Maximum Likelihood Estimation for Lossy Data Compression*

Matthew Harrison
Division of Applied Mathematics
Brown University
Providence, RI 02912 USA
Matthew_Harrison@Brown.EDU

Ioannis Kontoyiannis
Division of Applied Mathematics
& Department of Computer Science
Brown University
Providence, RI 02912 USA
yiannis@dam.brown.edu

Abstract

In *lossless* data compression, given a sequence of observations $(X_n)_{n\geq 1}$ and a family of probability distributions $\{Q_\theta\}_{\theta\in\Theta}$, the estimators $(\tilde{\theta}_n)_{n\geq 1}$ obtained by minimizing the ideal Shannon code-lengths over the family $\{Q_\theta\}_{\theta\in\Theta}$,

$$\tilde{\theta}_n := \arg\min_{\theta\in\Theta}\Big[-\log Q_\theta(X_1^n)\Big],$$

where $X_1^n := (X_1, X_2, \ldots, X_n)$, coincide with the classical maximum-likelihood estimators (MLEs). In the corresponding *lossy* compression setting, the ideal Shannon code-lengths are approximately $-\log Q_\theta(B(X_1^n, D))$ bits, where $B(X_1^n, D)$ is the distortion-ball of radius D around the source sequence X_1^n. In this work we consider the analogous estimators obtained by minimizing these lossy code-lengths,

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

The $\hat{\theta}_n$ are a lossy version of the MLEs, which we call “lossy MLEs”. We investigate the strong consistency of lossy MLEs when the Q_θ are i.i.d. and the sequence $(X_n)_{n\geq 1}$ is stationary and ergodic.

1 Introduction

Maximum likelihood (ML) estimation is an important tool in statistics and it is a useful intuitive and analytic starting point for more recent statistical developments like minimum description length (MDL) estimation and maximum a posteriori (MAP) Bayesian estimation. Many of the connections between statistics and information theory are naturally phrased in terms of ML. We will focus on the connections that lead to the celebrated MDL principle.

The present work is a first step in developing these connections in the context of *lossy* source coding. As will be described in the extensive treatment given in [6], the next step is the formulation of a lossy version of the MDL principle. Here we formalize

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

this connection by introducing the lossy maximum likelihood estimators (or lossy MLEs) and examining their strong consistency. Besides being a useful tool for proofs of finer properties, consistency is often a valuable heuristic, and, as discussed in [6], it can lead to heuristics for improving the performance of vector-quantizer design algorithms.

The starting point for our discussion is the Kraft inequality: Any uniquely decodable code with (finite binary string) code-lengths $\{l(\vec{x})\}$ satisfies $\sum_{\vec{x}} 2^{-l(\vec{x})} \leq 1$, and for any collection of $\{l(\vec{x})\}$ satisfying this inequality there is a uniquely decodable code with these code-lengths [4]. This leads immediately to the familiar codes-measures correspondence

$$l(\vec{x}) \approx -\log Q(\vec{x}) \quad \text{bits}, \tag{1.1}$$

namely, any probability measure Q leads to a code with code-lengths given by (1.1) and, conversely, and code induces a probability measure on the space of observations. [Throughout the paper, $\log := \log_2$.]

The main goal in lossless source coding it to design codes that minimize the code-lengths, that is, minimize the left side of (1.1). In statistics, particularly in ML estimation, the goal is to choose models that maximize the likelihood $Q(\vec{x})$, or, equivalently, minimize the right side of (1.1). This connection has led to useful insights in the literature, particularly in the context of the MDL principle; see [1] for a survey.

Turning to lossy source coding, this line of reasoning can be reproduced almost word-for-word for (variable-rate, fixed-distortion) lossy compression [8][9]. The starting point here is a lossy version of the Kraft inequality that leads to a lossy codes-measures correspondence. As it turns out,

$$l(\vec{x}) \approx -\log Q(B(\vec{x}, D)) \quad \text{bits}, \tag{1.2}$$

where $B(\vec{x}, D)$ denotes the distortion-ball of all reproduction strings that are within distortion D or less of the source string $\vec{x}$. Roughly speaking, (1.2) has the same interpretation as in the lossless case: To every measure Q there corresponds a code that compresses data with distortion D or less, and the performance of any such code can be quantified as in (1.2) in terms of an appropriately chosen Q; the precise statement is given in the next section.

The goal then is to choose codes that minimize the left side of this expression, or, by analogy to the lossless case reasoning, to find probability measures Q that make the right side as small as possible. In particular, we interpret $Q(B(\vec{x}, D))$ as a "lossy likelihood" and create lossy MLEs by maximizing this likelihood. These lossy MLEs are our main object of study. Specifically, we will ask and answer (in the affirmative) the following two questions: Do these estimators converge to a probability measure? Do the lossy code-lengths induced by this limiting measure have desirable compression properties?

2 The Lossy Codes-Measures Correspondence

Let S be the (general) source alphabet, where we assume that $(S, \mathcal{S})$ is an arbitrary measurable space; similarly $(T, \mathcal{T})$ is taken to be a measurable space corresponding to the reproduction alphabet T. The distortion between source and reproduction strings will be measured in terms of the single-letter distortion criterion induced by a (measurable) function $\rho : S \times T \to [0, \infty)$. Given a distortion level $D \geq 0$ and a source sequence $x_1^n := (x_1, \ldots, x_n) \in S^n$, we define the distortion-ball

$$B(x_1^n, D) := \left\{ y_1^n \in T^n : \frac{1}{n} \sum_{k=1}^{n} \rho(x_k, y_k) \leq D \right\}$$

to be the set of all reproduction sequences that are within distortion D from x_1^n.

For our purposes, a (fixed-distortion, variable-rate) lossy compression code consists of a map that takes source strings x_1^n to reproduction strings y_1^n, followed by a uniquely decodable lossless code that maps each y_1^n to a finite-length binary string. When each x_1^n is mapped to a $y_1^n \in B(x_1^n, D)$, we say that the code operates at distortion level D. The figure of merit is the code-length $l(x_1^n)$ bits assigned to x_1^n, where $l(x_1^n)$ is simply the lossless code-length for the y_1^n that was chosen to represent x_1^n.

In the remainder of the paper, we restrict ourselves to situations where it is *possible* to do fixed-distortion lossy compression, that is, to situations where there exists a lossy code such that the code-lengths $l(x_1^n)$ are finite. For example, if $S = T = \mathbb{R}$, $\rho(x, y) = I_{x \neq y}$ is Hamming distance, and $D = 0$, then we cannot do lossy compression. All of this can be made precise as in [9], but that is not necessary here. We can skip ahead to a result of [9, Theorem 1, Theorem 2]:

Theorem 2.1. (Lossy Kraft Inequality) [9] *For any code operating at distortion level D and having code-lengths $\{l(x_1^n)\}$, there is a probability measure Q such that*

$$l(x_1^n) \geq -\log Q(B(x_1^n, D)) \quad \textit{bits},$$

for all $x_1^n \in S^n$. Somewhat conversely, if $(Q_n)_{n\geq 1}$ is any admissible[1] *sequence of probability measures on T^n, then there is a sequence of codes with code-lengths $\{l_n(x_1^n)\}$ such that*

$$l_n(X_1^n) \overset{a.s.}{\leq} -\log Q_n(B(X_1^n, D)) + O(\log n), \tag{2.1}$$

as $n \to \infty$. Under additional conditions, finer bounds can be obtained.

In (2.1) the nth code is constructed based on a random codebook generated according to the measure Q_n. So it is natural to talk about lossy codebooks and codes induced by a probability measure Q. This remark and Theorem 2.1 motivate the heuristic,

$$l_n(x_1^n) \approx -\log Q_n(B(x_1^n, D)) \quad \text{bits},$$

which we refer to as the lossy codes-measures correspondence. Note that, like in the lossless case, in Theorem 2.1 *no* assumptions are made on the source (X_n). But the lossy correspondence is asymptotic, whereas the lossless correspondence is not. This turns out not to matter much here, because we are focusing on asymptotic results. In fact, in one aspect the lossy codes-measures correspondence is an improvement over the lossless version: For continuous alphabets the lossless correspondence breaks down. Although maximum likelihood estimation seems to make sense by maximizing over densities evaluated at the observations, there is no immediate way in which that value relates to code-lengths (unless one resorts to asymptotic quantization arguments). On the other hand, the lossy correspondence does not change as we move from discrete to continuous alphabets. Integrating over distortion-balls effectively changes a continuous alphabet into a discrete one and $-\log Q(B(x_1^n, D))$ still makes sense as an idealized code-length.

[1] A sequence of probability measures (Q_n) is "admissible" [9] if it satisfies the natural requirement that it yields codes with finite code-lengths, namely, if there is a constant R such that $\limsup_{n\to\infty} -n^{-1} \log Q_n(B(X_1^n, D)) \leq R < \infty$ a.s. This constraint turns out to be unimportant for the theory of lossy maximum likelihood estimation.

3 Convergence of Lossy Code Lengths

Motivated by the lossy codes-measures correspondence, especially (2.1), we can try to construct sequences of probability measures $(Q_n)_{n\geq 1}$ so that

$$\limsup_{n\to\infty} -\frac{1}{n}\log Q_n(B(X_1^n, D)) \overset{\text{a.s.}}{\leq} R.$$

This would guarantee the existence of a sequence of codes with asymptotic rate no greater than R bits per symbol. In particular, when the $(Q_n)_{n\geq 1}$ are the successive marginals of a (sufficiently rapidly) mixing process, this turns out to be possible [5][3]. For simplicity, we restrict ourselves to the i.i.d. case.

Theorem 3.1. (Generalized AEP) *Suppose $(X_n)_{n\geq 1}$ is stationary and ergodic with $X := X_1 \sim P$ and the measures $(Q_n)_{n\geq 1}$ are i.i.d. in the sense that $Q_n := Q^n$. Then,*

$$\lim_{n\to\infty} -\frac{1}{n}\log Q^n(B(X_1^n, D)) \overset{a.s.}{=} R(P, Q, D), \tag{3.1}$$

for some appropriate rate function $R(P, Q, D)$, except perhaps for a single value of D denoted D_{min}.[2]

To make the connection back to the lossless case, we note that (for discrete alphabets)

$$\lim_{n\to\infty} -\frac{1}{n}\log Q^n(X_1^n) \overset{\text{a.s.}}{=} H(P) + H(P||Q), \tag{3.2}$$

where $H(\cdot)$ is the entropy. This is one of the versions of the Asymptotic Equipartition Property (or AEP) and in many ways (3.2) plays the same role in lossless compression that (3.1) plays in lossy compression. To emphasize this analogy, (3.1) is sometimes referred to as the generalized AEP. Under the present conditions, Theorem 3.1 is a new result and its proof is based on large deviations techniques along the lines of corresponding arguments in [5]. Several other versions have also appeared in the literature; an extensive bibliography can be found in [5].

The rate-function $R(P, Q, D)$ also depends on the distortion function ρ, but this dependence is usually suppressed. There are several characterizations of $R(P, Q, D)$ other than (3.1) that can be found in [5]. Some are information-theoretic and others come from large deviations arguments. The following expression from [11] looks similar to the Shannon rate distortion function,

$$R(P, Q, D) = \inf_{(X,\hat{Y})} \left[I(X; \hat{Y}) + H(\hat{Q}||Q) \right],$$

where the infimum is over all jointly distributed random variables $(X, \hat{Y})$, such that $X \sim P$, $E\rho(X, \hat{Y}) \leq D$, and $\hat{Y} \sim \hat{Q}$. $I(\cdot;\cdot)$ denotes mutual information and $H(\cdot||\cdot)$ denotes relative entropy. In this paper we do not make use of any representation of $R(P, Q, D)$ other than (3.1).

[2] $R(P, Q, D)$ is nonincreasing in D, and the point $D = D_{\text{min}}$ is the transition point where $R(P, Q, D)$ becomes finite valued. Often, $D_{\text{min}} = 0$. In this case (and many others), (3.1) holds even when $D = D_{\text{min}}$.

4 Consistency of Lossy MLEs

From now on we assume that $(X_n)_{n\geq 1}$ is a stationary and ergodic source with first marginal given by P, i.e., $X := X_1 \sim P$, and, as in the previous section, we restrict attention to i.i.d. measures $Q_n = Q^n$: We start off with a given family Θ of probability measures Θ on T, and consider the corresponding i.i.d. mearures Q^n induced by $Q \in \Theta$.

Given such a family Θ, suppose that it is possible to find $Q^* \in \Theta$ such that

$$R(P, Q^*, D) = R(P, \Theta, D) := \inf_{Q\in\Theta} R(P, Q, D).$$

Then the generalized AEP in (3.1) and the lossy codes-measures would imply that there is a sequence of codes operating at distortion level D, with limiting rate no greater than $R(P, \Theta, D)$ bits/symbol. Namely, these codes would achieve the best rate possible within the class Θ.

This is especially interesting in cases where $R(P, \Theta, D) = R(D)$, the rate-distortion function of the source. Since no sequence of lossy codes can asymptotically beat $R(D)$, the sequence of codes induced by Q^* must be asymptotically optimal. For example, if $(X_n)_{n\geq 1}$ is i.i.d. and Θ is the set of all probability measures on T, then $R(P, \Theta, D) = R(D)$. In the special case when T is finite (as well as many other interesting cases), a minimizer Q^* can be shown to exist.

But even when $R(P, \Theta, D) > R(D)$, the codes induced by a minimizer Q^* are asymptotically optimal over a large class of codes, namely, those codes that look like they were induced by some $Q \in \Theta$. Moreover, identifying Q^* might be interesting from both a theoretical and a practical point of view, as discussed further in [6].

The question then becomes how to find such a Q^* when P is unknown. There are conceptually many different ways of accomplishing this. Motivated by the success of maximum likelihood (ML) estimation in the lossless case, we have chosen to estimate Q^* by maximizing the lossy likelihood $Q^n(B(X_1^n, D))$ or, equivalently, minimizing $-\log Q^n(B(X_1^n, D))$. [A somewhat related scheme is to approximate Q^* by a probability measure that minimizes $E\left[-\log Q^n(B(X_1^n, D))\right]$. These approximations are well-behaved [7], but they are not universal in the sense that the distribution of the source must be known *a priori*.]

Formally, let $\hat{Q}_n : S^n \to \Theta$, $n \geq 1$, be a sequence of estimators taking values in a collection Θ of probability measures on T. We say that the sequence $(\hat{Q}_n)_{n\geq 1}$ is *a sequence of approximate lossy MLEs* if,

$$-\frac{1}{n}\log \hat{Q}_n^n(B(X_1^n, D)) \overset{\text{a.s.}}{\leq} \inf_{Q\in\Theta} -\frac{1}{n}\log Q^n(B(X_1^n, D)) + \epsilon_n,$$

for some sequence $(\epsilon_n)_{n\geq 1}$ with $\epsilon_n \to 0$ as $n \to \infty$. Note that we have suppressed the dependence of $\hat{Q}_n$ on X_1^n.

We are interested in the asymptotic behavior of $\hat{Q}_n$ within the set Θ. In particular, we want to make statements of the form:

$$\hat{Q}_n \to Q^*.$$

For simplicity, we will only consider a single situation here.

Let $T \subset \mathbb{R}^d$ be compact and let Θ be the set of all probability measures on T. Suppose $\rho(x, \cdot)$ is continuous for each $x \in S$ and suppose that $\inf_{y\in T} \rho(x, y) \leq D$ for all x. Then:

Theorem 4.1. (Strong consistency of lossy MLEs) *The set of minimizers*

$$\mathcal{Q}^* := \{\textit{minimizers } Q \in \Theta \textit{ of } R(P, Q, D)\}$$

is not empty, and every sequence of approximate lossy MLEs $(\hat{Q}_n)_{n \geq 1}$ *a.s. satisfies*

$$\hat{Q}_n \xrightarrow{w} \mathcal{Q}^*.$$

The $\xrightarrow{w}$ convergence above means that the Q_n converge weakly to the set of minimizers $\mathcal{Q}^*$.[3] If there is a unique minimizer Q^*, then of course the $\hat{Q}_n$ converge weakly to Q^* a.s. Note that this particular setup includes the important special case when T is a finite alphabet.

To make the connection with classical maximum likelihood estimation, suppose $S = T$ is finite, $\rho(x, y) := I_{x \neq y}$ is Hamming distance and $D = 0$. Then $Q^n(B(x_1^n, D)) = Q^n(x_1^n)$, so lossy MLEs are just regular MLEs and $R(P, Q, D) = H(P) + H(P||Q)$. Since the relative entropy has a unique minimizer at $Q^* = P$, Theorem 4.1 reduces to the classical result that MLEs are strongly consistent.

4.1 Remarks about Theorem 4.1

Extensions. The assumptions used for Theorem 4.1 can be significantly relaxed in several directions, as described in [6]. First – and most importantly – it is possible to consider classes Θ of distributions with memory. The general scheme is to start with the memoryless case, and generalize to the case Q_n has memory using blocking arguments. Second, in the case of parametric estimation, we can often allow T to not be compact. For example, if Θ is a location-scale family on $T = \mathbb{R}^m$, Theorem 4.1 remains valid even though T is not compact. Finally, the assumption $\inf_{y \in T} \rho(x, y) \leq D$ can also be relaxed.

Taking a slightly different point of view, it is possible to rephrase the result entirely in terms of total variation convergence instead of weak convergence. In that setting ρ need not be continuous, but then it is difficult to consider the full nonparametric case (unless T is finite).

Proof. Here we briefly discuss some of the main steps in the proof of Theorem 4.1; see [6] for complete arguments. In the lossy case, the proof of Theorem 4.1 is considerably more difficult than the classical proof of the strong consistency of (lossless) MLEs.

The first step in the proof is to derive an appropriate version of the generalized AEP, as in Theorem 3.1. Because we are estimating over a large parameter space, it may be necessary to entertain measures Q that do not satisfy the typical assumptions found in the literature for (3.1). Furthermore, because of the problems at the point $D = D_{\min}$ (recall that $D_{\min}$ depends on Q), (3.1) may not hold for some Q is the parameter space. We must understand how these pathological cases behave. This uses results about the recurrence properties of random walks with stationary increments.

Once (3.1) and its pathologies have been understood, the next step is to prove a statement just like (3.1), except that a.s.-limits are replaced by a.s.-"epi-limits." Epi-convergence is a type of functional convergence that naturally characterizes the local

[3] To be pedantic, this means that the infimum of the Prohorov distances between Q_n and each $Q^* \in \mathcal{Q}^*$ converges to zero; see [2] for more details.

conditions needed for the consistency of minimizers [10]. This uses standard measure-theoretic techniques and requires taking careful note of the pathologies in (3.1), including the exceptional sets for a.s. convergence.

Penalization. The above two steps are sufficient for the proof of Theorem 4.1 under the assumptions presented here because Θ is compact (since T is compact). More generally, it is necessary to verify a global condition to complement the local condition provided by epi-convergence. This global condition is typically checked on a case-by-case basis, and, much as in the classical ML literature, one quickly finds that some important examples do not satisfy the necessary global properties. In fact, in many of these cases it is not just the conditions, it is the actually result that fails. Such cases require more advanced techniques, and often it is necessary to modify the MLEs appropriately, for example by including penalization terms for "overly complex" models Q. This line of thought parallels some of the analogous developments in the lossless case, and it is developed in more detail in [6]. A small taste of the corresponding results is given in the next section.

4.2 Consistency of Lossy MDLEs

Let $(\hat{Q}_n)_{n\geq 1}$ be any sequence of approximate lossy MLEs. Theorem 3.1 implies that

$$\limsup_{n\to\infty} -\frac{1}{n}\log \hat{Q}_n^n(B(X_1^n, D)) \overset{\text{a.s.}}{\leq} R(P,\Theta,D) \quad \text{bits/symbol.}$$

Naively applying the codes-measures correspondence of Theorem 2.1 to the sequence $(\hat{Q}_n)_{n\geq 1}$ suggests that we could use the sequence of lossy MLEs to build codes that achieve the optimal rate $R(P,\Theta,D)$ among all codes induced by measures $Q \in \Theta$. Unfortunately, this reasoning is faulty.

Theorem 2.1 applies to a fixed sequence of measures that is already known to both the encoder and the decoder. The sequence of lossy MLEs, however, depends on the specific realization of the source $(X_n)_{n\geq 1}$, which is unknown to the decoder. At this point, there are two ways to proceed. The first is an offline procedure. We observe the source, estimate the lossy MLE and then build the encoder and decoder based on this information and use them to encode a different source realization. If $R(P,Q,D)$ is continuous in Q, as it typically will be for simple examples like finite T, then Theorem 4.1 implies that we will have a good compression algorithm. For examples where $R(P,Q,D)$ is not continuous in Q, then our lossy MLE may not even induce a good codebook. This is analogous to the classic examples of overfitting when using MLEs.

Alternatively, we can overcome the problem by using a two-step code: First describe an appropriately chosen Q to the decoder, and then describe the data using the code induced by this Q. Following along the steps of the corresponding lossless coding developments, we consider a class of two-stage lossy codes and a corresponding family of lossy Minimum Description Length (MDL) estimates (MDLEs).

The formal setup for (approximate) lossy MDLEs is the same as for lossy MLEs, except that we require our estimators to be (approximate) minimizers (in Q over Θ) of

$$-\log Q(B(X_1^n, D)) \; + \; L(Q),$$

where $L(Q)$ is the code-length for a description of the measure Q according to a uniquely decodable code on Θ. This requirement of course forces $L(\cdot)$ to only be finite on countably many $Q \in \Theta$. Under appropriate regularity conditions on $L(\cdot)$, the techniques used to prove Theorem 4.1 can be used to prove the strong consistency of lossy MDLEs under similar assumptions:

Theorem 4.2. (Strong consistency of lossy MDLEs) *The set of minimizers*

$$\mathcal{Q}^* := \{\textit{minimizers } Q \in \Theta \textit{ of } R(P,Q,D)\}$$

is not empty, and every sequence of approximate lossy MDLEs $(\tilde{Q}_n)_{n\geq 1}$ *a.s. satisfies*

$$\tilde{Q}_n \xrightarrow{w} \mathcal{Q}^*.$$

Moreover,

$$\limsup_{n\to\infty} \frac{1}{n}\left[-\log \tilde{Q}_n^n(B(X_1^n,D)) + L(\tilde{Q}_n)\right] \overset{a.s.}{\leq} R(P,\Theta,D).$$

Therefore, for any stationary and ergodic source, lossy MDLEs provide a two-stage lossy compression algorithm which is asymptotically optimal over all codes induced by measures $Q \in \Theta$ (as well as all codes whose performance can be bounded below by such Q as in Theorem 2.1).

Although we do not explore lossy MDLEs any further here, note that, in the lossless case, MDLEs are consistent in many examples where MLEs are not, and there is extensive evidence that MDLEs can discover structures present in the data in finite time, whereas MLEs cannot. In [6] we consider these issues further in the lossy context.

5 Conclusions

We have extended a well-known connection between lossless compression and maximum likelihood estimation to lossy compression. After identifying a lossy codes-measures correspondence (Theorem 2.1) and a lossy version of the AEP (Theorem 3.1), introduced the lossy maximum likelihood estimators (MLEs) and showed that they converge to the set of reproduction probability distributions which minimize the asymptotic rate given by the lossy AEP (Theorem 4.1). This is the main result of the paper. Finally, we indicated that the theory extends (as in the lossless case) to lossy minimum description length estimators (MDLEs). In the lossless case this sequence of developments has had great practical and theoretical implications. It remains to be seen if corresponding results will arise from these developments in the lossy case.

References

[1] A. Barron, J. Rissanen and B. Yu. “The minimum description length principle in coding and modeling.” *IEEE Trans. Inform. Theory*, **44**, pp. 2743–2760, October 1998.

[2] P. Billingsley. *Convergence of Probability Measures*, 2nd Ed., New York: Wiley, 1999.

[3] Z. Chi. “The first order asymptotics of waiting times with distortion between stationary processes.” *IEEE Trans. Inform. Theory*, **47**, pp. 338–347, January 2001.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.

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

[6] M. Harrison, I. Kontoyiannis and M. Madiman. “The minimum description length principle in lossy data compression.” In preparation.

[7] J.C. Kieffer. “Sample converses in source coding theory.” *IEEE Trans. Inform. Theory*, **37**, pp. 263–268, March 1991.

[8] I. Kontoyiannis. “Model selection via rate-distortion theory.” *34th Annual Conference on Information Sciences and Systems*, March 2000.

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

[10] G. Salinetti. “Consistency of statistical estimators: the epigraphical view.” pp. 365–383. In *Stochastic Optimization: Algorithms and Applications.* S. Uryasev and P. M. Pardalos, Eds., Dordrecht: Kluwer Academic Publishers, 2001.

[11] E.-h. Yang and J.C. Kieffer. “On the performance of data compression algorithms based upon string matching.” *IEEE Trans. Inform. Theory*, **44**, pp. 47–65, January 1998.