^n}, Q^*, D) \leq R(\hat{P}_{X_1^n}, D) + D(\eta_D(\hat{P}_{X_1^n}) \| \eta_D(P)) \tag{88}$$

where the second term on the right is a measure of how different the optimal reproduction distributions corresponding to the real and empirical source distributions are.

Recall that the relative entropy between two nearby distributions belonging to a parametric family of probability measures on an alphabet can be expanded in a Taylor series in which the leading term is quadratic with a coefficient that depends on the Fisher information:

$$D(P_\sigma \| P) = \tfrac{1}{2}(\sigma - \sigma^*)^T J(\sigma^*)(\sigma - \sigma^*) + O(\|\sigma - \sigma^*\|^3) \tag{89}$$

This implies, for n large enough,

$$\begin{aligned} D(\eta_D(\hat{P}_{X_1^n}) \| \eta_D(P)) &\leq (\eta(\hat{\sigma_n}) - \eta(\sigma^*))^T J(\eta(\sigma^*))(\eta(\hat{\sigma_n}) - \eta(\sigma^*)) \\ &\leq C_5 \|\eta(\hat{\sigma_n}) - \eta(\sigma^*)\|^2 \end{aligned} \tag{90}$$

by the equivalence of Euclidean norms generated by quadratic forms involving non-singular matrices, and using the fact that $J(\theta^*)$ is invertible since $P \in \mathbb{S}(D)$. By (75) and the fact that $|Ax| \leq \|A\||x|$, there exists a constant C_4 such that

$$\|\eta(\hat{\sigma_n}) - \eta(\sigma^*)\| \leq C_4 \|\hat{\sigma_n} - \sigma^*\| \tag{91}$$

for n large enough. Plugging this into (88) and (90) gives

$$R(\hat{P}_{X_1^n}, Q^*, D) - R(\hat{P}_{X_1^n}, D) \leq C_5 C_4 \|\hat{\sigma_n} - \sigma^*\|^2 \tag{92}$$

which provides an alternative demonstration of (81) and hence of the pseudo-LMDL part of Theorem 2, since the squared norm is the sum of squares of the components and hence obeys a scalar LIL. □

Remark 10. The second proof emphasizes the role of the *multivariate* LIL in proving Theorem 4. Neither a scalar LIL nor a multivariate LIL which merely gave information about specific limit points rather than specified the entire limit set would have sufficed to analyze the pseudo-LML estimator satisfactorily, though as we saw above Proposition 2 (and hence the pseudo-LMDL estimator) can be studied in a simple way not involving Berning's LIL. Indeed, suppose $\Gamma = D\eta(\sigma^*)$ is non-singular, then $|\Gamma^{-1}y| \leq \|\Gamma^{-1}\||y| \Rightarrow |\Gamma x| \geq \frac{1}{\|\Gamma^{-1}\|}|x|$, so that (75) implies

$$C_3 \|\hat{\sigma_n} - \sigma^*\| \leq \|\eta(\hat{\sigma_n}) - \eta(\sigma^*)\| \tag{93}$$

for n large enough. This can be combined with the lower bound of (85) to get

$$\begin{aligned} R(\hat{P}_{X_1^n}, Q^*, D) - R(\hat{P}_{X_1^n}, D) &\geq C_1 \|\tilde{\theta}_n^{\mathrm{LML}} - \theta^*\|^2 \\ &\geq C_1 C_3 \|\hat{\sigma_n} - \sigma^*\|^2 \end{aligned} \tag{94}$$

which, using a scalar LIL, would yield

$$R(\hat{P}_{X_1^n}, Q^*, D) - R(\hat{P}_{X_1^n}, D) \geq C_* \frac{\log\log n}{n} > 0 \quad \text{i.o. w.p.1} \tag{95}$$

for some $C_* > 0$. In other words, this only proves the fluctuation property of the pseudo-LML estimator when the simplest set L_{s^*} containing θ^* is exactly $\{\theta^*\}$. It does indicate that $\tilde{\theta}_n^{\mathrm{LML}}$ fluctuates around θ^* for ever, but does not show that this never-ending fluctuation happens in *every possible direction*, which is necessary to prove Theorem 5.

5.3 Behavior of the pseudo-LMDL estimator

Set

$$\begin{aligned} l_1(\theta) &= R(\hat{P}_{X_1^n}, Q_\theta, D), \\ l_2(\theta) &= k(\theta) c(n), \\ l(\theta) &= l_1(\theta) + l_2(\theta), \end{aligned} \tag{96}$$

where the complexity coefficient $k(\theta)$ is defined as $k(\theta) = \min\{1 \leq i \leq s : \theta \in L_i\}$ in terms of the nested sequence of models. Recall that the pseudo-LMDL estimator is defined as $\tilde{\theta}_n^{\mathrm{LMDL}} = \arg\min_{\theta\in\Theta} l(\theta)$. By Proposition 2, we have for any $\delta > 0$,

$$l_1(\theta^*) < l_1(\tilde{\theta}_n^{\mathrm{LML}}) + \delta c(n) \quad \text{eventually w.p.1.} \tag{97}$$

Since we will use a sample-path argument, let us fix $\delta \in (0, 1)$ and then fix our attention on any realization for which (97) holds. For this realization, define the sequence of integers $\{\alpha_n\}$ by

$$\alpha_n = k(\tilde{\theta}_n^{\mathrm{LML}}) - k(\theta^*). \tag{98}$$

The sequence $\{\alpha_n\}$ is the union of the subsequences defined by the index sets $I_- = \{n : \alpha_n \leq 0\}$ and $I_+ = \{n : \alpha_n > 0\}$. When $n \in I_-$, $k(\tilde{\theta}_n^{\mathrm{LML}}) \leq k(\theta^*)$, and hence

$$k(\tilde{\theta}_n^{\mathrm{LMDL}}) \leq k(\theta^*), \tag{99}$$

because the only way the LMDL estimator can improve on the LML estimator is through the penalty function.

However, from the previous section, $k(\tilde{\theta}_n^{\mathrm{LML}}) = m - 1$ i.o. for almost every realization. For the fixed realization of interest, this indicates that the index set I_+ is non-empty, and this is the case for which something remains to be proved. When $n \in I_+$, we obtain a relationship between $l(\theta^*)$ and $l(\tilde{\theta}_n^{\mathrm{LML}})$ using (97):

$$\begin{aligned} l(\theta^*) &= l_1(\theta^*) + c(n) k(\theta^*) \\ &< l_1(\tilde{\theta}_n^{\mathrm{LML}}) + c(n)[k(\theta^*) + \delta] \\ &= \left[l_1(\tilde{\theta}_n^{\mathrm{LML}}) + c(n) k(\tilde{\theta}_n^{\mathrm{LML}}) \right] - c(n)[\alpha_n - \delta] \\ \Rightarrow \quad l(\theta^*) &< l(\tilde{\theta}_n^{\mathrm{LML}}) - c(n)[\alpha_n - \delta] \end{aligned} \tag{100}$$

Since, by definition, $l(\tilde{\theta}_n^{\mathrm{LMDL}}) \le l(\theta)$ for all $\theta \in \Theta$, we have, in fact, a relationship between $l(\tilde{\theta}_n^{\mathrm{LMDL}})$ and $l(\tilde{\theta}_n^{\mathrm{LML}})$:

$$l(\tilde{\theta}_n^{\mathrm{LMDL}}) \le l(\theta^*) < l(\tilde{\theta}_n^{\mathrm{LML}}) - c(n)[\alpha_n - \delta] \tag{101}$$

Now,

$$\begin{aligned} l_2(\tilde{\theta}_n^{\mathrm{LMDL}}) - l_2(\tilde{\theta}_n^{\mathrm{LML}}) &= c(n)[k(\tilde{\theta}_n^{\mathrm{LMDL}}) - k(\tilde{\theta}_n^{\mathrm{LML}})] \\ &< \left[l_1(\tilde{\theta}_n^{\mathrm{LML}}) - l_1(\tilde{\theta}_n^{\mathrm{LML}})\right] - c(n)[\alpha_n - \delta] \\ &\le -c(n)[\alpha_n - \delta] \end{aligned} \tag{102}$$

using the fact that $l_1(\tilde{\theta}_n^{\mathrm{LML}}) \le l_1(\theta)$ for all $\theta \in \Theta$. Thus

$$k(\tilde{\theta}_n^{\mathrm{LMDL}}) < k(\tilde{\theta}_n^{\mathrm{LML}}) - [\alpha_n - \delta]. \tag{103}$$

Since $\delta < 1$ and $k(\cdot)$ must be an integer, we have shown that

$$k(\tilde{\theta}_n^{\mathrm{LMDL}}) \le k(\tilde{\theta}_n^{\mathrm{LML}}) - \alpha_n = k(\theta^*). \tag{104}$$

Combining (99) and (104), we have that $k(\tilde{\theta}_n^{\mathrm{LMDL}}) \le k(\theta^*)$ eventually along the fixed realization, and hence along almost every realization.

Finally by the definition of $k(\cdot)$ and the fact that the sequence of sets is nested, we have

$$\tilde{\theta}_n^{\mathrm{LMDL}} \in L_{k(\tilde{\theta}_n^{\mathrm{LMDL}})} \subset L_{k(\theta^*)}. \tag{105}$$

Hence, $\tilde{\theta}_n^{\mathrm{LMDL}}$ lies in L_{s^*}, which proves Theorem 5.

5.4 The LMDL estimator

Observe that by Theorem 2,

$$\begin{aligned} \frac{1}{n}|L_n(Q^*, X_1^n) - L_n(Q_{\hat{\theta}_n^{\mathrm{LML}}}, X_1^n)| &= |R(\hat{P}_{X_1^n}, Q^*, D) - R(\hat{P}_{X_1^n}, Q_{\hat{\theta}_n^{\mathrm{LML}}}, D) + O\left(\frac{1}{n}\right)| \\ &\le |R(\hat{P}_{X_1^n}, Q^*, D) - R(\hat{P}_{X_1^n}, Q_{\tilde{\theta}_n^{\mathrm{LML}}}, D)| + \\ &\quad |R(\hat{P}_{X_1^n}, Q_{\tilde{\theta}_n^{\mathrm{LML}}}, D) - R(\hat{P}_{X_1^n}, Q_{\hat{\theta}_n^{\mathrm{LML}}}, D)| + O\left(\frac{1}{n}\right) \\ &\le O\left(\frac{\log\log n}{n}\right) + D(Q_{\tilde{\theta}_n^{\mathrm{LML}}} \| Q_{\hat{\theta}_n^{\mathrm{LML}}}) + O\left(\frac{1}{n}\right) \end{aligned} \tag{106}$$

where we used (81) to estimate the first term, and (87) to obtain the relative entropy bound. To estimate the second term, we need the following lemma.

Lemma 4. *Let $g_1(\theta)$ and $g_2(\theta)$ be two real-valued functions on $\Theta = \mathbb{R}^m$. Suppose g_1 has a unique minimizer denoted by θ_1, and that $\sup_\theta |g_1 - g_2| < \Delta$. If Δ is small enough so that $g_1(\theta') - g_1(\theta_1) > 2\Delta$ for any local minimizer $\theta' \neq \theta_1$ of g_1, and if g_1 is C^2 with non-singular Hessian, then*

$$\|\theta_1 - \theta_2\|^2 \le \textit{(const.)}\Delta \tag{107}$$

Proof. Since g_1 and g_2 are at most Δ apart, we have that the minimum value of g_2 satisfies

$$g_2(\theta_2) \leq g_2(\theta_1) \leq g_1(\theta_1) + \Delta, \tag{108}$$

where θ_2 is any minimizer of g_2. Note that

$$\begin{aligned} \theta_2 &\in \{\theta : g_2(\theta) \leq g_1(\theta_1) + \Delta\} \\ &\subset \{\theta : g_1(\theta) - \Delta \leq g_1(\theta_1) + \Delta\} \\ &= \{\theta : g_1(\theta) \leq g_1(\theta_1) + 2\Delta\} \\ &= \Theta_{\text{allowed}}, \text{ say.} \end{aligned} \tag{109}$$

Since $g_1(\theta') > g_1(\theta_1) + 2\Delta$ for every local minimizer $\theta' \neq \theta_1$, $\theta' \notin \Theta_{\text{allowed}}$ and the continuity of g_1 implies that g_2 must achieve its minimum in a neighborhood of θ_1 (because if no local non-global minimum of g_1 is close enough to $g_2(\theta_2)$, then neither is any point between two adjacent local, non-global minima). It remains only to determine the size of this neighborhood. We observe that for $\theta_2 \in (\theta_1 - \delta, \theta_1 + \delta)$, g_1 when Taylor expanded has a quadratic leading term since its first derivative vanishes at the minimum:

$$g_1(\theta_2) - g_1(\theta_1) = \tfrac{1}{2}(\theta_2 - \theta_1)^T \text{Hess}(g_1(\theta_1))(\theta_2 - \theta_1) + O(\theta_2 - \theta_1)^3. \tag{110}$$

Since $\theta_2 \in \Theta_{\text{allowed}}$, we need to choose δ small enough so that $|g_1(\theta_2) - g_1(\theta_1)| < 2\Delta$, which can be done because g_1 is continuous. Then,

$$2\Delta \geq \tfrac{1}{2}(\theta_2 - \theta_1)^T \text{Hess}(g_1(\theta_1))(\theta_2 - \theta_1) \geq C\|\theta_2 - \theta_1\|^2 \tag{111}$$

since in Euclidean space, a norm generated by a quadratic form involving a symmetric (invertible) matrix is equivalent to the original norm with bounds given by the smallest and largest eigenvalues. □

Let us apply Lemma 4 to the functions $g_1(\theta) = \frac{1}{n}\tilde{L}_n(Q_\theta, X_1^n) = R(\hat{P}_{X_1^n}, Q_\theta, D) + \frac{1}{2}\log n$ and $g_2(\theta) = \frac{1}{n}L_n(Q_\theta, X_1^n)$. By Theorem 2, there exists $C < \infty$ such that

$$\sup_\theta |g_1(\theta) - g_2(\theta)| \leq \frac{C}{n} \quad \text{eventually w.p.1.} \tag{112}$$

Since $P \in \mathbb{S}(D)$, we also have $\hat{P}_{X_1^n} \in \mathbb{S}(D)$ for large enough n, and Lemma 2 tells us that $\text{Hess}(g_1(\theta_1))$ is not only well-defined but is positive-definite because g_1 is minimized at θ_1. Furthermore the unique minimizer of g_1 is $\tilde{\theta}_n^{\text{LML}}$. Thus, the conditions of Lemma 4 are verified, and for any minimizer $\hat{\theta}_n^{\text{LML}}$ of g_2 we have

$$\|\tilde{\theta}_n^{\text{LML}} - \hat{\theta}_n^{\text{LML}}\|^2 = O\left(\frac{1}{n}\right) \quad \text{w.p.1.} \tag{113}$$

Combining this with (89) leads to

$$D(\eta_D(\hat{P}_{X_1^n})\|Q_{\hat{\theta}_n^{\text{LML}}}) = O\left(\frac{1}{n}\right) \quad \text{w.p.1.} \tag{114}$$

Therefore, going back to (106), we have

$$\frac{1}{n}|L_n(Q^*, X_1^n) - L_n(Q_{\hat{\theta}_n^{\text{LML}}}, X_1^n)| = O\left(\frac{\log\log n}{n}\right) \quad \text{eventually w.p.1.} \tag{115}$$

This is the equivalent of Proposition 2 for the LML estimator. Repeating the argument of the previous section with the pseudo-lossy likelihood $l_1(\theta)$ replaced by the lossy likelihood $l_1'(\theta) = L_n(Q_\theta, X_1^n)$ completes the proof of Theorem 6.

6 Conclusion

We have used three kinds of length functions to characterize the performance of lossy codes. These are:

1. the actual length of the codeword $\text{len}_n(X_1^n)$;
2. the idealized lossy codelength of the data given the coding distribution, or the lossy likelihood of a coding distribution given the data, namely $-\log Q^n(B(X_1^n, D))$;
3. the idealized pseudo-lossy codelength of the data given the coding distribution, or the pseudo-lossy likelihood of a coding distribution given the data, namely $nR(\hat{P}_{X_1^n}, Q, D)$.

In essence, much of the work is devoted to understanding the links between these various notions of codelength and how we can use these links along with the successively greater tractability of the last two notions to shed light on the problem of universal lossy coding. The main contribution is the result that even for lossy compression, appropriately defined MDL codes work better than codes corresponding to maximum likelihood estimators in the sense that they choose the optimal model eventually. More generally, this work extends the emerging statistically-motivated theoretical framework for lossy coding suggested by [17].

There are many problems that this work leaves open. This ranges from problems of generalization, such as to non-finite alphabets or beyond the i.i.d. case (all of which involve dealing with daunting technicalities), to conceptual problems, such as the question of whether there is an analog of the Cramer-Rao bound for "lossy" estimators. The most important open area, though, is the question of whether and how the principles discussed can be used to construct practical codes.

A Proof of Theorem 3

Unlike for Theorem 2, we cannot just use the non-uniform version of Theorem 3 to prove the uniform version, since we do not have explicit tractable bounds on the error term. However, the structure of the proof is similar to the non-uniform case, though some additional ingredients like a uniform law of the iterated logarithm are required. A significantly more intricate use of tools from the Vapnik-Chervonenkis theory should yield the theorem for compact source alphabets (this would involve performing a series of increasingly fine finite discretizations of the alphabet, using the finite alphabet result proved here, and justifying the approximation via discretization using smoothness arguments); however we do not detail that proof. It is not clear how to extend this to general alphabets like the real numbers.

The structure of the proof is as follows:

1. First we will show that for n large enough,

$$R(P, Q_\theta, D) - R(\hat{P}_{X_1^n}, Q_\theta, D) + \frac{1}{n}\sum_{i=1}^{n} g_\theta(X_i) = \inf_{|\kappa|<\delta} \left[\frac{1}{n}\sum_{i=1}^{n} \{f(\theta, \kappa, X_i) - f(\theta, 0, X_i)\} \right] \quad (116)$$

   where

$$f(\theta, \kappa, a) \equiv \Lambda_a(Q_\theta, \lambda_\theta + \kappa) - (\lambda_\theta + \kappa)D \quad (117)$$

2. We will then perform a Taylor expansion and a sequence of manipulations to reduce Theorem 3 to a pair of propositions.
3. We will then prove these propositions, using a uniform law of the iterated logarithm.

A.1 Part 1

We need the following lemmas.

Lemma A.1: If Q_{θ_0} lies in the interior of the simplex $\mathbb{P}(\hat{A})$ of i.i.d. probability measures on finite alphabet $\hat{A}$, then:

(1) $D_{\min}(\theta_0) = 0$

(2) There exists $\delta > 0$ such that $D^{(n)}_{\min}(\theta) \to D_{\min}(\theta)$ uniformly w.p.1 for $\theta \in \mathrm{B}(\theta_0, \delta)$.

(3) If $D > 0$, then there exists $\delta > 0$ such that $\lambda^{(n)}_\theta \to \lambda_\theta$ uniformly w.p.1 for $\theta \in B(\theta_0, \delta)$.

Proof: Let

$$m_\theta(a) \equiv \operatorname*{ess\,inf}_{Y \sim Q_\theta} \rho(a, Y) = \min_{y \in \mathrm{supp}(Q_\theta)} \rho(a, y) \tag{118}$$

so that $D_{\min}(\theta) = E_P[m_\theta(X)]$. Since Q_{θ_0} lies in the interior of $\mathbb{P}(A)$, $\mathrm{supp}(Q_{\theta_0}) = A$ which implies $m_{\theta_0}(a) = 0$ for each $a \in A$. Thus, irrespective of what P is, we have $D_{\min}(\theta_0) = 0$.

If we choose δ so that the neighborhood $B(\theta_0, \delta)$ lies in the interior of the simplex, then by the above, we have $D^{(n)}_{\min}(\theta) = D_{\min}(\theta) = 0$ for every n and every θ in this neighborhood. Thus the uniform convergence in (2) is trivial.

It is evident that we can pick $\delta > 0$ so that $D > D_{\min}(\theta) = 0$ for all θ in the neighborhood $B(\theta_0, \delta)$. Thus the definitions of $\lambda^{(n)}_\theta$ and λ_θ make sense. Now suppose

$$\limsup_{n \to \infty} \sup_{\theta \in \mathrm{B}(\theta_0, \delta)} (\lambda^{(n)}_\theta - \lambda_\theta) \geq \epsilon \tag{119}$$

Then there exists a subsequence n_k along which $\sup_{\theta \in \mathrm{B}(\theta_0, \delta)} (\lambda^{(n)}_\theta - \lambda_\theta) \geq \frac{\epsilon}{2}$ Focussing on this subsequence, we choose θ so that $(\lambda^{(n)}_\theta - \lambda_\theta) \geq \frac{\epsilon}{4}$. Now,

$$\begin{aligned}
D &= \liminf_{n \to \infty} \Lambda'(\hat{P}_{X_1^n}, Q_\theta, \lambda^{(n)}_\theta) \\
&\leq \limsup_{n \to \infty} \Lambda'(\hat{P}_{X_1^n}, Q_\theta, \lambda_\theta - \frac{\epsilon}{4}) \\
&= \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \Lambda'_{X_i}(Q_\theta, \lambda_\theta - \frac{\epsilon}{4}) \\
&= \Lambda'(Q_\theta, \lambda_\theta - \frac{\epsilon}{4}) \\
&< \Lambda'(Q_\theta, \lambda_\theta) = D
\end{aligned}$$

which shows that (119) leads to a contradiction. Similarly, the assumption

$$\liminf_{n \to \infty} \inf_{\theta \in \mathrm{B}(\theta_0, \delta)} (\lambda^{(n)}_\theta - \lambda_\theta) \leq \epsilon \tag{120}$$

also leads to a contradiction. Together, these prove the lemma.

Choose $N = N(X_1^\infty, \delta)$ such that $|\lambda^{(n)}_\theta - \lambda_\theta| < \delta$ for every $n > N$ and for every $\theta \in \mathrm{B}(\theta_0, \delta)$. By Lemma A.1, we know that $N < \infty$ w.p.1. Using the definitions of R and g_θ, we can rewrite the LHS of (116) for $n > N$ as

$$
\begin{aligned}
&[\Lambda^*(P,Q_\theta,D)-\Lambda^*(\hat{P}_{X_1^n},Q_\theta,D)]+[\Lambda(P,Q_\theta,\lambda_\theta)-\Lambda(\hat{P}_{X_1^n},Q_\theta,\lambda_\theta)]\\
&=^{(a)}\ \lambda_\theta D-\Lambda(P,Q_\theta,\lambda_\theta)-\sup_{\kappa\in(-\delta,\delta)}[(\lambda_\theta+\kappa)D-\Lambda(\hat{P}_{X_1^n},Q_\theta,\lambda_\theta+\kappa)]\\
&\qquad\qquad+\Lambda(P,Q_\theta,\lambda_\theta)-\Lambda(\hat{P}_{X_1^n},Q_\theta,\lambda_\theta)\\
&=\inf_{|\kappa|<\delta}-[(\lambda_\theta+\kappa)D-\Lambda(\hat{P}_{X_1^n},Q_\theta,\lambda_\theta+\kappa)]+\lambda_\theta D-\Lambda(\hat{P}_{X_1^n},Q_\theta,\lambda_\theta)\\
&=\inf_{|\kappa|<\delta}\left[\frac{1}{n}\sum_{i=1}^n\{f(\theta,\kappa,X_i)-f(\theta,0,X_i)\}\right]
\end{aligned}
\tag{121}
$$

which proves (116). Note that in the equality (a), the restriction of the supremum in the definition of $R(\hat{P}_{X_1^n},Q_\theta,D)$ to the small interval is valid for $n>N(\omega)$ for a.s-ω. We are also using the fact here that $\Lambda(\hat{P}_{X_1^n},Q_\theta,\lambda_\theta)=\frac{1}{n}\sum_{i=1}^n\Lambda_{X_i}(Q_\theta,\lambda_\theta)$, which is a consequence of the definitions.

A.2 Part 2

Now, by Taylor's theorem, for some $\psi_n(\theta,\kappa)$ between $-\kappa$ and κ,

$$
\left[\frac{1}{n}\sum_{i=1}^n\{f(\theta,\kappa,X_i)-f(\theta,0,X_i)\}\right]=\kappa A_n(\theta)+\frac{\kappa^2}{2}B_n(\theta,\psi_n(\theta,\kappa)) \tag{122}
$$

where, if we use $'$ to denote differentiation with respect to λ,

$$
A_n(\theta)\equiv\frac{1}{n}\sum_{i=1}^n\zeta(\theta,X_i) \tag{123}
$$

$$
\zeta(\theta,a)\equiv\frac{\partial}{\partial\kappa}\bigg|_{\kappa=0}f(\theta,\kappa,a)=\Lambda_a'(Q_\theta,\lambda_\theta)-D \tag{124}
$$

$$
B_n(\theta,\gamma)\equiv\frac{1}{n}\sum_{i=1}^n\frac{\partial^2}{\partial\kappa^2}\bigg|_{\kappa=\gamma}f(\theta,\kappa,a)=\frac{1}{n}\sum_{i=1}^n\Lambda_a''(Q_\theta,\lambda_\theta+\gamma) \tag{125}
$$

By combining (116) and (122), it is clear that to prove Proposition 2, it suffices to show that

$$
\sup_\theta\left|\inf_{|\kappa|<\delta}\left[\kappa A_n(\theta)+\frac{\kappa^2}{2}B_n(\theta,\psi_n(\theta,\kappa))\right]\right|=O(\frac{\log\log n}{n}) \tag{126}
$$

Note that since the expression in square brackets is 0 for $\kappa=0$, the infimum must necessarily be non-positive. In other words, we only need to prove a one-sided version of (126).

The following simple estimate using completion-by-squares is useful:

$$
\begin{aligned}
&\kappa A_n(\theta)+\frac{\kappa^2}{2}B_n(\theta,\psi_n(\theta,\kappa))\\
&=\left(\sqrt{\frac{B_n(\theta,\psi_n(\theta,\kappa))}{2}}\kappa+\frac{A_n}{\sqrt{2B_n(\theta,\psi_n(\theta,\kappa))}}\right)^2-\frac{A_n^2}{2B_n(\theta,\psi_n(\theta,\kappa))}\\
&\geq-\frac{A_n^2}{2B_n(\theta,\psi_n(\theta,\kappa))}
\end{aligned}
\tag{127}
$$

This estimate implies

$$\inf_{|\kappa|<\delta}\left[\kappa A_n(\theta)+\frac{\kappa^2}{2}B_n(\theta,\psi_n(\theta,\kappa))\right]\geq -\frac{A_n^2}{2\inf_{|\kappa|<\delta}B_n(\theta,\kappa)} \tag{128}$$

since we know that $|\psi_n(\theta,\kappa)|<\delta$.

Combining (126) and (128) and noting the comment after the former, we see that to prove Proposition 2, it is enough to show that for n large enough and for some constant $C<\infty$ independent of θ,

$$\sup_{\theta\in \mathrm{B}(\theta^*,\delta')}\frac{A_n^2}{2\inf_{|\kappa|<\delta}B_n(\theta,\kappa)}\leq\frac{C\log\log n}{n} \tag{129}$$

or equivalently

$$\sup_{\theta\in \mathrm{B}(\theta^*,\delta')}\frac{(\frac{nA_n^2}{\log\log n})}{2(\inf_{|\kappa|<\delta}B_n(\theta,\kappa))}\leq C \tag{130}$$

To show this, it is sufficient to prove the following 2 statements:
Proposition A.1:

$$\limsup_{n\to\infty}\sup_{\theta\in \mathrm{B}(\theta^*,\delta')}\frac{nA_n^2}{\log\log n}<\infty \quad \text{w.p.1} \tag{131}$$

Proposition A.2:

$$\liminf_{n\to\infty}\inf_{\theta\in \mathrm{B}(\theta^*,\delta')}\inf_{|\kappa|<\delta}B_n(\theta,\kappa)>0 \quad \text{w.p.1} \tag{132}$$

A.3 Part 3

We need the following lemmas.

Lemma A.2: (Properties of $\zeta(\cdot,\cdot)$) If $\hat{A}$ is finite and the distortion measure ρ is bounded,

1. The collection of functions $\mathbb{F}=\{\zeta(\theta,\cdot):\theta\in \mathrm{B}(\theta^*,\delta')\}$ is a class of measurable functions from A to $\mathbb{R}$. Also, $\zeta(\theta,X)$ is a real-valued random variable with mean 0 for any θ.
2. For fixed $x\in A$, $\zeta(\theta,x)$ is a continuous function of θ.
3. $\mathbb{F}$ forms a bounded separable subset of $L^2(A,P)$.
4. Define $S(x)=\sup_{\theta\in \mathrm{B}(\theta^*,\delta')}|\zeta(\theta,x)|$ and $L(z)=\log(\max(z,e))$. If A is finite, then $S(X)$ is a bounded random variable, and

$$E\left[\frac{S^2(X)}{LLS(X)}\right]<\infty \tag{133}$$

Proof:

1. Recall

$$\zeta(\theta, a) = \frac{\partial}{\partial\lambda}|_{\lambda_\theta}(\log E_\theta[e^{\lambda\rho(a,Y)}]) - D = \frac{E_\theta[\rho(a,Y)e^{\lambda_\theta\rho(a,Y)}]}{E_\theta[e^{\lambda_\theta\rho(a,Y)}]} - D \tag{134}$$

where the interchange of derivative and expectation was permissible because the finiteness of the alphabet causes E_θ to be just a weighted sum. Since ρ is bounded, all the summands in the numerator are finite, while those in the denominator are positive. Hence ζ is a finite-valued function.

For $\zeta(\theta, X)$ to be a random variable, we need $\zeta(\theta, \cdot)$ to be a measurable function.

To see that $\zeta(\theta, X)$ has mean zero, we merely note that as before we can interchange derivative and expectation, so

$$E_P[\zeta(\theta, X)] = \frac{\partial}{\partial\lambda}E_P[\Lambda_X(Q_\theta, \lambda_\theta)] - D = \frac{\partial}{\partial\lambda}\Lambda(P, Q_\theta, \lambda_\theta) - D = 0$$

2. $h(\lambda, \theta) = \frac{\partial}{\partial\lambda}\Lambda_x(Q_\theta, \lambda)$ is a continuous function of both θ and λ. Noting that $\zeta(\theta, x) = h(\lambda_\theta, \theta) - D$, it only remains to observe:

$$|h(\lambda_\theta, \theta) - h(\lambda_{\theta'}, \theta')| \leq |h(\lambda_\theta, \theta) - h(\lambda_{\theta'}, \theta)| + |h(\lambda_{\theta'}, \theta) - h(\lambda_{\theta'}, \theta')|$$

Q.E.D.

3. The bounded distortion function implies that $d_k(P, Q)$ is bounded for any P and Q. Consequently,

$$\bar{d_2} \equiv \sup_{\theta\in \mathrm{B}(\theta^*,\delta')} E_P[\zeta(\theta, X)]^2 < \infty$$

and thus $\mathbb{F}$ is bounded in $L^2(A, P)$. $\mathbb{F}$ is also a separable subset since continuity of $\zeta(\theta, x)$ in θ implies that considering only rational θ yields a dense subclass.

4. Note that

$$\zeta(\theta, a) = \Lambda'_a(Q_\theta, \lambda_\theta) - D \in [D_{min}(a, \theta) - D, D_{av}(a, \theta) - D]$$

if we define $D_{min}(a, \theta) = D_{min}(1_a, Q_\theta)$ and $D_{av}(a, \theta) = D_{av}(1_a, Q_\theta)$, and so $\zeta(\cdot, a)$ is bounded. Further, by continuity of the bounds in θ, and the fact that $S(\cdot)$ is the supremum of $\zeta(\theta, \cdot)$ over a ball of θ's that is contained in a compact set in Euclidean space, it is clear that $S(a)$ is a finite number for each a. Finiteness of A implies that $S(\cdot)$ is a bounded function. This, together with the fact that $LLS(x) \geq 1$ by definition, yields (43).

Lemma A.3: If A is finite, $\mathbb{F}$ is a countably determined Vapnik-Chervonenkis graph class.

Proof. Consider the set of subgraphs of $\mathbb{F}$. Each subgraph is a point-subset of a set of $|A| = m$ lines, namely $A \times \mathbb{R}$. We need to show that for some j, NO j-element subset of $A \times \mathbb{R}$ is shattered by the subgraph class of $\mathbb{F}$.

First we note that by the continuity of ζ (Lemma A.2), the extremal points of the subgraphs on each of the m lines are continuous in θ; thus they populate an interval on the real line. Let $j = m + 1$, so that at least one line has two points of the set to be shattered. Consider this line. If the 2 points lie on the same side of 0, then the one farther away cannot be isolated by any subgraph. If they lie on opposite sides of 0, the pair set cannot be isolated by any subgraph. □

Proof of Proposition A.1: From Lemmas A.2 and A.3, we see that $\mathbb{F}$ satisfies the conditions of Alexander and Talagrand, and thus a uniform law of the iterated logarithm holds for the i.i.d.variables $\zeta(\theta, X)$. □

Proof of Proposition A.2:

$$\begin{aligned}\liminf_{n\to\infty} \inf_{\theta\in \mathrm{B}(\theta^*,\delta')} \inf_{|\kappa|<\delta} B_n(\theta,\kappa) &\geq \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \left\{ \inf_{\theta\in \mathrm{B}(\theta^*,\delta')} \inf_{|\kappa|<\delta} \Lambda''_{X_i}(Q_\theta, \lambda_\theta+\kappa) \right\} \\ &= E_P\left[\inf_{\theta\in \mathrm{B}(\theta^*,\delta')} \inf_{|\kappa|<\delta} \Lambda''_{X_i}(Q_\theta, \lambda_\theta+\kappa) \right] \quad \text{w.p.1}\end{aligned} \tag{135}$$

where we used Birkhoff's ergodic theorem. We want to show that this is positive if we make δ and δ' small enough. But this follows from continuity of $\Lambda''_{X_i}(Q_\theta, \lambda_\theta+\kappa)$ in θ and κ (see Lemma A.2). □

B Connections to the Method of Types

B.1 Background

In a series of papers ([35], [32], [33]), Yang, Zhang and Wei pursued the use of the method of types to study lossy data compression with known and unknown statistics (the latter referring to the analysis of universal lossy codes). This appendix uses their comprehensive framework to make some comments on second-order properties of the lossy likelihood (cf., Section 3).

First, we outline the notation needed for the rest of this section. If $M = \{e_1, ..., e_m\}$, then $t \in \mathcal{P}(M)$ is an n-type if $t(e) \in \{0, \frac{1}{n}, \frac{2}{n}, ..., 1\}$ for each $e \in M$. The set of all n-types of M is denoted $T_n(M)$.

The type of a string $z_1^n \in M^n$ is $t(z_1^n) = (t(z_1^n, e_1), t(z_1^n, e_2), ..., t(z_1^n, e_m))$, where $t(z_1^n, e_i) = \frac{1}{n}|\{j : z_j = e_i\}|$ is the fraction of entries in z_1^n at which e_i occurs. The type class of a type $t \in T_n(M)$ is the set of strings

$$T_M^n(t) = \{z_1^n \in M^n : t(z_1^n) = t\} \tag{136}$$

If $t(e_i) > 0$ for every $e_i \in M$, the following facts hold:

$$\begin{aligned}\log |T_M^n(t)| &= nH(t) - \frac{m-1}{2}\log n + O(1) \\ -\log p^n(T_M^n(t)) &= nD(t\|p) + \frac{m-1}{2}\log n + O(1)\end{aligned} \tag{137}$$

where the $O(1)$ term in both expressions can be uniformly bounded over a set of types whose components are uniformly bounded away from 0.

With these basic notions, the theory of lossy data compression for finite alphabets can be cast in the language of types. Recall that the source alphabet is denoted A and the reproduction alphabet by $\hat{A}$, and that for any $x_1^n \in A^n$, we defined $B(x_1^n, D) = \{y_1^n \in \hat{A}^n : \rho_n(x_1^n, y_1^n) \leq D\}$. The main point of difference between the papers cited above and the way in which we use types is that we study distortion balls consisting of strings on the reproduction alphabet, whereas [35], [32] and [33] study distortion balls consisting of strings on the source

alphabet. In the rest of this appendix, we use t to denote a probability distribution on A, r to denote a probability distribution on $\hat{A}$, and s to denote a probability distribution on $A \times \hat{A}$.

For $r \in T_n(\hat{A})$, define

$$B(x_1^n, r, D) = B(x_1^n, D) \cap T_Y^n(r) \tag{138}$$

Since we are using a single-letter distortion function, symmetry implies that the cardinality of this set of strings depends only on $t(x_1^n)$; thus we define

$$F_n(t, r, D) = |B(x_1^n, r, D)|, \quad \text{where } x_1^n \in T_X^n(t) \tag{139}$$

As in [35], let the "upper joint entropy" and the "lower mutual information" be defined by

$$H_u(t, r, D) = \sup_{s \in S(t,r,D)} H(s) \tag{140}$$

and

$$I_\ell(t, r, D) = \inf_{s \in S(t,r,D)} I(t; r) = H(t) + H(r) - H_u(t, r, D) \tag{141}$$

where

$$S(t, r, D) = \{ s \in \mathcal{P}(A \times \hat{A}) : E_s \rho(X, Y) \leq D, \text{ and } s \text{ has } t \text{ and } r \text{ as its marginals } \} \tag{142}$$

Then, following the computations done in [35] and the refinements of the same in [36], it is easy to see the following result.

Theorem B.1: If $K = |\hat{A}|$, then for sufficiently large n, and for all (t, r, D) in a neighborhood of (t_0, r_0, D_0) such that t and r are n-types,

$$\log F_n(t, r, D) = nH_u(t, r, D) - nH(t) - \frac{K}{2} \log n + O(1) \tag{143}$$

Also,

$$-\log q^n(B(x_1^n, r, D)) = nD(r\|q) + nI_\ell(t, r, D) + \frac{K}{2} \log n + O(1) \tag{144}$$

where $x_1^n \in T_X^n(t)$, and the $O(1)$ error term depends on r and t but not on q.

B.2 The second-order generalized AEP using types

Having stated Theorem B.1, we can now use rough heuristic arguments to obtain the second-order generalized AEP. While these arguments can be made rigorous, we do not describe the laborious computations that would entail since this section merely provides an alternative proof for a result that has already been rigorously proven in Section 3.

According to Theorem B.1, $q^n(B(X_1^n, r, D)) = e^{-nf(r)}$, where

$$f(r) = D(r\|q) + I_\ell(\hat{P}_{X_1^n}, r, D) + \frac{K \log n}{2n} + O\left(\frac{1}{n}\right) \tag{145}$$

Thus we have

$$\begin{aligned} q^n(B(x_1^n, D)) &= \Sigma_{r \in T_n(\hat{A})} e^{-nf(r)} \\ &\approx n^{K-1} \int_{r \in \mathcal{P}(\hat{A})} e^{-nf(r)} dr \\ &\approx n^{K-1} e^{-nf(\hat{r})} \end{aligned} \tag{146}$$

where we approximated the sum by an integral, and then used Laplace's method of integration to estimate the value of the integral. Here, $\hat{r}$ is the minimizer of $f(r)$, which is $O(\frac{1}{n})$ close to what is known as the "favorite type" (assuming f is differentiable and using Proposition 4). Since the error is irrelevant for our approximate computation, we simply use $\hat{r}$ to denote the favorite type itself henceforth. The favorite type- which represents the empirical probability distribution of the first codeword in the random codebook that matches the data- is defined and studied in [34]. It is a fascinating object, since iteratively finding favorite types results in identifying the optimal Q^* because of Blahut-Arimoto-type convexity considerations. It is thus satisfying to see the favorite type make an appearance in these computations. Note that $\hat{r}$ depends on q, since it is the minimizer of a functional in which q is a parameter.

Now,

$$\begin{aligned} -\log q^n(B(X_1^n, D)) &\approx nf(\hat{r}) - (K-1)\log n \\ &= nD(\hat{r}\|q) + nI_\ell(\hat{P}_{X_1^n}, \hat{r}, D) + \left(1 - \frac{K}{2}\right)\log n + O(1) \end{aligned} \tag{147}$$

But from [34],

$$D(\hat{r}\|q) + I_\ell(p, \hat{r}, D) = R(p, q, D) \tag{148}$$

so that (147) can be rewritten as

$$-\log q^n(B(X_1^n, D)) \approx R(\hat{P}_{X_1^n}, q, D) + \left(1 - \frac{K}{2}\right)\log n + O(1) \tag{149}$$

which would be precisely the same as Proposition 1 if the coefficient of the $\log n$ term were $\frac{1}{2}$ instead of $(1 - \frac{K}{2})$! Note that the error arose because of the rough approximation made in (146). If that calculation were done very carefully, one ought to get an exponent of $\frac{K-1}{2}$ there and consequently recover Proposition 1 exactly here.

To obtain the full second-order generalized AEP using the method of types, one needs to carry out a Taylor expansion of $I_\ell(\hat{P}_{X_1^n}, \hat{r}, D)$ in its first variable about p. (This can be performed since Lemma 2 in [35] indicates that I_ℓ is at least second-order differentiable.) Then a simple observation based on (148) yields the second-order generalized AEP.

C Remarks on Asymptotic Normality

In the usual setting of parametric estimation, a large variety of estimators are asymptotically normally distributed about the true value of the parameter. Indeed, under regularity conditions, the maximum likelihood estimator (MLE) is not only asymptotically normal but efficient, in the sense that the limiting variance is given by the inverse Fisher information and is optimal according to the Cramér-Rao bound. A natural, interesting question in our framework for lossy compression is whether the LML estimator has any such properties, and

what they mean. In this Appendix, the classical approach to proving asymptotic normality of the MLE, (see, e.g., [19]) is examined in the context of lossy estimators; this indicates the kind of problems that crop up in extending the analogy.

The classical method is based on Taylor's theorem assuming smoothness of the likelihood function. Suppose $l(\theta) = l(\theta|X_1^n)$ denotes either the lossy likelihood function or the pseudo-lossy likelihood function. Let θ_n denote the maximizer of $l(\theta)$– this is $\hat{\theta}_n^{\mathrm{LML}}$ when l is the lossy likelihood and $\tilde{\theta}_n^{\mathrm{LML}}$ when l is the pseudo-lossy likelihood.

Expanding $\nabla_\theta l(\theta_n)$ in a Taylor series about the optimal parameter θ^* yields

$$\nabla_\theta l(\theta_n) = \nabla_\theta l(\theta^*) + \mathrm{Hess}_\theta l(\theta^*) \cdot (\theta_n - \theta^*) + \tfrac{1}{2}(\theta_n - \theta^*)^T \cdot \frac{\partial^3 l(\bar{\theta_n})}{\partial \theta^3} \cdot (\theta_n - \theta^*), \tag{150}$$

where $\bar{\theta_n}$ lies between θ_n and θ^*. Assuming that the optimal θ^* is in the interior of Θ, the derivative of $l(\theta)$ at the point where it achieves its maximum (namely, θ_n) must be 0. Thus we can write:

$$\sqrt{n}(\theta_n - \theta^*) = \left[-\frac{1}{n}\mathrm{Hess}_\theta l(\theta^*) - \frac{1}{2n}(\theta_n - \theta^*)^T \cdot \frac{\partial^3 l(\tilde{\theta_n})}{\partial \theta^3} \right]^{-1} \cdot \left[\frac{1}{\sqrt{n}} \nabla_\theta l(\theta^*) \right]. \tag{151}$$

Guided by the proof for the lossless MLE, we would hope to prove that each of these terms converges in an appropriate sense, by using the second-order properties of $l(\theta)$, and the consistency results for the lossy estimators. If $Q^* = Q_{\theta^*}$ is in the interior of the simplex, and if (1) can be validly differentiated term-by-term with respect to θ, with *the error terms remaining asymptotically insignificant*,

$$\begin{aligned} \frac{1}{\sqrt{n}} \nabla_\theta l(\theta^*) &= \sqrt{n} \nabla_\theta \Big|_{\theta^*} \left\{ -\frac{1}{n} \log Q_\theta^n(B(X_1^n, D)) \right\} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_\theta \ g_{\theta^*}(X_i) + \frac{\log n}{2\sqrt{n}} + O\left(\frac{\log \log n}{\sqrt{n}} \right) \end{aligned} \tag{152}$$

The first term on the right is a normalized sum of i.i.d. random vectors, which have mean 0 since the finiteness of A justifies the interchange of derivative and expectation. It therefore converges weakly to a multivariate normal with covariance matrix $J_1(\theta)$ given by the covariance matrix of $\nabla_\theta|_{\theta^*} g_\theta(X)$:

$$J_1(\theta) = \mathrm{Cov}_P \left[\frac{\partial}{\partial \theta_i} g_\theta(X), \frac{\partial}{\partial \theta_j} g_\theta(X) \right]. \tag{153}$$

The other terms are lower order and converge to 0 w.p.1, so that

$$\frac{1}{\sqrt{n}} \nabla_\theta l(\theta^*) \to N\big(0, J_1(\theta)\big). \tag{154}$$

(The above is valid when $l(\theta)$ is the lossy likelihood, but the result is unchanged when it is the pseudo-lossy likelihood, since only the less significant terms that converge to 0 w.p.1 are different.)

Formally differentiating the second-order lossy AEP again,

$$\begin{aligned} \frac{1}{n}\mathrm{Hess}_\theta l(\theta^*) &= \mathrm{Hess}_\theta \Big|_{\theta^*} \left[R(P, Q_\theta, D) + \frac{1}{n} \sum_{i=1}^n g_\theta(X_i) + o(1) \right] \\ &\to \mathrm{Hess}_\theta \ R(P, Q_{\theta^*}, D) \\ &=: J_2(\theta^*) \end{aligned} \tag{155}$$

where the convergence is w.p.1. The boundedness in probability of the norm of the third derivative tensor term in (151) would follow similarly.

Thus, under the assumption that the expansion (1) can be differentiated thrice term-by-term in such a way that the error term remains $o(\frac{\log n}{n})$, the LML estimator is asymptotically normal around the optimal θ^*, and furthermore the covariance matrix of the limiting normal is $J_2^{-1} J_1 (J_2^{-1})^T$ (with a similar result for the pseudo-LML estimator). However, this assumption is a major problem, since we know very little about the error term. Although the left-hand side of (126) in Appendix A is an explicit expression for the error term, it is opaque and it is unclear how to differentiate it. All that is explicitly known about the error term is its order, which does not say anything about the order of its derivatives. To see this, consider the sequence of functions $f_n(x) = \sin nx \frac{\log \log n}{n}$ which is $O(\frac{\log \log n}{n})$ (and hence converges to 0 quite rapidly) but whose derivatives are unbounded as n increases.

References

[1] D. W. K. Andrews. Consistency in nonlinear econometric models: A generic uniform law of large number. *Econometrica*, 55(6):1465–1471, 1987.

[2] Z. D. Bai, C. R. Rao, and Y. Wu. Model selection with data-oriented penalty. *J. Statist. Plann. Inference*, 77(1):103–117, 1999.

[3] A. Barron. *Logically Smooth Density Estimation*. PhD thesis, Dept. of Electrical Engineering, Stanford University, 1985.

[4] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. (Information theory: 1948–1998). *IEEE Trans. Inform. Theory*, 44(6):2743–2760, 1998.

[5] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, 37(4):1034–1054, 1991.

[6] J. A. Berning Jr. On the multivariate law of the iterated logarithm. *Ann. Prob.*, 7:980–988, 1979.

[7] T. Cover and J. Thomas. *Elements of Information Theory*. J. Wiley, New York, 1991.

[8] A. Dembo and I. Kontoyiannis. The asymptotics of waiting times between stationary processes, allowing distortion. *Ann. Appl. Probab.*, 9:413–429, 1999.

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

[10] M. Harrison. The convergence of lossy maximum likelihood estimators. *Brown University APPTS Report # 03-5*, April 2003.

[11] M. Harrison. The first-order asymptotics of waiting times between stationary processes under non-standard conditions. *Brown University APPTS Report # 03-3*, April 2003.

[12] M. Harrison. A lossy generalized Asymptotic Equipartition Property under nonstandard conditions. *Preprint*, 2005.

[13] M. Harrison, I. Kontoyiannis, and M. Madiman. The Minimum Description Length principle in lossy data compression. *Preprint*, 2004.

[14] J. Kieffer. Sample converses in source coding theory. *IEEE Trans. Inform. Theory*, 37(2):263–268, 1991.

[15] I. Kontoyiannis. An implementable lossy version of the Lempel-Ziv algorithm – Part I: Optimality for memoryless sources. *IEEE Trans. Inform. Theory*, 45(7):2293–2305, November 1999.

[16] I. Kontoyiannis. Pointwise redundancy in lossy data compression and universal lossy data compression. *IEEE Trans. Inform. Theory*, 46(1):136–152, January 2000.

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

[18] L. G. Kraft. A device for quantizing, grouping and coding amplitude modulated pulses. Master's thesis, Department of Electrical Engineering, MIT, Cambridge MA, 1949.

[19] E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1998.

[20] T. Luczak and W. Szpankowski. A suboptimal lossy data compression algorithm based on approximate pattern matching. *IEEE Trans. Inform. Theory*, 43(5):1439–1451, 1997.

[21] G. Matz and P. Duhamel. Information geometric formulation and interpretation of accelerated Blahut-Arimoto-type algorithms. In *Proc. IEEE Inform. Th. Workshop 2004, San Antonio*, October 2004.

[22] B. McMillan. Two inequalities implied by unique decipherability. *IEEE Trans. Inform. Th.*, 2:115–116, 1956.

[23] B. M. Pötscher and I. R. Prucha. A uniform law of large numbers for dependent and heterogeneous data processes. *Econometrica*, 57(3):675–683, 1989.

[24] J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

[25] J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Inform. Theory*, 42(1):40–47, 1996.

[26] C. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656, 1948.

[27] C. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec.*, part 4:142–163, 1959. Reprinted in D. Slepian (ed.), *Key Papers in the Development of Information Theory*, IEEE Press, 1974.

[28] Y. Steinberg and M. Gutman. An algorithm for source coding subject to a fidelity criterion, based upon string matching. *IEEE Trans. Inform. Theory*, 39(3):877–886, 1993.

[29] V. N. Vapnik and A. Y. Chervonenkis. Necessary and sufficient conditions for the uniform convergence of empirical means to their true values. *Teor. Veroyatnost. i Primenen.*, 26(3):543–563, 1981.

[30] E.-H. Yang and J. Kieffer. On the performance of data compression algorithms based upon string matching. *IEEE Trans. Inform. Theory*, 44(1):47–65, 1998.

[31] E.-H. Yang and Z. Zhang. On the redundancy of lossy source coding with abstract alphabets. *IEEE Trans. Inform. Theory*, 45(4):1092–1110, 1999.

[32] E.-H. Yang and Z. Zhang. The redundancy of source coding with a fidelity criterion – Part II: Coding at a fixed rate level with unknown statistics. *IEEE Trans. Inform. Theory*, 47(1):126–145, 2001.

[33] E.-H. Yang and Z. Zhang. The redundancy of source coding with a fidelity criterion – Part III: Coding at a fixed distortion level with unknown statistics. *Preprint*, 2002.

[34] R. Zamir and K. Rose. Natural type selection in adaptive lossy compression. *IEEE Trans. Inform. Theory*, 47(1):99–111, 2001.

[35] Z. Zhang, E.-H. Yang, and V. Wei. The redundancy of source coding with a fidelity criterion – Part I: Known statistics. *IEEE Trans. Inform. Theory*, 43(1):71–91, 1997.

[36] G. Zhou and Z. Zhang. On the redundancy of trellis lossy source coding. *IEEE Trans. Inform. Theory*, 48(1):205–218, 2002.