The Convergence of Lossy Maximum Likelihood Estimators

Matthew Harrison

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

July 30, 2003

Abstract

Given a sequence of observations $(X_n)_{n\geq 1}$ and a family of probability distributions $\{Q_\theta\}_{\theta\in\Theta}$, the lossy likelihood of a particular distribution Q_θ given the data $X_1^n := (X_1, X_2, \ldots, X_n)$ is defined as

$$Q_\theta(B(X_1^n, D)),$$

where $B(X_1^n, D)$ is the distortion-ball of radius D around the source sequence X_1^n. Here we investigate the convergence of maximizers of the lossy likelihood.

1 Introduction

Consider a random data source $(X_n)_{n\geq 1}$ and a collection of probability measures $\{P_\theta\}_{\theta\in\Theta}$ on the sequence space. In statistics, the likelihood of a particular distribution P_θ given the empirical data $X_1^n := (X_1, \ldots, X_n)$ is defined by

$$P_\theta(X_1^n).$$

The maximizer (over Θ) of the likelihood is called a maximum likelihood estimator (MLE). In many situations, the sequence of MLEs (in n) converges to $\theta^* \in \Theta$, where P_{θ^*} is the distribution of the source. An MLE is also a minimizer of

$$-\log P_\theta(X_1^n). \tag{1.1}$$

When written in this form, we notice that the negative log-likelihood is exactly the ideal Shannon code length for the data X_1^n and the source P_θ. So we can conceptualize the MLE as searching for probability measures that would induce short codewords for the data. Indeed P_{θ^*} would give the optimal first order *lossless* compression performance.

Kontoyiannis and Zhang (2002) [15] argue that an analog of (1.1) for fixed distortion *lossy* data compression is

$$-\log Q_\theta(B(X_1^n, D)), \tag{1.2}$$

where $B(X_1^n, D)$ is the distortion ball around X_1^n of radius D and where $\{Q_\theta\}_{\theta\in\Theta}$ are probability measures on the reproduction sequence space (see below for precise definitions). Reversing the analogy, we define the *lossy likelihood* as

$$Q_\theta(B(X_1^n, D))$$

and we are interested in the asymptotic behavior of maximizers of this quantity, or equivalently, of minimizers of (1.2). We call these *lossy maximum likelihood estimators* or lossy MLEs. We can conceptualize the lossy MLE as searching for probability measures that would induce short codewords for the data allowing for distortion. Here we give conditions under which a sequence of lossy MLEs converges to a limit (or a limiting set). This limiting probability distribution will be optimal in that it induces the shortest codewords among all the distributions that are under consideration.

The connection between statistics and lossless data compression resulting from the dual interpretations of (1.1) has led to many interesting insights and applications. Perhaps similar connections exist for lossy data compression. See Harrison and Kontoyiannis (2002) [11] (where some of these results were reported without proof) and Kontoyiannis (2000) [14] for a more detailed discussion of the motivations and possible applications.

We always assume that the source sequence $(X_n)_{n\geq 1}$ is stationary and ergodic and that the reproduction measures Q_θ satisfy certain strong mixing conditions. We only consider the case of single letter distortion (see the definition of $B(x_1^n, D)$ below), but we allow for arbitrary alphabets and arbitrary distortion functions. Naturally, we also need some assumptions about how the probability distributions $\{Q_\theta\}_{\theta\in\Theta}$ are related to the topology of the parameter space Θ.

2 Epi-convergence

We take the epi-convergence approach for studying the convergence of minimums and minimizers [1, 19], where we think of the lossy MLE as a minimizer of (1.2). Let Θ be a metric space and let $(f_n)_{n\geq 1}$ be a sequence of functions $f_n : \Theta \to [-\infty, \infty]$. We say that f_n epi-converges to a function $f : \Theta \to [-\infty, \infty]$ at the point θ if

$$\liminf_{n\to\infty} f_{m_n}(\theta_n) \geq f(\theta), \text{ for any } \theta_n \to \theta \text{ and any subsequence } m_n \to \infty, \text{ and}$$
$$\limsup_{n\to\infty} f_n(\theta'_n) \leq f(\theta), \text{ for some } \theta'_n \to \theta.$$

If these conditions hold for every $\theta \in \Theta$, then we say that f_n epi-converges to f and we write $f = \text{epi-lim}_n f_n$. In this case, the convergence of minimizers (minima) of f_n to minimizers (minima) of f simplifies to a compactness condition as the following result shows:

Proposition 2.1. [1, 2] Let Θ be a metric space and let $(f_n)_{n\geq 1}$ be a sequence of functions $f_n : \Theta \to [-\infty, \infty]$ such that $f := \text{epi-lim} f_n$ exists on Θ. Then $f : \Theta \to [-\infty, \infty]$ is lower semicontinuous (l.sc.) and

$$\limsup_{n\to\infty} \inf_{\theta\in\Theta} f_n(\theta) \leq \inf_{\theta\in\Theta} f(\theta). \tag{2.1}$$

Let $(\theta_n)_{n\geq 1}$ be a sequence of points from Θ satisfying

$$\limsup_{k\to\infty} f_{n_k}(\theta_{n_k}) \leq \limsup_{k\to\infty} \inf_{\theta\in\Theta} f_{n_k}(\theta), \quad \text{for all subsequences } n_k \to \infty. \tag{2.2}$$

If $(\theta_n)_{n\geq 1}$ is relatively compact, then

$$\theta_n \to \arg\inf_{\Theta} f := \left\{\theta \in \Theta : f(\theta) = \inf_{\theta' \in \Theta} f(\theta')\right\}, \tag{2.3}$$

$$\lim_{n\to\infty} f_n(\theta_n) = \lim_{n\to\infty} \inf_{\theta\in\Theta} f_n(\theta) = \inf_{\theta\in\Theta} f(\theta). \tag{2.4}$$

On the other hand, if $(\theta_n)_{n\geq 1}$ satisfies (2.3) and $\arg\inf_\Theta f$ is compact, then $(\theta_n)_{n\geq 1}$ is relatively compact and (2.4) holds.

In fact, every sequence $(\theta_n)_{n\geq 1}$ satisfying (2.2) is relatively compact if and only if $\arg\inf_\Theta f$ is compact and every sequence $(\theta_n)_{n\geq 1}$ satisfying (2.2) satisfies (2.3).

In Proposition 2.1, we think of the sequence $(\theta_n)_{n\geq 1}$ as a sequence of minimizers of f_n. Indeed, (2.2) is just about the weakest possible notion of a sequence of minimizers. Any sequence of $(\theta_n)_{n\geq 1}$ satisfying

$$f_n(\theta_n) \leq -M_n \vee \inf_{\theta\in\Theta} f_n(\theta) + \epsilon_n,$$

for some sequences $\epsilon_n \to 0$ and $M_n \to \infty$, satisfies (2.2). Such sequences always exist. Under the condition that $f = \text{epi-lim}_n f_n$, every cluster point of a sequence of minimizers of f_n is a minimizer of f. An easy way to ensure cluster points is with a compactness assumption. A subset of a metric space is *relatively compact* if it has compact closure. For a sequence $(\theta_n)_{n\geq 1}$, this is equivalent to saying that every subsequence has a convergent subsequence. When a sequence of minimizers of f_n is relatively compact, then (2.3) says that these minimizers of f_n converge to the set of minimizers of f. By converging to a subset A of a metric space Θ with metric ν, we mean that $\nu(\theta_n, A) \to 0$, where $\nu(\theta, A) := \inf_{\theta' \in A} \nu(\theta, \theta')$. Since we always define $\inf \emptyset = +\infty$, $\theta_n \to \arg\inf_\Theta f$ implies that $\arg\inf_\Theta f$ is not empty, that is, minimizers of f exist.

The epi-convergence approach essentially splits convergence of minimizers into a local and a global component. The local component is epi-convergence. For the case of lossy MLEs (and several variants), the required epi-convergence results are given in Harrison (2003) [10]. The global component is a compactness requirement. We want to prevent the sequence of lossy MLEs from "wandering to infinity" and to ensure that they are eventually contained in a compact set. Then we can use Proposition 2.1 to show that the sequence of lossy MLEs converges to minimizers of the limiting function. The main results of this paper consist of describing some examples where this global compactness condition holds.

3 Lossy MLEs

We begin with the setup used throughout the remainder of the paper. $(S, \mathcal{S})$ and $(T, \mathcal{T})$ are standard measurable spaces.[1] $(X_n)_{n\geq 1}$ is a stationary and ergodic random process on $(S^{\mathbb{N}}, \mathcal{S}^{\mathbb{N}})$ with distribution P which is assumed to be complete. $\rho : S \times T \to [0, \infty)$ is an $\mathcal{S} \times \mathcal{T}$-measurable function ($\mathcal{S} \times \mathcal{T}$ denotes the smallest product σ-algebra).

Let $(\Theta, \mathcal{B})$ be a separable metric space with its Borel σ-algebra and metric ν. We use $O(\theta, \epsilon) := \{\theta' \in \Theta : \nu(\theta, \theta') < \epsilon\}$ to denote the ϵ-neighborhood of θ. To each $\theta \in \Theta$ we associate a probability measure Q_θ on $(T^{\mathbb{N}}, \mathcal{T}^{\mathbb{N}})$. We use $(Y_n)_{n\geq 1}$ to denote a random

[1] Standard measurable spaces include Polish spaces and let us avoid uninteresting pathologies while working with random sequences [8].

sequence on $T^{\mathbb{N}}$. Typically, its distribution will be one of the Q_θ and this will be clear from the context. We use E_θ to denote E_{Q_θ}, the expectation with respect to (w.r.t.) Q_θ.

We allow for two different ways that the topology on Θ is related to the measures Q_θ. Let $Q_{\theta,n}$ be the nth marginal of Q_θ, i.e., the distribution on $(T^n, \mathcal{T}^n)$ of $(Y_1, \ldots, Y_n)$ under Q_θ. We assume that either

$$\theta_m \to \theta \text{ implies } Q_{\theta_m,n} \xrightarrow{\tau} Q_{\theta,n} \text{ as } m \to \infty \text{ for each } n, \tag{3.1}$$

or

$$(T, \mathcal{T}) \text{ is a separable metric space with its Borel } \sigma\text{-algebra}, \tag{3.2a}$$

$$\rho(x, \cdot) \text{ is continuous for each } x \in S, \tag{3.2b}$$

$$\theta_m \to \theta \text{ implies } Q_{\theta_m,n} \xrightarrow{w} Q_{\theta,n} \text{ as } m \to \infty \text{ for each } n. \tag{3.2c}$$

τ-Convergence is setwise convergence of probability measures.[2] w-Convergence is weak convergence of probability measures.[3] When T is discrete, assumptions (3.1) and (3.2) are equivalent. When each Q_θ is independent and identically distributed (i.i.d.), then (3.1) and (3.2c) will hold whenever they hold for $n = 1$.

Fix $D \geq 0$. For each $x_1^n \in S^n$, $y_1^n \in T^n$ and $\theta \in \Theta$, define

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

$$L_n(\theta, x_1^n) := -\frac{1}{n}\log Q_\theta(B(x_1^n, D)),$$

$$\Lambda_n(\theta, \lambda) := \frac{1}{n} E_P \log E_\theta e^{\lambda n \rho_n(X_1^n, Y_1^n)}, \qquad \Lambda_\infty(\theta, \lambda) := \limsup_{n\to\infty} \Lambda_n(\theta, \lambda),$$

$$\Lambda_n^*(\theta) := \sup_{\lambda \leq 0}[\lambda D - \Lambda_n(\theta, \lambda)], \quad n = 1, \ldots, \infty,$$

where log denotes the natural logarithm $\log_e$. $B(x_1^n, D)$ is called the single-letter distortion ball of radius D around x_1^n and ρ is called the single-letter distortion function. L_n is just (1.2) normalized to give per-symbol code lengths. Many properties of these quantities can be found in the literature [9, 10]. An important property here is that $L_n(\cdot, x_1^n)$ is l.sc. on Θ for each x_1^n [9].

Several recent papers [5, 7, 10] give conditions for which

$$\lim_{n\to\infty} L_n(\theta, X_1^n) \overset{\text{a.s.}}{=} \Lambda_\infty^*(\theta), \tag{3.3}$$

which Dembo and Kontoyiannis (2002) [7] call the *generalized AEP (Asymptotic Equipartition Property)*. Because of this, we are interested in approximating minimizers of Λ_∞^* via minimizers of $L_n(\cdot, X_1^n)$. We say that a sequence of mappings $(\hat{\theta}_n)_{n\geq 1}$ with $\hat{\theta}_n : S^{\mathbb{N}} \to \Theta$ is a sequence of *lossy MLEs* if

$$\text{Prob}\left\{\limsup_{k\to\infty} L_{n_k}(\hat{\theta}_{n_k}(X_1^\infty), X_1^{n_k}) \leq \limsup_{k\to\infty} \inf_{\theta\in\Theta} L_{n_k}(\theta, X_1^{n_k}), \quad \forall n_k \to \infty\right\} = 1. \tag{3.4}$$

[2] $Q_m \xrightarrow{\tau} Q$ if $E_{Q_m} f \to E_Q f$ for all bounded, measurable f, or equivalently, if $Q_m(A) \to Q(A)$ for all measurable A.

[3] $Q_m \xrightarrow{w} Q$ if $E_{Q_m} f \to E_Q f$ for all bounded, continuous f, or equivalently, if $Q_m(A) \to Q(A)$ for all measurable A with $Q(\partial A) = 0$.

(3.4) is just a probabilistic translation of (2.2) into the lossy MLE terminology. $(n_k)_{k\geq 1}$ is any infinite subsequence. In the special case where

$$L_n(\hat{\theta}_n(x_1^\infty), x_1^n) = \inf_{\theta\in\Theta} L_n(\theta, x_1^n), \quad \text{for each } n \text{ and each } x_1^\infty \in S^{\mathbb{N}},$$

we say that $\hat{\theta}_n$ is an *exact lossy MLE*. Notice that exact lossy MLEs are also lossy MLEs. Lossy MLEs always exist, but exact lossy MLEs may not exist. If Θ is compact, however, exact lossy MLEs do exist because L_n is l.sc. on Θ.

In this paper we do not need to assume that the lossy MLEs are measurable functions. The completeness of P takes care of a lot of problems. More detailed analysis of lossy MLEs, such as distributional properties, will require the assumption that the lossy MLEs are measurable (as will a relaxation of the completeness assumption). It also seems reasonable to require that lossy MLEs are predictable and independent of P, in the sense that the value of $\hat{\theta}_n$ only depends on x_1^n, the data up to time n, and that (3.4) holds for every realization x_1^∞, not just for almost every realization. In the Appendix we prove that lossy MLEs with these properties always exist as long as Θ is σ-compact, and whenever possible, we can suppose that this lossy MLE is exact.[4]

Proposition 3.1. Suppose Θ is σ-compact and $\epsilon_n : S^n \to (0,\infty)$ is (Borel) measurable. Then there exists an $\mathcal{S}^n/\mathcal{B}$-measurable mapping $\hat{\theta}_n : S^n \to \Theta$ such that for each $x_1^n \in S^n$

$$L_n(\hat{\theta}_n(x_1^n), x_1^n) = \min_{\theta\in\Theta} L_n(\theta, x_1^n) \text{ if the minimum exists, and} \tag{3.5}$$

$$L_n(\hat{\theta}_n(x_1^n), x_1^n) \leq \inf_{\theta\in\Theta} L_n(\theta, x_1^n) + \epsilon_n(x_1^n) \text{ otherwise.} \tag{3.6}$$

4 Consistency of Lossy MLEs

For easy reference, we translate Proposition 2.1 into the lossy MLE terminology.[5] Henceforth, we suppress the dependence of $\hat{\theta}_n$ on x_1^∞. If we are making a probabilistic statement about $\hat{\theta}_n$, then we mean $\hat{\theta}_n(X_1^\infty)$. Define the (possibly empty) set

$$\Theta^* := \arg\inf_{\theta\in\Theta} \Lambda_\infty^*(\theta) := \{\theta \in \Theta : \Lambda_\infty^*(\theta) = \Lambda_\infty^*(\Theta)\}, \quad \text{where } \Lambda_\infty^*(\Theta) := \inf_{\theta\in\Theta} \Lambda_\infty^*(\theta).$$

Corollary 4.1. Suppose

$$\text{Prob}\left\{\operatorname*{epi-lim}_{n\to\infty} L_n(\cdot, X_1^n) = \Lambda_\infty^*\right\} = 1. \tag{4.1}$$

Then every sequence of lossy MLEs $(\hat{\theta}_n)_{n\geq 1}$ satisfies

$$\text{Prob}\left\{\limsup_{n\to\infty} L_n(\hat{\theta}_n, X_1^n) \leq \limsup_{n\to\infty} \inf_{\theta\in\Theta} L_n(\theta, X_1^n) \leq \Lambda_\infty^*(\Theta)\right\} = 1. \tag{4.2}$$

If $(\hat{\theta}_n)_{n\geq 1}$ is a sequence of lossy MLEs such that

$$\text{Prob}\left\{(\hat{\theta}_n)_{n\geq 1} \text{ is relatively compact}\right\} = 1, \tag{4.3}$$

[4] We do not need the assumption that P is complete to show the existence of measurable lossy MLEs (Proposition 3.1). A metric space is σ-compact if it is a countable union of compact sets. Every locally compact, separable metric space is σ-compact.

[5] Notice that Corollary 4.1 does not actually make use of assumptions (3.1) or (3.2), however, these are used later to establish the hypotheses of the Corollary and also to establish the existence of measurable lossy MLEs.

then

$$\operatorname{Prob}\left\{\hat{\theta}_n \to \Theta^*\right\} = 1, \tag{4.4}$$

$$\operatorname{Prob}\left\{\lim_{n\to\infty} L_n(\hat{\theta}_n, X_1^n) = \lim_{n\to\infty}\inf_{\theta\in\Theta} L_n(\theta, X_1^n) = \Lambda_\infty^*(\Theta)\right\} = 1. \tag{4.5}$$

If $(\hat{\theta}_n)_{n\geq 1}$ is a sequence of lossy MLEs satisfying (4.4) and Θ^* is compact, then $(\hat{\theta}_n)_{n\geq 1}$ satisfies (4.3) and (4.5).

In fact, every sequence of lossy MLEs satisfies (4.3) if and only if Θ^* is compact and every sequence of lossy MLEs satisfies (4.4).

The epi-limit in (4.1) is a functional convergence and must hold at each point in Θ. More specifically, (4.1) says that the set of x_1^∞ for which the sequence of functions $L_n(\cdot, x_1^n)$, $n \geq 1$, epi-converges to the function Λ_∞^* at every point $\theta \in \Theta$ has P-probability 1. (4.4) is the main result that we want to prove in this paper. The next two sections are devoted to establishing the hypotheses of the Corollary, in particular, epi-convergence of L_n and relative compactness of lossy MLEs, so that we can conclude (4.4). We often need to assume that $\Lambda_\infty^*(\Theta) < \infty$ to conclude that lossy MLEs are relatively compact. For the purpose of establishing (4.4), however, this causes no loss of generality. When $\Lambda_\infty^*(\Theta) = \infty$, then $\Theta^* = \Theta$ and (4.4) is trivially true.

If Λ_∞^* has a unique minimizer θ^*, that is $\Theta^* = \{\theta^*\}$, then (4.4) becomes

$$\operatorname{Prob}\left\{\hat{\theta}_n \to \theta^*\right\} = 1.$$

In a typical statistical setting, θ^* is the unique point corresponding to the distribution of the source and (4.4) is called *(strong) consistency* of the estimator $\hat{\theta}_n$. Consistency is often a starting point for more detailed analysis, such as asymptotic normality. One of the nice properties of MLEs is consistency under a wide variety of conditions [12]. As we will show, this property carries over to lossy MLEs. We call (4.4) the *(strong) consistency of lossy MLEs*. Note, however, that in the lossy setting θ^* need not be unique and even if it is unique it need not correspond with the source distribution, so this is not consistency in the usual sense.

Harrison (2003) [9] gives conditions for which (4.1) holds, including necessary and sufficient conditions when Θ is a convex family of probability measures with Q_θ i.i.d. θ. We cannot ensure that (4.3) holds for every sequence of lossy MLEs without further assumptions. The simplest assumption to add is that Θ is compact. Then (4.3) is trivially true and epi-convergence (4.1) implies consistency (4.4). When Θ is not compact, (4.3) seems difficult to verify with any generality and we must verify it for specific settings.[6]

In the presence of epi-convergence, there are other characterizations and implications of consistency. Here is a useful one. A proof can be found in the Appendix.

Proposition 4.2. Suppose (4.1) holds. For each $\epsilon > 0$, let $\Theta_\epsilon^* := \bigcup_{\theta\in\Theta^*} O(\theta, \epsilon)$ be the ϵ-neighborhood of Θ^* (and empty if Θ^* is empty). If

$$\operatorname{Prob}\left\{\liminf_{n\to\infty} \inf_{\theta\notin\Theta_\epsilon^*} L_n(\theta, X_1^n) > \Lambda_\infty^*(\Theta)\right\} = 1 \quad \text{for each } \epsilon > 0, \tag{4.6}$$

then every sequence of lossy MLEs is consistent (4.4) and $\Lambda_\infty^*(\Theta) < \infty$. If Θ^* is compact, then the converse is also true.

[6] This seems to be a common situation in statistical minimization. The local convergence conditions can be verified in great generality and the global compactness conditions must be verified on a case-by-case basis.

4.1 A special nonparametric case

We begin with a special case where the reproduction distributions Q_θ are i.i.d. and every possible (i.i.d.) distribution is indexed by the parameter space Θ. Assume (3.2a) and (3.2b) and let Θ be the set of all probability measures on $(T, \mathcal{T})$ with a metric that metrizes weak convergence of probability measures, such as the Prohorov metric. For each $\theta \in \Theta$, let Q_θ be i.i.d. with distribution θ, that is $Q_{\theta,1} = \theta$. Θ is convex and separable and (3.2c) holds [4].

Assume that either

$$E\left[\inf_{y\in T}\rho(X_1,y)\right] \neq D \text{ or } \inf_{y\in T}\rho(X_1,y) \text{ is a.s. constant or } \Lambda^*_\infty(\Theta)=\infty. \tag{4.7}$$

Under these conditions (and the rest of the setup from Section 3) (4.1) holds [9].

Assume that for each $\epsilon > 0$ and each $M > 0$ there exists a $K \in \mathcal{S}$ such that

$$P_1(K) > 1-\epsilon \text{ and } B(K,M) := \bigcup_{x\in K} B(x,M) \text{ is relatively compact.} \tag{4.8}$$

P_1 is the first marginal of P, that is, the distribution of X_1. These assumptions give (4.3) and the existence of measurable lossy MLEs (Proposition 3.1) as shown in the Appendix.

Proposition 4.3. (4.1) is true and Θ is σ-compact. If $\Lambda^*_\infty(\Theta) < \infty$, then every sequence of lossy MLEs satisfies (4.3), (4.4) and (4.5) and Θ^* is nonempty, convex and compact.

Combining Corollary 4.1 and Proposition 4.3 shows that lossy MLEs are consistent.

Theorem 4.4. Every sequence of lossy MLEs is consistent (4.4).

The setup described here includes several standard cases. If $\inf_{y\in T}\rho(x,y) \le D$ for all x, then (4.7) is trivially true. If T is compact, then (4.8) is trivially true. This includes the case where T is a finite alphabet. If S and T are both subsets of (the same) finite dimensional Euclidean space, T is complete and $\inf_{\|x-y\|>m}\rho(x,y) \to \infty$ as $m \to \infty$, then (4.8) is true. For example, if $S = T \subset \mathbb{R}^d$ is complete and $\rho(x,y) = \|x-y\|_p$ is L^p-error distortion for some $1 \le p \le \infty$, then both (4.7) and (4.8) are true. Notice that this last case includes $S = T \subset \mathbb{Z}$ with L^p-error distortion.

Since each Q_θ is i.i.d. Λ^*_n does not depend on n, $1 \le n \le \infty$, and we have [10]

$$\Lambda^*_\infty(\theta) = R(P_1,\theta,D) := \inf_W H(W\|P_1\times\theta),$$

where the infimum is taken over all probability measures W on $(S\times T, \mathcal{S}\times\mathcal{T})$ such that W has S-marginal P_1 and $E_W\rho(X,Y) \le D$. $H(\mu\|\nu)$ denotes the relative entropy in nats

$$H(\mu\|\nu) := \begin{cases} E_\mu \log\frac{d\mu}{d\nu} & \text{if } \mu \ll \nu, \\ \infty & \text{otherwise.}\end{cases}$$

Any $\theta^* \in \Theta^*$ thus satisfies [7, 21]

$$R(P_1,\theta^*,D) = \inf_{\theta\in\Theta} R(P_1,\theta,D) = R(P_1,D) := \inf_W I(X;Y),$$

where the final infimum is over the same set as in the definition of $R(P_1,\theta,D)$ and $I(X;Y)$ is the mutual information in nats between the random variables X on S and Y on T

which have joint distribution W. $R(P_1, D)$ is the (information) rate distortion function in nats for a memoryless source with distribution P_1. If the source $(X_n)_{n\geq 1}$ is actually memoryless, then $R(P_1, D) = R(P, D) = R(D)$ is the rate distortion function for this source and any $\theta^* \in \Theta^*$ "achieves" the optimal rate [10, 15]. (See Section A.2 in the Appendix.)

We can relax (4.7). Indeed, if (4.7) does not hold, then (4.1) still holds along the subsequence (which is a.s. infinite) where $\inf_{\theta\in\Theta} L_n(\theta, X_1^n) < \infty$ [9]. If we insist that lossy MLEs do not change value if $L_n(\theta, X_1^n) = \infty$ for all θ, then any such sequence will still be relatively compact and consistent. In situations, such as those described above, where (4.8) does not depend on the structure of P_1, lossy MLEs are *always* consistent (with this caveat when $L_n \equiv \infty$) regardless of the source statistics P_1 or the distortion level D.

4.2 General parametric conditions

Now we give some general conditions for which both (4.1) and (4.3) are true. Unlike the previous section, we allow the reproduction distributions Q_θ to have some memory and we allow more freedom in the parameterization Θ. We continue to assume everything from Section 3, in particular we assume that either (3.1) or (3.2) holds.

We begin with conditions that control the mixing properties of the Q_θ. Assume that each Q_θ is stationary and assume that there exists a finite $C \geq 1$ such that for each $\theta \in \Theta$ we have

$$Q_\theta(A \cap B) \leq C Q_\theta(A) Q_\theta(B), \tag{4.9}$$

for all $A \in \sigma(Y_1^n)$ and $B \in \sigma(Y_{n+1}^\infty)$ and any n. Variants of this mixing condition arise when extending the generalized AEP to cases with memory [5, 9, 10]. If each Q_θ is i.i.d., then this is trivially true ($C = 1$). When the Q_θ are allowed to have memory, then the assumption that C is fixed independent of θ is quite restrictive.

Define

$$\Theta_{\lim} := \left\{\theta : \limsup_{n\to\infty} L_n(\theta, X_1^n) \leq \Lambda_\infty^*(\theta)\right\}, \tag{4.10}$$

$$\Theta_r := \{\theta : \Lambda_\infty^*(\theta) < r\}, \quad r \leq \infty, \tag{4.11}$$

and assume that

$$\Theta_{\lim} \cap \Theta_r \text{ is dense in } \Theta_r \text{ for each } r. \tag{4.12}$$

This assumption is discussed further in Harrison (2003) [9] where it is shown that (4.9) and (4.12) imply (4.1).

Now we introduce a condition that gives (4.3). Assume that there exists a $\Delta > 0$ and a $K \in \mathcal{S}$ with $P_1(K) > D/(D + \Delta)$ such that for every $\epsilon > 0$ the set

$$A_\epsilon := \{\theta : Q_\theta(B(K, D + \Delta)) \geq \epsilon\} \text{ is relatively compact,} \tag{4.13}$$

where $B(K, M) := \bigcup_{x\in K} B(x, M)$. If Θ is compact, then this condition is trivially true (take $K = S$). Note that this property only depends on the distribution P_1 of X_1 and the first marginals $Q_{\theta,1}$, i.e., the allowed distributions of Y_1. The strong mixing condition (4.9) lets us handle cases with memory using only the first marginals. The intuition for this assumption is illustrated somewhat by the next two results, the proofs of which can be found in the Appendix.

Proposition 4.5. Suppose that $S = T := \mathbb{R}^d$ is finite dimensional Euclidean space and that $\lim_{m\to\infty} \inf_{\|x-y\|>m} \rho(x,y) > D$. Then we can choose $K \in \mathcal{S}$ and $\Delta > 0$ with $P(K) > D/(D+\Delta)$ so that $B(K, D+\Delta)$ is bounded.

Proposition 4.6. Let Z be a random vector on $\mathbb{R}^d$ with a distribution that is absolutely continuous w.r.t. d-dimensional Lebesgue measure. For each vector $\mu \in \mathbb{R}^d$ and each $d \times d$ matrix $M \in \mathbb{R}^{d\times d}$, let $q_{\mu,M}$ be the distribution of $MZ + \mu$. Then for each $\epsilon > 0$ and each bounded subset $B \subset \mathbb{R}^d$,

$$\left\{(\mu, M) \in \mathbb{R}^{d+d\times d} : q_{\mu,M}(B) \geq \epsilon\right\}$$

is relatively compact.

In the Appendix we show that this compactness assumption gives (4.3).

Proposition 4.7. (4.1) is true. If $\Lambda^*_\infty(\Theta) < \infty$, then every sequence of lossy MLEs satisfies (4.3), (4.4) and (4.5) and Θ^* is nonempty and compact.

Combining Corollary 4.1 and Proposition 4.7 gives the consistency of lossy MLEs.

Theorem 4.8. Every sequence of lossy MLEs is consistent (4.4).

4.3 Examples

We now give some examples that satisfy the assumptions needed for Theorem 4.8. We always assume that $(S, \mathcal{S})$ and $(T, \mathcal{T})$ are standard measurable spaces, that $(X_n)_{n\geq 1}$ is stationary and ergodic, taking values in S, with a distribution P that is complete and that $\rho : S \times T \to [0, \infty)$ is $\mathcal{S} \times \mathcal{T}$-measurable.[7] The rest of the assumptions that are needed are addressed on a case by case basis. We closely follow the examples in Harrison (2003) [9], which satisfy (4.1). Some useful notation is

$$m(\theta, x) := \operatorname*{ess\,inf}_{Q_\theta} \rho(x, Y_1), \qquad D_{\min}(\theta) := Em(\theta, X_1).$$

4.3.1 Example: memoryless families, compact parameter space

Suppose that Θ is a compact, separable metric space, that each Q_θ is i.i.d. (memoryless) and that either (3.1) or (3.2) holds. (4.9) and (4.13) are trivially true. If (4.12) is true, then Theorem 4.8 gives the consistency of lossy MLEs. There are many situations where (4.12) holds, including the case where $D_{\min}(\theta) < D$ for all $\theta \in \Theta$ [9]. The following examples in Harrison (2003) [9] all work: Example 2.2.3 (the class of all probability measures on T with the weak convergence topology) with the modification that T is compact (and thus Θ is compact [4]) and the finite dimensional cases of Examples 2.2.4 (finite alphabet T) and 2.2.5 (finite mixing coefficients).

4.3.2 Example: memoryless location-scale families

Suppose $S = T := \mathbb{R}^d$ is finite dimensional Euclidean space, that ρ is continuous and satisfies the hypotheses of Proposition 4.5, that each Q_θ is i.i.d. (memoryless) and that $\{Q_{\theta,1}\}_{\theta\in\Theta}$ is a location-scale family (scale family, location family, etc.) whose canonical

[7] If $\rho(\cdot, y)$ is measurable for each $y \in T$ and $\rho(x, \cdot)$ is continuous for each $x \in S$ (which is trivial if T is finite), then ρ is product measurable.

member has a density w.r.t. Lebesgue measure. Most typical parameterizations of Θ will give (3.2) and if we include degenerate cases, so that the parameter space is closed, Proposition 4.6 can often be used to show that (4.13) holds. All of this depends on the parameterization, of course, and needs to be checked for specific cases. (4.9) is trivially true. If (4.12) is true, then Theorem 4.8 gives the consistency of lossy MLEs. In the next example, we give a simple example of a location-scale family where everything works out.

4.3.3 Example: memoryless Gaussian families, squared-error distortion

Take $S = T := \mathbb{R}$ with the Euclidean metric and with squared-error distortion $\rho(x, y) := |x - y|^2$. Let $\Theta := \mathbb{R} \times [0, \infty)$ and write $\theta := (\mu, \sigma)$ for $\theta \in \Theta$. Define $Q_{(\mu,\sigma)}$ to be i.i.d. Normal(μ,σ^2), where we define Normal(μ,0) to be the point mass at μ. (3.2), (4.9) and (4.12) are all valid [9][Example 2.2.2]. Propositions 4.5 and 4.6 show that (4.13) is satisfied as well, so Theorem 4.8 gives the consistency of lossy MLEs.

4.3.4 Example: finite state Markov chains

Let T be a finite set and let $\{Q_\theta\}_{\theta \in \Theta_{\text{irr}}}$ be the class of stationary, first-order, irreducible Markov chains on T. Let Θ_{irr} be the corresponding set of probability transition matrices, which we can think about as a subset of $\mathbb{R}^{T \times T}$, and let ν be a metric on Θ_{irr} that is equivalent to the Euclidean metric when Θ_{irr} is viewed as a subset of $\mathbb{R}^{T \times T}$. Suppose $\Theta \subset \Theta_{\text{irr}}$ is closed (and thus compact). (4.13) holds. For example, let Θ correspond to the set of all Q_θ that have stationary probabilities bounded below by a fixed $\epsilon > 0$. If $E[\min_{y \in T} \rho(X_1, y)] \neq D$ or $D = 0$, then the remaining conditions for Theorem 4.8 are valid [9][Example 2.2.6] and we have the consistency of lossy MLEs. In the special case where $S = T$ and $\rho(x, x) = 0$ (such as Hamming distortion), then lossy MLEs are consistent regardless of the source statistics.

Notice that we had to artifically make Θ compact to apply Theorem 4.8. The set of all possible transition probabilities is also compact and would be a more natural parameter space, however, (4.9) is no longer true, and more importantly, we do not know if (4.1) holds. Redefine Θ to be the set of all probability transition matrices with the same metric as before and let each Q_θ be a Markov chain as before except with uniform initial distribution. Under the same conditions on D, Harrison (2003) [9] shows that $L_n(\theta, X_1^n)$ epi-converges to $\operatorname{lsc} \Lambda_\infty^*(\theta)$, the l.sc. envelope of Λ_∞^*. This shows that lossy MLEs converge to minimizers of $\operatorname{lsc} \Lambda_\infty^*$. We do not know if these minimizers always agree with minimizers of Λ_∞^*. If Λ_∞^* is l.sc. on all of Θ, then it is equal to its l.sc. envelope and lossy MLEs are consistent.

4.3.5 Example: maximizing the approximate lossy likelihood

Assume everything from Section 3. Define

$$R_n(\theta, x_1^n) := \sup_{\lambda \leq 0} \left[\lambda D - \frac{1}{n} \log E_\theta e^{\lambda n \rho_n(x_1^n, Y_1^n)} \right].$$

We think of R_n as an approximation to L_n. This can be a useful analytic approximation [21] and can sometimes be simpler to compute than L_n in applications [M. Madiman, personal communication].

We are interested in the behavior of minimizers of R_n. We can define a sequence of *lossy R-minimizers* $(\hat{\theta}_n)_{n \geq 1}$ exactly like in (3.4) except with L_n replaced by R_n. The proof of Proposition 3.1 in the Appendix shows that Proposition 3.1 holds with L_n replaced by R_n. Because Corollary 4.1 comes exactly from Proposition 2.1, we can replace L_n by R_n and "lossy MLEs" by "lossy R-minimizers" in Corollary 4.1. The proof of Proposition 4.2 shows that we can make the same changes there as well.

Suppose that (4.9) and (4.12) hold. Then (4.1) holds with L_n replaced by R_n [9][Example 2.2.9]. Suppose that (4.13) also holds. The proof of Proposition 4.7 shows that it holds with "lossy MLEs" replaced by "lossy R-minimizers", so Theorem 4.8 holds with this replacement as well.

Notice that in each of the above examples where we demonstrated the consistency of lossy MLEs, we also have the consistency of lossy R-minimizers. This is also true in Section 4.1, where in the proofs we actually prove relative compactness (and thus consistency) for lossy R-minimizers first and then infer the consistency of lossy MLEs.

Consider the situation and assumptions in Section 4.1. In this case

$$R_n(\theta, x_1^n) = R(P_{x_1^n}, \theta, D) \text{ and } \Lambda_\infty^*(\theta) = R(P_1, \theta, D)$$

which implies that

$$\inf_{\theta \in \Theta} R_n(\theta, x_1^n) = R(P_{x_1^n}, D) \text{ and } \Lambda_\infty^*(\Theta) = R(P_1, D),$$

where $R(\tilde{P}, \tilde{Q}, D)$ and $R(\tilde{P}, D)$ are defined in Section A.2 for probability measures $\tilde{P}$ and $\tilde{Q}$ on S and T, respectively. $P_{x_1^n}$ is the empirical probability distribution on S defined by x_1^n. P_1 is the first marginal of P. See Section A.2 and the proof of Proposition 4.3 for details.

The important point here is that $R(\tilde{P}, D)$ is the rate distortion function for an i.i.d. source with distribution $\tilde{P}$. The lossy R-minimizer version of Proposition 4.3, in particular (4.5), implies that whenever $R(P_1, D) < \infty$ we have

$$\text{Prob}\left\{R(P_{X_1^n}, D) \to R(P_1, D)\right\} = 1.$$

If the source $(X_n)_{n\geq 1}$ is i.i.d., then the rate distortion function computed from the data converges to the true rate distortion function of the source.

4.3.6 Example: penalized lossy MLEs

Assume everything from Section 3. Let $(F_n)_{n\geq 1}$ be a sequence of l.sc. functions $F_n : \Theta \to [0, \infty]$. We think of F_n as a penalty and we want to minimize $L_n(\cdot, X_1^n) + F_n(\cdot)$ over Θ. Just like in the previous example we can define *penalized lossy MLEs* by replacing L_n with $L_n + F_n$. Propositions 3.1 and 4.2 and Corollary 4.1 continue to hold with the corresponding changes.

Suppose that

$$\Theta_{\text{lim}} \cap \Theta_r \cap \{\theta : F_n(\theta) \to 0\} \text{ is dense in } \Theta_r \text{ for each } r$$

and that (4.1) holds for L_n. Then (4.1) holds with L_n replaced by $L_n + F_n$ [9][Example 2.2.8]. If all lossy MLEs are consistent, Θ^* is compact and $\Lambda_\infty^*(\Theta) < \infty$, then (4.6) holds for L_n and thus it holds with L_n replaced by $L_n + F_n$. So penalized lossy MLEs are consistent as well. Notice that in each of the above examples where we demonstrated the consistency of lossy MLEs, we also have the consistency of penalized lossy MLEs.[8]

[8] All of this can be extended to the more general case where F_n is allowed to depend on X_1^n. Appropriate conditions for ensuring (4.1) with L_n replaced by $L_n + F_n$ can be found in Harrison (2003) [9][Example 2.2.8]. In this case and if $F_n \geq 0$, then the arguments (via Proposition 4.2) that the consistency of lossy MLEs implies the consistency of penalized lossy MLEs continue to hold.

One of the reasons for adding a penalty is to ensure that (4.3) holds for any sequence of penalized lossy MLEs even though it may not hold for all lossy MLEs. We will now give a contrived example of this phenomenon. We begin with a class of examples where an exact lossy MLE is not consistent.

Fix $D \geq 0$. Let $S = T := \mathbb{N}$ with the discrete topology and let

$$\rho(x, y) := \begin{cases} 0 & \text{if } x \leq y, \\ f(x) & \text{otherwise,} \end{cases}$$

for some nonnegative, real-valued function f. Take $(X_n)_{n\geq 1}$ i.i.d. with $Ef(X_1) = \infty$. Define $\Theta := \mathbb{N} \cup \{0\}$ with the discrete topology. Let Q_0 be any i.i.d. probability with a generalized AEP with a finite limit. (For example, if P has finite entropy, we can take $Q_0 = P$.) Let Q_θ be i.i.d. point masses on θ for each $\theta \geq 1$ (i.e., $Y_k \overset{\text{a.s.}}{=} \theta$ for all k w.r.t. Q_θ).

For $\theta \geq 1$, we have $D_{\min}(\theta) := E[f(X_1)I_{(\theta,\infty)}(X_1)] = \infty$ and thus $L_n(\theta, X_1^n) \overset{\text{a.s.}}{\to} \Lambda_\infty^*(\theta) = \infty$ [10]. So $\Theta = \Theta_{\lim}$ and $\Theta^* = \Theta_\infty = \{0\}$. Notice that (3.1), (4.9) and (4.12) all hold, so (4.1) holds.

Define

$$\hat{\theta}_n(x_1^n) := \max_{1\leq k\leq n} x_k.$$

We have $L_n(\hat{\theta}_n(x_1^n), x_1^n) = 0$, so $\hat{\theta}_n$ is an exact lossy MLE. Notice that $\hat{\theta}_n \uparrow \infty$ a.s. It does not converge and cannot be consistent. This shows the importance of the compactness assumption (4.3) for convergence of minimizers.

Now we will show that a penalty can correct things, at least in a specific instance. Consider the same setup. Let X_1 have distribution $p(x) := 2^{-x}$, define $f(x) := 2^{2^{2^x}}$ and take $F_n(\theta) := f(\theta)/n$. Notice that $F_n(\theta) \to 0$ for each θ so (4.1) holds with L_n replaced by $L_n + F_n$. We will show that every sequence of penalized lossy MLEs is consistent with this penalty. Before going through the details, however, we have two remarks.

First, since we are using the discrete topology on Θ, consistency actually means that our estimator is eventually a.s. equal to a minimizer of Λ_∞^*. The convergence happens in finite time. Second, the penalty that we chose satisfies

$$F_n(\theta) = \frac{1}{n} F(\theta) \text{ with } \sum_{\theta\in\Theta} 2^{-F(\theta)} \leq 1. \tag{4.14}$$

Barron (1985) [3] shows that penalties satisfying (4.14) lead to consistent estimators in the penalized (lossless) MLE setting under great generality if the source distribution P is in the parameter space and $F(P) < \infty$. We do not know if an equivalent result holds for penalized lossy MLEs. We suspect that a special reproduction distribution Q^* will need to be in the parameter space to take the place of P in the lossless setting. That some such assumption is needed is demonstrated at the end of this section, where we show that in this particular example we can choose a penalty satisfying (4.14) for which penalized lossy MLEs are not consistent.

Define $\hat{\theta}_n$ as before. For n large enough we can bound

$$\beta_n := \text{Prob}\left\{\hat{\theta}_n := \max_{1\leq k\leq n} X_k \leq \log_2 \log_2 n\right\} = \left(1 - 2^{-\log_2 \log_2 n}\right)^n$$
$$= \left[\left(1 - \frac{1}{\log_2 n}\right)^{\log_2 n}\right]^{n/\log_2 n} \leq e^{-n/\log_2 n}.$$

So for n large enough

$$2^n \beta_{2^n} \leq 2^n e^{-2^n/n} \leq \left[\frac{2}{e}\right]^n,$$

which is summable. This implies that β_n is summable [18][Theorem 3.27] and the Borel-Cantelli Lemma gives

$$\text{Prob}\left\{\hat{\theta}_n > \log_2 \log_2 n \text{ eventually}\right\} = 1$$

which implies

$$\text{Prob}\left\{\frac{1}{n} f(\hat{\theta}_n) > \frac{2^n}{n} > D \text{ eventually}\right\} = 1. \tag{4.15}$$

Suppose that $1 \leq \theta < \hat{\theta}_n(x_1^n)$ for some x_1^n. Then

$$\frac{1}{n}\sum_{k=1}^n \rho(x_k, \theta) \geq \frac{1}{n}\rho(\hat{\theta}_n(x_1^n), \theta) = \frac{1}{n} f(\hat{\theta}_n(x_1^n)).$$

If the left side is greater than D then $L_n(\theta, x_1^n) = \infty$, so (4.15) gives

$$\text{Prob}\left\{L_n(\theta, X_1^n) = \infty,\ 1 \leq \theta < \hat{\theta}_n,\ \text{eventually}\right\} = 1.$$

Since $L_n(\hat{\theta}_n(x_1^n), x_1^n) = 0$ for each x_1^n and since the penalty is increasing in θ, we have

$$\text{Prob}\left\{\inf_{\theta \geq 1}[L_n(\theta, X_1^n) + F_n(\theta)] = F_n(\hat{\theta}_n) \text{ eventually}\right\} = 1. \tag{4.16}$$

The only fact about the penalty that we used to derive (4.16) is that F_n is increasing in θ. When $F_n(\theta) := f(\theta)/n$, we can combine (4.15) and (4.16) to get

$$\text{Prob}\left\{\liminf_{n\to\infty} \inf_{\theta \geq 1}[L_n(\theta, X_1^n) + F_n(\theta)] = \infty\right\} = 1.$$

The (penalized modification of) Proposition 4.2 shows that every sequence of penalized lossy MLEs is consistent with this penalty as claimed.

On the other hand, suppose we were using the penalty $F_n(\theta) := [2\log_2(\theta + 2)]/n$, which also satisfies (4.14). Since it is increasing in θ, (4.16) holds. We have

$$\text{Prob}\left\{\hat{\theta}_n := \max_{1 \leq k \leq n} X_k > 3\log_2 n\right\} \leq n\,\text{Prob}\left\{X_1 > 3\log_2 n\right\} = n 2^{-3\log_2 n} = n^{-2}$$

which is summable, so the Borel-Cantelli Lemma gives

$$\text{Prob}\left\{\hat{\theta}_n \leq 3\log_2 n \text{ eventually}\right\} = 1.$$

Combining this with (4.16) gives

$$\text{Prob}\left\{\liminf_{n\to\infty} \inf_{\theta \geq 1}[L_n(\theta, X_1^n) + F_n(\theta)] = 0\right\} = 1.$$

Proposition 4.2 implies the existence of an inconsistent sequence of penalized lossy MLEs with this penalty. If $\Lambda_\infty^*(\Theta) > 0$, then every sequence of penalized lossy MLEs is inconsistent with this penalty.

A Appendix

The Appendix begins with justification of Proposition 2.1 and its Corollary 4.1. Then we prove some results needed for Section 4.1 and Proposition 4.3. Some of these may have independent interest so we allow for slightly more generality than is needed in the text. Next, we state some measurability results that are needed for establishing the existence of measurable lossy MLEs (Proposition 3.1) and also for showing that lossy likelihoods are well-behaved from a stochastic minimization perspective (Proposition A.11). The end of the Appendix is devoted to the proofs of the propositions found in the text.

A.1 Epi-convergence

Proposition 2.1 is well known, but I did not find a reference that states it in the form given here, particularly when minimizers are defined like (2.2). The proof is simple. The l.sc. of f and (2.1) can be found in any reference on epi-convergence [1]. Let $\Theta^* := \arg\inf_\Theta f$, let $f(\Theta) := \inf_\Theta f$ and let ν be the metric on Θ.

Let $(\theta_n)_{n\geq 1}$ be a relatively compact sequence satisfying (2.2). Suppose that (2.3) does not hold. Choose $(\theta_{n_k})_{k\geq 1}$ such that $\nu(\theta_{n_k}, \Theta^*) > \epsilon$ for all k and some $\epsilon > 0$ and such that $\theta_{n_k} \to \theta$ for some $\theta \in \Theta$. Clearly $\nu(\theta, \Theta^*) \geq \epsilon$, so $\theta \notin \Theta^*$. However, epi-convergence, (2.1) and (2.2) imply that $f(\theta) \leq \liminf_k f_{n_k}(\theta_{n_k}) \leq f(\Theta)$. So $\theta \in \Theta^*$ and (2.3) must hold. Similarly, supposing that (2.4) does not hold, lets us have $\lim_k f_{n_k}(\theta_{n_k}) < f(\Theta)$ and $\theta_{n_k} \to \theta$. Epi-convergence, (2.1) and (2.2) give us the same contradiction.

Now let $(\theta_n)_{n\geq 1}$ satisfy (2.2) and suppose that Θ^* is compact. Choose $\theta_n^* \in \Theta^*$ so that $\nu(\theta_n, \theta_n^*) < \nu(\theta_n, \Theta^*) + 1/n$. Since $(\theta_n^*)_{n\geq 1}$ is relatively compact and $\nu(\theta_n, \theta_n^*) \to 0$, we see that $(\theta_n)_{n\geq 1}$ is relatively compact.

Now suppose that every sequence satisfying (2.2) is relatively compact, but that Θ^* is not compact. Choose $\theta_n^* \in \Theta^*$ and $\epsilon_n > 0$ such that $(\theta_n^*)_{n\geq 1}$ has no convergent subsequence, $\epsilon_n \downarrow 0$ and the $O(\theta_n^*, \epsilon_n)$ are mutually disjoint. For each $n \geq 1$, use epi-convergence to choose a $\theta_n \in O(\theta_n^*, \epsilon_n)$ such that $f_n(\theta_n) \leq -n \vee f(\theta_n^*) + 1/n$. Notice that $(\theta_n)_{n\geq 1}$ is not relatively compact and that $\limsup_n f_n(\theta_n) \leq f(\Theta) = \lim_n \inf_\Theta f_n$, where the last equality comes from (2.4). But this means that $(\theta_n)_{n\geq 1}$ satisfies (2.2), which is a contradiction and Θ^* must be compact. This completes the proof of Proposition 2.1.

Translating Proposition 2.1 into Corollary 4.1 is straightforward. We just apply Proposition 2.1 along each realization x_1^∞ where epi-convergence and either relative compactness or consistency hold. Such realizations have probability 1. The only possible problem is the statement that if every sequence of lossy MLEs satisfies (4.3), then Θ^* is compact. There may be no particular realization x_1^∞ such that each sequence of lossy MLEs is actually a minimizer, much less relatively compact, so we cannot immediately apply Proposition 2.1. We can, however, imitate the proof in the preceeding paragraph.

Suppose that every sequence of lossy MLEs satisfies (4.3). Pick one, say $(\hat{\theta}_n)_{n\geq 1}$. Pick a realization x_1^∞ such that (4.1) holds and such that (3.4) and (4.3) hold for x_1^∞ and $\hat{\theta}_n(x_1^\infty)$. Notice that (4.5) also holds for this x_1^∞ and that the set of such x_1^∞ has probability 1. Assuming that Θ^* is not compact and repeating the proof from Proposition 2.1, shows that we can choose a sequence $(\theta_n)_{n\geq 1}$ that is not relatively compact but that has $\limsup_n L_n(\theta_n, x_1^n) \leq \Lambda_\infty^*(\Theta) = \lim_n \inf_{\theta\in\Theta} L_n(\theta, x_1^n)$. The sequence of mappings $x_1^\infty \mapsto \theta_n$ thus defines a sequence of lossy MLEs that is relatively compact with probability 0. This is a contradiction and Θ^* must be compact.

A.2 Minimizers of a rate distortion function

Let $(S,\mathcal{S})$ and $(T,\mathcal{T})$ be measurable spaces. Let $\rho : S\times T \to [0,\infty)$ be $\mathcal{S}\times\mathcal{T}$-measurable. For each probability measure P on $(S,\mathcal{S})$ and $D \geq 0$ define

$$W(P,D) := \left\{\text{probability measures } W \text{ on } (S\times T, \mathcal{S}\times\mathcal{T}) : W^S = P \text{ and } E_W\rho \leq D\right\},$$

where we use the notation W^S and W^T to denote the marginal distribution of W on S and T, respectively. The following definitions and equivalences are well known: [7, 21]

$$\begin{aligned} R(P,Q,D) &:= \inf_{W\in W(P,D)} H(W\|P\times Q) = \inf_{W\in W(P,D)} \left[H(W\|W^S\times W^T) + H(W^T\|Q)\right], \\ R(P,D) &:= \inf_{W\in W(P,D)} H(W\|W^S\times W^T) = \inf_Q R(P,Q,D), \\ \Lambda(P,Q,\lambda) &:= E_P \log E_Q e^{\lambda\rho(X,Y)}, \qquad \Lambda^*(P,Q,D) := \sup_{\lambda\leq 0}[\lambda D - \Lambda(P,Q,D), \end{aligned}$$

where Q denotes an arbitrary probability measure on $(T,\mathcal{T})$ and X and Y denote random variables on S and T, respectively. $H(\cdot\|\cdot)$ is the relative entropy in nats (see Section 4.1). If (X,Y) has joint distribution W, then $H(W\|W^S\times W^T) = I(X;Y)$, the mutual information (in nats) between X and Y. So $R(P,D)$ is the (information) rate distortion function (in nats) for the memoryless source with distribution P. Notice that we can replace W^S with P in each of the above definitions because $W\in W(P,D)$. As is typical, we define the infimum of the empty set to be $+\infty$.

The infimum is actually achieved in the definition of $R(P,Q,D)$.

Proposition A.1. [10] $R(P,Q,D) = \Lambda^*(P,Q,D)$. If $W(P,D)$ is not empty, then there exists a $W\in W(P,D)$ such that $R(P,Q,D) = H(W\|P\times Q)$.

Here we give some conditions for which the infimum is achieved in the two representations of $R(P,D)$ given above. This issue is addressed in detail by Csiszár (1974) [6]. The assumptions here are more general, although Csiszár allows ρ to be infinite valued and we do not.

We further assume that $(T,\mathcal{T})$ is a metric space with its Borel σ-algebra and that $\rho(x,\cdot)$ is l.sc. for each $x\in S$. In the appropriate topologies, $R(P,Q,D)$ is sequentially l.sc.

Proposition A.2. If $P_n \xrightarrow{\tau} P$ and $Q_n \xrightarrow{w} Q$ and $D_n \to D$, then $\liminf_n R(P_n,Q_n,D_n) \geq R(P,Q,D)$.

(Note that in this section P_n and Q_n refer to sequences of probability measures on S and T, respectively, and not the nth marginals of probability measures on a sequence space.) We use $\xrightarrow{\tau}$ to denote setwise convergence of probability measures and $\xrightarrow{w}$ to denote weak convergence of probability measures (see footnotes 2 and 3). If $\mathcal{Q}$ is a set of probability measures on $(T,\mathcal{T})$, we use $Q_n \xrightarrow{w} \mathcal{Q}$ to mean that every open neighborhood of $\mathcal{Q}$ (in the topology of weak convergence) contains each Q_n for large enough n (depending on the neighborhood). Since the topology of weak convergence is metrizable [20], this is convergence to a set in the usual manner for metric spaces.

A sequence of probability measures $(Q_n)_{n\geq 1}$ on $(T,\mathcal{T})$ is said to be *tight* if

$$\sup_{\substack{F\subset T \\ F \text{ compact}}} \liminf_{n\to\infty} Q_n(F) = 1.$$

If the sequence Q_n is tight, then Prohorov's Theorem states that the sequence Q_n is relatively compact in the topology of weak convergence of probability measures [13]. In particular, every subsequence has a subsequence that converges weakly to a probability measure.

Now we state the crucial assumption. This takes the place of the typical assumption that T is compact (c.f. Csiszár, 1974 [6]) and is trivial if T is compact. Assume that for each $\epsilon > 0$ and each $M > 0$ there exists a $K \subset S$ such that $P(K) > 1 - \epsilon$ and $B(K, M) \subset T$ is relatively compact, where

$$B(K, M) := \{y \in T : \rho(x, y) \leq M \text{ for some } x \in K\} = \bigcup_{x \in K} B(x, M)$$

in the notation of the text. Notice that this immediately implies that T is σ-compact. Section 4.1 describes some common situations where this assumption is valid. The key technical result is

Proposition A.3. If $P_n \xrightarrow{\tau} P$ and $D_n \to D$ and $\limsup_n R(P_n, Q_n, D_n) \leq R(P, D) < \infty$, then the sequence Q_n is tight.

We can use it to easily deduce the following:

Corollary A.4. The set of minimizers of $R(P, \cdot, D)$

$$\begin{aligned} \arg\inf_Q R(P, Q, D) &:= \{Q : R(P, Q, D) = \inf_{Q'} R(P, Q', D)\} \\ &= \{Q : R(P, Q, D) = R(P, D)\} \end{aligned}$$

is not empty. If $R(P, D) < \infty$, then $\arg\inf_Q R(P, Q, D)$ is compact in the topology of weak convergence of probability measures.

Corollary A.5. If $P_n \xrightarrow{\tau} P$ and $D_n \to D$ and $\limsup_n R(P_n, Q_n, D_n) \leq R(P, D)$, then $Q_n \xrightarrow{w} \arg\inf_Q R(P, Q, D)$.

Corollary A.6. There exists a Q such that $R(P, D) = R(P, Q, D)$. If $W(P, D)$ is not empty, then there exists a $W \in W(P, D)$ such that $R(P, D) = H(W \| W^S \times W^T)$.

A.2.1 Proof of Proposition A.2

Fix $\lambda \leq 0$ and $x \in S$. Since $e^{\lambda\rho(x,\cdot)}$ is bounded and u.sc. and since $Q_n \xrightarrow{w} Q$, we have $\limsup_n E_{Q_n} e^{\lambda\rho(x,Y)} \leq E_Q e^{\lambda\rho(x,Y)}$ [20][pp.313]. A generalization of Fatou's Lemma [17][p.269] gives

$$\liminf_n E_{P_n}\left[-\log E_{Q_n} e^{\lambda\rho(X,Y)}\right] \geq E_P\left[-\log E_Q e^{\lambda\rho(X,Y)}\right],$$

which implies

$$\liminf_n \left[\lambda D_n - \Lambda(P_n, Q_n, \lambda)\right] \geq \lambda D - \Lambda(P, Q, \lambda).$$

Taking the supremum over $\lambda \leq 0$ first inside the lim inf on the left and then on the right gives $\liminf_n \Lambda^*(P_n, Q_n, D) \geq \Lambda^*(P, Q, D)$. Proposition A.1 completes the proof.

A.2.2 Proof of Proposition A.3

For n large enough, $R(P_n, Q_n, D_n) < \infty$, so $W(P_n, D_n)$ is not empty and we can use Proposition A.1 to choose $W_n \in W(P_n, D_n)$ with $R(P_n, Q_n, D_n) = H(W_n \| P_n \times Q_n)$. Thus

$$\begin{aligned} R(P,D) &\geq \limsup_n R(P_n, Q_n, D_n) = \limsup_n H(W_n \| P_n \times Q_n) \\ &= \limsup_n \left[H(W_n \| P_n \times W_n^T) + H(W_n^T \| Q_n) \right] \\ &\geq \liminf_n H(W_n \| P_n \times W_n^T) + \limsup_n H(W_n^T \| Q_n) \\ &\geq \liminf_n R(P_n, W_n^T, D_n) + \limsup_n H(W_n^T \| Q_n). \end{aligned} \tag{A.1}$$

Suppose that W_n^T is tight. Then every subsequence has a subsequence that converges weakly. So we can choose a subsequence n_k such that $R(P_{n_k}, W_{n_k}^T, D_{n_k}) \to \liminf_n R(P_n, W_n^T, D_n)$ and such that $W_{n_k}^T \xrightarrow{w} Q$ for some probability measure Q on $(T, \mathcal{T})$. Applying Proposition A.2 to (A.1) gives

$$R(P,D) \geq R(P,Q,D) + \limsup_n H(W_n^T \| Q_n) \geq R(P,D) + \limsup_n H(W_n^T \| Q_n).$$

$R(P,D) < \infty$ so $\limsup_n H(W_n^T \| Q_n) = 0$. Since W_n^T is tight, Q_n is also tight.

To complete the proof, we will use the compactness assumption to show that W_n^T is tight. Fix $\epsilon > 0$ and $M > 2(D+\epsilon)/\epsilon$. Choose $K \subset S$ such that $P(K) > 1 - \epsilon/2$ and such that $B(K,M)$ is relatively compact. We can choose N large enough that $\sup_{n \geq N} D_n < D + \epsilon$, $\sup_{n \geq N} R(P_n, Q_n, D_n) < \infty$ and $\inf_{n \geq N} P_n(K) > 1 - \epsilon/2$.

For $n \geq N$, we have

$$D + \epsilon > D_n \geq E_{W_n} \rho(X,Y) \geq M W_n(K \times B(K,M)^c) \geq 2(D+\epsilon) W_n(K \times B(K,M)^c)/\epsilon.$$

This implies that $W_n(K \times B(K,M)^c) < \epsilon/2$ and we can bound

$$\begin{aligned} W_n^T(\overline{B(K,M)}) &\geq W_n^T(B(K,M)) \geq W_n(K \times B(K,M)) \\ &= P_n(K) - W_n(K \times B(K,M)^c) > 1 - \epsilon/2 - \epsilon/2 = 1 - \epsilon. \end{aligned}$$

Since $B(K,M)$ is relatively compact, it has compact closure $F := \overline{B(K,M)}$. We have just shown that $\liminf_n W_n^T(F) \geq 1 - \epsilon$. Since ϵ is arbitrary, W_n^T is tight and the proof is complete.

A.2.3 Proof of Corollaries

If $R(P,D) = \infty$, then $R(P,D) = R(P,Q,D) = H(W \| W^S \times W^T)$ for every Q and every $W \in W(P,D)$ (if there are any). So each of the Corollaries is trivially true.

Suppose $R(P,D) < \infty$. Consider the situation in Corollary A.5. Proposition A.3 shows that the sequence Q_n is tight. Suppose that $Q_n \not\to \arg\inf_Q R(P,Q,D)$. Then we can pick a subsequence Q_{n_k} and an open neighborhood $\mathcal{Q}$ of $\arg\inf_Q R(P,Q,D)$ such that $Q_{n_k} \in \mathcal{Q}^c$ for all k and such that $Q_{n_k} \xrightarrow{w} Q^*$ for some probability measure Q^*. (If $\arg\inf_Q R(P,Q,D) = \emptyset$, we can take $\mathcal{Q} = \emptyset$.) Since $\mathcal{Q}^c$ is closed, $Q^* \in \mathcal{Q}^c$ and thus $Q^* \notin \arg\inf_Q R(P,Q,D)$. On the other hand,

$$R(P,D) \geq \limsup_n R(P_n, Q_n, D_n) \geq \limsup_k R(P_{n_k}, Q_{n_k}, D_{n_k}) \geq R(P, Q^*, D) \geq R(P,D)$$

where the next to last inequality comes from Proposition A.2. So we have $R(P, Q^*, D) = R(P, D)$, which means $Q^* \in \arg\inf_Q R(P, Q, D)$. This is a contradiction, so $Q_n \overset{w}{\to} \arg\inf_Q R(P, Q, D)$ which therefore cannot be empty. This proves Corollary A.5.

Taking $P_n = P$ and $D_n = D$ and using the representation $R(P, D) = \inf_Q R(P, Q, D)$, shows that we can always satisfy the hypotheses of Corollary A.5, so $\arg\inf_Q R(P, Q, D)$ is not empty. Since $R(P, \cdot, D)$ is l.sc. (Proposition A.2), $\arg\inf_Q R(P, Q, D)$ is closed. If we choose a sequence of Q_n from $\arg\inf_Q R(P, Q, D)$, then the sequence Q_n must be tight. So every subsequence has a convergent subsequence and the limit must be in $\arg\inf_Q R(P, Q, D)$ because it is closed. This implies that $\arg\inf_Q R(P, Q, D)$ is sequentially compact and thus compact, because the topology of weak convergence is metrizable. This proves Corollary A.4.

We have already shown that $R(P, D) = R(P, Q, D)$ for some Q. For this Q, Proposition A.1 shows that we can choose $W \in W(P, D)$ (which is not empty since $R(P, D) < \infty$) such that

$$\begin{aligned} R(P, D) = R(P, Q, D) &= H(W \| P \times Q) = H(W \| P \times W^T) + H(W^T \| Q) \\ &= H(W \| W^S \times W^T) + H(W^T \| Q) \geq R(P, D) + H(W^T \| Q) \geq R(P, D). \end{aligned}$$

So $H(W^T \| Q) = 0$ and $R(P, D) = H(W \| W^S \times W^T)$. This proves Corollary A.6.

A.3 Measurability lemmas

For the lemmas given here let $(\Theta, \mathcal{B})$ be a separable metric space with its Borel σ-algebra and metric ν and let $(S, \mathcal{S})$ be an arbitrary measurable space. $\mathcal{S} \times \mathcal{B}$ denotes the smallest σ-algebra containing the measurable rectangles. If $f : \Theta \to [-\infty, \infty]$ is any function we use

$$\operatorname{lsc} f(\theta) := \sup_{\epsilon > 0} \inf_{\theta' \in O(\theta, \epsilon)} f(\theta') \quad \text{and} \quad \operatorname{usc} f(\theta) := \inf_{\epsilon > 0} \sup_{\theta' \in O(\theta, \epsilon)} f(\theta')$$

to denote the l.sc. and u.sc. envelopes of f, respectively. When $f : S \times \Theta \to [-\infty, \infty]$ we use $\operatorname{lsc} f$ to denote the l.sc. envelope of f w.r.t. $\theta \in \Theta$ and similarly for $\operatorname{usc} f$.

Lemma A.7. If $f : S \times \Theta \to [-\infty, \infty]$ satisfies

1. $f(s, \cdot)$ is u.sc. for each $s \in S$,
2. $f(\cdot, \theta)$ is $\mathcal{S}$-measurable for each $\theta \in \Theta$,

then

a. $\inf_{\theta \in U} f(\cdot, \theta)$ is $\mathcal{S}$-measurable for any $U \subset \Theta$,

b. $(s, \theta) \mapsto \inf_{\theta' \in O(\theta, \epsilon)} f(s, \theta)$ is $\mathcal{S} \times \mathcal{B}$-measurable for each $\epsilon > 0$,

c. $\operatorname{lsc} f$ is $\mathcal{S} \times \mathcal{B}$-measurable.

Proof. For part a, fix $U \subset \Theta$ and let $U_0 \subset U$ be countable and dense (w.r.t. U). Then $\inf_{\theta \in U} f(\cdot, \theta) = \inf_{\theta_0 \in U_0} f(\cdot, \theta_0)$, which is measurable.

For part b fix $\epsilon > 0$ and let $\Theta_0 \subset \Theta$ be countable and dense. Notice that

$$\inf_{\theta' \in O(\theta, \epsilon)} f(s, \theta') = \inf_{\theta_0 \in \Theta_0 \cap O(\theta, \epsilon)} f(s, \theta_0) = \inf_{\theta_0 \in \Theta_0} \left[f(s, \theta_0) I_{O(\theta_0, \epsilon)}(\theta) + \infty \cdot I_{O(\theta_0, \epsilon)^c}(\theta) \right],$$

which is $\mathcal{S} \times \mathcal{B}$-measurable in (s, θ). Part c follows by letting $\epsilon \downarrow 0$. □

Lemma A.8. If for each $n = 1, 2, \ldots$, $f_n : S \times \Theta \to [-\infty, \infty]$ satisfies

1. $f_n(s, \cdot)$ is l.sc. for each $s \in S$,
2. $f_n(\cdot, \theta)$ is $\mathcal{S}$-measurable for each $\theta \in \Theta$,
3. $\operatorname{usc} f_n \leq \sup_m f_m$,

then $f := \sup_m f_m$ satisfies

a. $f(s, \cdot)$ is l.sc. for each s,

b. f is $\mathcal{S} \times \mathcal{B}$-measurable,

c. $\inf_{\theta \in U} f(\cdot, \theta)$ is $\mathcal{S}$-measurable for any $U \subset \Theta$ such that U is a countable union of compact sets.

Proof. Part a is trivial. By redefining $f_n = \max_{k \leq n} f_n$, we can assume that $f_n \uparrow f$.

Since $-f_n$ satisfies the hypotheses of Lemma A.7, we know that $\operatorname{usc} f_n = \operatorname{lsc}(-f_n)$ is $\mathcal{S} \times \mathcal{B}$-measurable. So $\operatorname{usc} f_n$ also satisfies the hypotheses of Lemma A.7. Furthermore, $f_n \leq \operatorname{usc} f_n \leq f$, so $\operatorname{usc} f_n \uparrow f$. This proves part b.

For part c, first let $U \subset \Theta$ be compact. Clearly

$$\inf_{\theta \in U} f(\cdot, \theta) \geq \sup_n \inf_{\theta \in U} \operatorname{usc} f_n(\cdot, \theta),$$

the latter of which is measurable from Lemma A.7. To show the reverse inequality, fix s and choose $\theta_n \in U$ such that $\operatorname{usc} f_n(s, \theta_n) < \max\{-n, \inf_{\theta \in U} \operatorname{usc} f_n(s, \theta)\} + 1/n$. The compactness of U implies that $\theta_{n_k} \to \theta^*$ for some subsequence and some $\theta^* \in U$. So for each m,

$$\begin{aligned}\sup_n \inf_{\theta \in U} \operatorname{usc} f_n(s, \theta) &= \lim_{n \to \infty} \operatorname{usc} f_n(s, \theta_n) = \lim_{k \to \infty} \operatorname{usc} f_{n_k}(s, \theta_{n_k}) \\ &\geq \liminf_{k \to \infty} f_m(s, \theta_{n_k}) \geq f_m(s, \theta^*).\end{aligned}$$

Letting $m \to \infty$ gives

$$\sup_n \inf_{\theta \in U} \operatorname{usc} f_n(s, \theta) \geq f(s, \theta^*) \geq \inf_{\theta \in U} f(s, \theta)$$

and completes the argument for U compact. If $U = \bigcup_{n=1}^{\infty} U_n$, where U_n is compact, then $\inf_{\theta \in U} f(\cdot, \theta) = \inf_n \inf_{\theta \in U_n} f(\cdot, \theta)$, the latter of which is measurable. □

Lemma A.9. Let $f : S \times \Theta \to [-\infty, \infty]$ and $g : \Theta \to [-\infty, \infty]$ satisfy

1. $f(s, \cdot)$ is l.sc. for each $s \in S$,
2. $\inf_{\theta \in U} f(\cdot, \theta)$ is $\mathcal{S}$-measurable for each compact $U \subset \Theta$,
3. g is either l.sc. or u.sc.,
4. $f(s, \theta) + g(\theta)$ is well-defined for all s and θ.

Then $\inf_{\theta \in U}[f(\cdot, \theta) + g(\theta)]$ is $\mathcal{S}$-measurable for each $U \subset \Theta$ such that U is a countable union of compact sets.

Proof. We need only establish the case where g is l.sc. and bounded below. Indeed, if g is just l.sc., then $\max(g, -n)$ is l.sc. and bounded below and $\inf[f+g] = \inf_n \inf[f + \max(g, -n)]$, the latter of which is measurable and well-defined. Similarly, if g is u.sc., define

$$g_n(\theta) := \sup_{\epsilon>0} \inf_{\theta' \in O(\theta,\epsilon)} \sup_{\theta'' \in O(\theta', 1/n)} g(\theta'').$$

g_n is l.sc. and $g_n \downarrow g$. So $\inf[f+g] = \inf_n \inf[\max(f, -n) + g_n]$, the latter of which is measurable and well-defined.

Also, we need only establish the result for compact U (see the proof of Lemma A.8.c) and the case where U is compact follows directly from Pfanzagl (1969) [16, Lemma 3.8]. □

Lemma A.10. If $f : S \times \Theta \to [-\infty, \infty]$ and $\epsilon, \delta : S \to (0, \infty]$ satisfy

1. $f(s, \cdot)$ is l.sc. for each $s \in S$,

2. $\inf_{\theta \in U} f(\cdot, \theta)$ is $\mathcal{S}$-measurable for each compact $U \subset \Theta$,

3. ϵ and δ are $\mathcal{S}$-measurable,

then for each $U \subset \Theta$ such that U is a countable union of compact sets, there exists an $\mathcal{S}/\mathcal{B}$-measurable function $\hat{\theta} : S \to U$ such that for each $s \in S$

$$f(s, \hat{\theta}(s)) = \min_{\theta \in U} f(s, \theta) \text{ if the minimum exists, and}$$

$$f(s, \hat{\theta}(s)) \leq \max\left\{-\delta(s)^{-1}, \inf_{\theta \in U} f(s, \theta)\right\} + \epsilon(s) \text{ otherwise.}$$

Proof. Choose $U_n \uparrow U$, $n = 1, 2, \ldots$, where each U_n is compact. Since $f(s, \cdot)$ is l.sc., the minimum $\min_{\theta \in U_n} f(s, \theta)$ exists for each s. Pfanzagl (1969) [16, Theorem 3.10] shows that there exists an $\mathcal{S}/\mathcal{B}$-measurable function $\hat{\theta}_n : S \to U_n$ such that $f(s, \hat{\theta}_n(s)) = \min_{\theta \in U_n} f(s, \theta)$ for all $s \in S$.

Notice that $\inf_{\theta \in U} f(\cdot, \theta) = \inf_n \inf_{\theta \in U_n} f(\cdot, \theta)$ is $\mathcal{S}$-measurable, so

$$E_n := \left\{s : \inf_{\theta \in U_n} f(s, \theta) = \inf_{\theta \in U} f(s, \theta)\right\} \in \mathcal{S}$$

and $E_n \uparrow E := \{s : \text{the minimum } \min_{\theta \in U} f(s, \theta) \text{ exists}\} \in \mathcal{S}$. Let $(E, \mathcal{E})$ be the restriction of $(S, \mathcal{S})$ to E. Pfanzagl (1969) [16, Theorem 3.10] shows that there exists an $\mathcal{E}/\mathcal{B}$-measurable function $\hat{\theta}_E : E \to U$ such that $f(s, \hat{\theta}_E(s)) = \min_{\theta \in U} f(s, \theta)$ for all $s \in E$.

Define

$$A_n := \left\{s : \inf_{\theta \in U_n} f(s, \theta) \leq \max\left\{-\delta(s)^{-1}, \inf_{\theta \in U} f(s, \theta)\right\} + \epsilon(s)\right\}.$$

Each $A_n \in \mathcal{S}$ and $A_n \uparrow S$. Define $A_0 := \emptyset$.

Let $S_n := A_n \sim A_{n-1}$, $n = 1, 2, \ldots$. Then $S = \bigcup_n S_n$ and the S_n are disjoint. The function $\hat{\theta}(s) := \hat{\theta}_n(s)$, if $s \in S_n \cap E^c$, $n = 1, 2, \ldots$, and $\hat{\theta}(s) := \hat{\theta}_E(s)$, if $s \in E$, is $\mathcal{S}/\mathcal{B}$-measurable and has the desired properties. □

A.4 Proofs

Throughout the proofs we assume everything from Section 3 and also any of the specifics from the context where a proposition is stated.

A.4.1 Proof of Proposition 3.1

We begin with a result that is a common requirement for many statistical applications of epi-convergence.

Proposition A.11. $L_n(\cdot, x_1^n)$ is l.sc. for each $x_1^n \in S^n$. L_n is $\mathcal{B} \times \mathcal{S}^n$-measurable. For $U \subset \Theta$ such that U is a countable union of compact sets, $\inf_{\theta \in U} L_n(\theta, \cdot)$ is $\mathcal{S}^n$-measurable.

Proof. We use different methods for situations (3.1) and (3.2). In both cases, $L_n(\theta, \cdot)$ is measurable [10].

Suppose (3.1) holds. $L_n(\cdot, x_1^n)$ is continuous [9], so $\operatorname{lsc} L_n(\cdot, x_1^n) = \operatorname{usc} L_n(\cdot, x_1^n) = L_n(\cdot, x_1^n)$. Lemma A.7 completes the proof.

Now suppose (3.2) holds. Since $(T, \mathcal{T})$ is a separable metric space with its Borel σ-algebra, $(T^n, \mathcal{T}^n)$ is also. Similarly, ρ_n is product measurable and $\rho_n(x_1^n, \cdot)$ is continuous for each x_1^n. The structure of the problem does not change with n, so without loss of generality we will prove the result for $n = 1$. Let $O(y, r)$ denote the open r-neighborhood around $y \in T$.

For $m \geq 1$ define the functions

$$\rho^m(x, y) := \sup_{\epsilon > 0} \inf_{y' \in O(y, \epsilon + 1/m)} \rho(x, y').$$

$\rho^m(x, \cdot)$ is l.sc. for each x, ρ^m is $\mathcal{S} \times \mathcal{T}$-measurable (Lemma A.7) and $\rho^m \uparrow \rho$ as $m \uparrow \infty$. Define $B^m(x, D) := \{y \in T : \rho^m(x, y) \leq D\}$ and $L^m(\theta, x) := -\log Q_\theta(B^m(x, D))$. To complete the proof we need only show that $f_m(x, \theta) := L^m(\theta, x)$ and $f(x, \theta) := L(\theta, x)$ satisfy the assumptions for Lemma A.8.

Since $\rho^m(x, \cdot)$ is l.sc., $B^m(x, D)$ is closed and $\theta \mapsto Q_\theta(B^m(x, D))$ is u.sc. from a property of weak convergence [20][pp.311]. This shows that $L^m(\cdot, x)$ is l.sc. for each x. Since ρ^m is product measurable, $L^m(\theta, \cdot)$ is measurable [10]. Since $\rho^m \uparrow \rho$, $B^m(x, D) \downarrow B(x, D)$ and $Q_\theta(B^m(x, D)) \downarrow Q_\theta(B(x, D))$. This shows that $L^m \uparrow L$.

All that remains to prove is $\operatorname{usc} L^m \leq L$, where the u.sc. envelope is taken over Θ. This is equivalent to showing that

$$\operatorname{lsc} E_\theta I_{B^m(x,D)}(Y) \geq E_\theta I_{B(x,D)}(Y), \tag{A.2}$$

where the l.sc. envelope is taken over $\theta \in \Theta$. Since $B^m(x, D)$ is closed and ρ^m is product measurable, $I_{B^m(x,D)}(y)$ is u.sc. in y and product measurable in (x, y). Lemma A.7 shows that $\operatorname{lsc} I_{B^m(x,D)}(y)$ is product measurable, where the l.sc. envelope is taken over $y \in T$. We have $\operatorname{lsc} I_{B^m(x,D)}(y) \geq I_{B(x,D)}(y)$, so

$$E_\theta I_{B^m(x,D)}(Y) \geq E_\theta \left[\operatorname{lsc} I_{B^m(x,D)}(Y)\right] \geq E_\theta I_{B(x,D)}(Y).$$

But the middle expression is l.sc. in θ from a property of weak convergence [20][pp.313], so it is equal to its l.sc. envelope over $\theta \in \Theta$. This gives (A.2) and completes the proof. □

Now we prove Proposition 3.1. If Θ is σ-compact, then it is a countable union of compact sets. Proposition 3.1 follows from Proposition A.11 and Lemma A.10. We can ignore δ because L_n is nonnegative.

Lemma A.9 lets us derive the same result for certain types of penalties, namely l.sc. functions $F : \Theta \to [0, \infty)$. See Example 4.3.6. It is not hard to prove Proposition A.11 for R_n as defined in Example 4.3.5. Indeed, R_n is a supremum of continuous, measurable functions [9]. The functions are concave in the variable that is being maximized over [9], so the supremum can be taken over a fixed, countable set. Lemma A.8 will give the desired results.

A.4.2 Proof of Proposition 4.2

Suppose (4.6) is true. Combining it with (4.2) shows that every sequence of lossy MLEs is eventually contained in Θ^*_ϵ a.s. Since $\epsilon > 0$ was arbitrary, (4.4) holds. Obviously, (4.6) can only hold when $\Lambda^*_\infty(\Theta) < \infty$.

Now suppose that Θ^* is compact, $\Lambda^*_\infty(\Theta) < \infty$ and every sequence of lossy MLEs satisfies (4.4). Corollary 4.1 shows that every sequence of lossy MLEs satisfies (4.3). If (4.6) is not true for some $\epsilon > 0$, then we can find realizations x_1^∞ (the collection of which has positive probability) such that (4.4) and (4.3) hold and such that

$$\liminf_{n\to\infty} \inf_{\theta\notin\Theta^*_\epsilon} L_n(\theta, x_1^n) \le \Lambda^*_\infty(\Theta) = \lim_{n\to\infty} \inf_{\theta\in\Theta} L(\theta, x_1^n), \tag{A.3}$$

where the last equality comes from the second part of (4.5). But this final result implies that we can find a sequence of lossy MLEs that are not eventually in Θ^*_ϵ with positive probability, contradicting (4.4). The reason we need $\Lambda^*_\infty(\Theta) < \infty$ is to ensure that Θ^*_ϵ has a nonempty complement via (A.3); otherwise, the left side would be infinite.

A.4.3 Proof of Proposition 4.3

(4.7) implies (4.1) [9]. Since Λ^*_n does not depend on n, we will just write Λ^*. Notice that $\Lambda^*(\theta) = \Lambda^*(P_1, \theta, D) = R(P_1, \theta, D)$ and $\inf_{\theta\in\Theta} \Lambda^*(\theta) = R(P_1, D)$ in the notation of Section A.2. We will make frequent use of the results and methods of that section.

Henceforth we assume (4.8) and that $\inf_{\theta\in\Theta} \Lambda^*(\theta) < \infty$. It is not hard to see that Θ is σ-compact by covering it with a countable collection of the $\overline{B(K, M)}$. Corollary A.4 shows that Θ^* is nonempty, convex and compact.

For each $x_1^n \in S^n$, define the empirical probability measure $P_{x_1^n}$ on $(S, \mathcal{S})$ by

$$P_{x_1^n}(A) := \frac{1}{n} \sum_{k=1}^n I_A(x_k), \qquad A \in \mathcal{S}.$$

Using the notation of Section A.2, (4.1) implies [9][Example 2.2.9]

$$\text{Prob}\left\{ \underset{n\to\infty}{\text{epi-lim}}\, \Lambda^*(P_{X_1^n}, \theta, D) = \Lambda^*(\theta),\ \forall\theta \in \Theta \right\} = 1. \tag{A.4}$$

The equivalence in Proposition A.1 shows that we can rewrite this as

$$\text{Prob}\left\{ \underset{n\to\infty}{\text{epi-lim}}\, R(P_{X_1^n}, \theta, D) = R(P_1, \theta, D),\ \forall\theta \in \Theta \right\} = 1. \tag{A.5}$$

For each $\epsilon > 0$ and each $M > 0$, let $K(\epsilon, M)$ be the set in (4.8). Then the ergodic theorem gives

$$\text{Prob}\left\{ \lim_{n\to\infty} P_{X_1^n}(K(\epsilon, M)) = P_1(K(\epsilon, M)),\ \text{for all rational } \epsilon > 0,\ M > 0 \right\} = 1. \tag{A.6}$$

Let $(\hat\theta_n)_{n\ge1}$ be a sequence of lossy MLEs. (4.1) implies (4.2). Fix a realization x_1^∞ of X_1^∞ such that (4.1), (4.2), (A.4), (A.5) and (A.6) each hold. Let $\hat\theta_n$ denote $\hat\theta_n(x_1^n)$. We will show that the sequence $\hat\theta_n$ is tight in a manner analogous to the proof of Proposition A.3 and we use the notation found there. Although $P_{x_1^n}$ does not τ-converge to P_1 (unless P_1 is discrete), (A.6) and (A.5) are sufficient for what we need here.

Using (4.2), Chebyshev's inequality ($L_n \geq \Lambda^*$) and the different representations already mentioned gives

$$\begin{aligned} R(P,D) &= \inf_{\theta\in\Theta} \Lambda^*(\theta) \geq \limsup_n L_n(\hat{\theta}, x_1^n) \geq \limsup_n \Lambda^*(P_{x_1^n}, \hat{\theta}_n, D) \\ &= \limsup_n R(P_{x_1^n}, \hat{\theta}_n, D). \end{aligned} \tag{A.7}$$

Using (A.7) and repeating the steps of (A.1) with P_1, $P_{x_1^n}$, $\hat{\theta}_n$ and D taking the place of P, P_n, Q_n and D_n, respectively, gives

$$R(P_1, D) \geq \liminf_n R(P_{x_1^n}, W_n^T, D) + \limsup_n H(W_n^T \| \hat{\theta}_n). \tag{A.8}$$

Θ is the class of all probability measures on $(T, \mathcal{T})$ with the topology of weak convergence, so each W_n^T corresponds to some $\theta_n \in \Theta$. Suppose that W_n^T is tight. Then θ_n is relatively compact and we can choose a subsequence such that $\theta_{n_k} \to \theta$ for some $\theta \in \Theta$ and such that $R(P_{x_1^{n_k}}, \theta_{n_k}, D) \to \liminf_n R(P_{x_1^n}, W_n^T, D)$. Using (A.5) and the definition of epi-convergence gives

$$\liminf_n R(P_{x_1^n}, W_n^T, D) = \lim_k R(P_{x_1^{n_k}}, \theta_{n_k}, D) \geq R(P_1, \theta, D) \geq R(P_1, D).$$

Combining this with (A.8) gives

$$R(P_1, D) \geq R(P_1, D) + \limsup_n H(W_n^T \| \hat{\theta}_n) \geq R(P_1, D).$$

So $\limsup_n H(W_n^T \| \hat{\theta}_n) = 0$ and $\hat{\theta}_n$ is tight. This gives (4.3). Corollary 4.1 gives the rest of the results.

To complete the proof, we need only show that W_n^T is tight. The steps are identical to those in the proof of Proposition A.3 except that we must choose ϵ and M rational and use (A.6) instead of τ-convergence.

A.4.4 Proof of Proposition 4.5

Fix m and $\Delta > 0$ so that $\inf_{\|x-y\|\geq m} \rho(x,y) > D + \Delta$. Since $D/(D+\Delta) < 1$, we can choose N large enough so that $P(K) > D/(D+\Delta)$, where $K := \left\{x \in \mathbb{R}^d : \|x\| < N\right\}$ is the ball of radius N in $\mathbb{R}^d$.

Suppose $y \in B(K, D+\Delta)$. Then there exists an x with $\|x\| < N$ such that $\rho(x,y) \leq D + \Delta$. But this implies that $\|x - y\| < m$. The triangle inequality gives $\|y\| < N + m$. So $B(K, D+\Delta) \subset \mathbb{R}^d$ is bounded.

A.4.5 Proof of Proposition 4.6

The proof proceeds by establishing the proposition first for uniformly distributed Z, then for bounded Z with bounded density and finally for Z with arbitrary density. Note that the proposition is not true if Z is allowed to have point masses. In the proof, we use $\|\cdot\|$ to denote both the Euclidean norm for vectors and the operator norm for matrices, i.e., $\|M\| := \sup_{z:\|z\|=1} \|Mz\|$ for a matrix M and vectors z. Define $B : \{z \in \mathbb{R}^d : \|z\| \leq 1\}$ to be the closed unit ball at the origin, so that $rB + z$ is the ball of radius r centered at z. We use $I_B(z)$ to denote the indicator function that $z \in B$.

Fix $F \subset \mathbb{R}$ bounded and $\epsilon > 0$. Choose p so that $F \subset pB$. We will first prove the proposition under the added assumption that Z has uniform distribution on the unit ball, that is, the density of Z is $kI_B(z)$, where the k^{-1} is the volume of the unit ball in $\mathbb{R}^d$.

For any matrix A with $\|A\| = 1$, there exists a unit vector ϕ_A with $\|\phi_A\| = \|\phi_A^T A\| = 1$. Let $Z_1 \in \mathbb{R}$ denote the first coordinate of Z. For $m > 0$, we have

$$\begin{aligned}
&\sup_{\mu\in\mathbb{R}^d; M\in\mathbb{R}^{d\times d}:\|M\|=m} q_{\mu,M}(F) = \sup_{\mu;M:\|M\|=m} \text{Prob}\left\{MZ+\mu \in F\right\} \\
&\quad\le \sup_{\mu;A:\|A\|=1} \text{Prob}\left\{AZ \in \frac{p}{m}B - \mu\right\} \le \sup_{\mu;A:\|A\|=1} \text{Prob}\left\{\phi_A^T AZ \in \frac{p}{m}\phi_A^T B - \phi_A^T\mu\right\} \\
&\quad\le \sup_{c\in\mathbb{R};a,b\in\mathbb{R}^d:\|a\|=\|b\|=1} \text{Prob}\left\{a^T Z \in \frac{p}{m}b^T B - c\right\} \\
&\quad\le \sup_{c;a:\|a\|=1} \text{Prob}\left\{a^T Z \in \left[-\frac{p}{m} - c, \frac{p}{m} - c\right]\right\} = \sup_c \text{Prob}\left\{Z_1 \in \left[-\frac{p}{m} - c, \frac{p}{m} - c\right]\right\} \\
&\quad\le \text{Prob}\left\{Z_1 \in \left[-\frac{p}{m}, \frac{p}{m}\right]\right\} \downarrow 0 \quad \text{as } m \uparrow \infty. \qquad\qquad \text{(A.9)}
\end{aligned}$$

We used the fact that Z is uniform over B to reason that $a^T Z$ has the same distribution for all unit vectors a and therefore the same distribution as Z_1.

(A.9) implies that we can choose m_ϵ large enough so that $\|M\| > m_\epsilon$ implies $q_{\mu,M}(F) < \epsilon$ for all μ. Suppose M has $\|M\| \le m_\epsilon$. Then $\|MZ\| \le m_\epsilon$ a.s. If μ has $\|\mu\| > m_\epsilon + p$, then $\|MZ + \mu\| > p$ a.s. and

$$q_{\mu,M}(F) \le \text{Prob}\left\{MZ + \mu \in pB\right\} = \text{Prob}\left\{\|MZ+\mu\| \le p\right\} = 0.$$

So we have proved that

$$\{(\mu, M) : q_{\mu,M}(F) \ge \epsilon\} \subset \{(\mu, M) : \|M\| \le m_\epsilon,\ \|\mu\| \le m_\epsilon + p\}$$

which is compact. This completes the proof for the case when Z is uniform on the unit ball.

Now suppose that Z is uniformly distributed on some ball $rB + z$ for $r > 0$. Then $Z' := (Z - z)/r$ is uniformly distributed on B and we have

$$\begin{aligned}
&\sup_{\mu\in\mathbb{R}^d; M\in\mathbb{R}^{d\times d}:\|M\|=m} q_{\mu,M}(F) = \sup_{\mu;M:\|M\|=m} \text{Prob}\left\{MZ+\mu \in F\right\} \\
&\quad= \sup_{\mu;M:\|M\|=m} \text{Prob}\left\{rMZ' + Mz + \mu \in F\right\} \\
&\quad= \sup_{\mu;\|M\|=rm} \text{Prob}\left\{MZ' + \mu \in F\right\} \downarrow 0 \quad \text{as } m \uparrow \infty \qquad\qquad \text{(A.10)}
\end{aligned}$$

from (A.9). Again we can choose m_ϵ large enough so that $\|M\| > m_\epsilon$ implies $q_{\mu,M}(F) < \epsilon$ for all μ. If $\|M\| \le m_\epsilon$ but $\|\mu\| > m_\epsilon(r + z) + p$, then $\|MZ + \mu\| \ge \|\mu\| - \|MZ\| > m_\epsilon(r + z) + p - m_\epsilon\|Z\| \overset{\text{a.s.}}{\ge} m_\epsilon(r + z) + p - m_\epsilon(r + z) = p$. So just as before we see that the proposition holds.

Now suppose that Z has a density f_Z that is bounded with compact support. Let Z' be a random variable that is uniformly distributed on a ball that contains the support of Z and let $f_{Z'}$ be its density. Since f_Z is bounded we can choose $k > 0$ large enough that $f_Z \le kf_{Z'}$. So for any set $E \subset \mathbb{R}^d$, we have $\text{Prob}\{Z \in E\} \le k\,\text{Prob}\{Z' \in E\}$. In particular,

$$\begin{aligned}
&\text{Prob}\left\{MZ + \mu \in F\right\} = \text{Prob}\left\{Z \in \{z : Mz + \mu \in F\}\right\} \\
&\quad\le k\,\text{Prob}\left\{Z' \in \{z : Mz + \mu \in F\}\right\} = k\,\text{Prob}\left\{MZ' + \mu \in F\right\}.
\end{aligned}$$

Applying the proposition to Z' and ϵ/k gives the proposition for Z.

Finally, suppose that Z has a probability distribution q that is absolutely continuous w.r.t. d-dimensional Lebesgue measure. It has a density f_Z. We can choose N large enough that $q(A) < \epsilon/3$, where $A := \{z : f_Z(z) > N\}$, and we can choose N' large enough that $q(A') < \epsilon/3$, where $A' := N'B$. We have

$$\text{Prob}\{MZ + \mu \in F\} \le \text{Prob}\{MZ + \mu \in F | Z \notin A \cup A'\} + 2\epsilon/3.$$

Now the conditional density of Z given that $Z \notin A \cup A'$ is bounded with compact support. Applying the proposition to this conditional random variable and with $\epsilon/3$ gives the proposition for Z and completes the proof.

A.4.6 Proof of Proposition 4.7

We assume everything from Section 3. (4.9) and (4.12) imply (4.1) [9][Theorem 2.1]. Fix $\Delta > 0$ and $K \subset S$ so that $P(K) > D/(D + \Delta)$ and so that (4.13) holds for each $\epsilon > 0$. We will prove the following: for every finite M, there exists $\epsilon > 0$ such that

$$\text{Prob}\left\{\sup_{\lambda \le 0}\left[\liminf_{n\to\infty}\inf_{\theta\in A_\epsilon^c}\left[\lambda D - \frac{1}{n}\sum_{k=1}^{n}\log E_\theta e^{\lambda\rho(X_k,Y_1)}\right]\right] > M\right\} = 1. \tag{A.11}$$

First, we show how (A.11) gives (4.3).

The stationarity and mixing properties (4.9) of Q_θ show that

$$\frac{1}{n}\log E_\theta e^{\lambda n\rho_n(X_1^n,Y_1^n)} \le \frac{1}{n}\sum_{k=1}^{n}\log E_\theta e^{\lambda\rho(X_k,Y_1)} + \log C,$$

where $1 \le C < \infty$ does not depend on θ. (A.11) then implies the following: for every finite M, there exists $\epsilon > 0$ such that

$$\text{Prob}\left\{\sup_{\lambda\le 0}\left[\liminf_{n\to\infty}\inf_{\theta\in A_\epsilon^c}\left[\lambda D - \frac{1}{n}\log E_\theta e^{\lambda n\rho_n(X_1^n,Y_1^n)}\right]\right] > M\right\} = 1.$$

This gives

$$\text{Prob}\left\{\liminf_{n\to\infty}\inf_{\theta\in A_\epsilon^c}\sup_{\lambda\le 0}\left[\lambda D - \frac{1}{n}\log E_\theta e^{\lambda n\rho_n(X_1^n,Y_1^n)}\right] > M\right\} = 1.$$

And Chebyshev's inequality gives

$$\text{Prob}\left\{\liminf_{n\to\infty}\inf_{\theta\in A_\epsilon^c} L_n(\theta, X_1^n) > M\right\} = 1.$$

Choosing $\epsilon > 0$ corresponding to some $M > \Lambda_\infty^*(\Theta)$, which is finite by assumption, and using (4.2) shows that no sequence of lossy MLEs can be in A_ϵ^c infinitely often with positive probability. So every sequence of lossy MLEs is contained in A_ϵ eventually with probability one. Since A_ϵ has compact closure, (4.3) holds. Corollary 4.1 gives the rest of the results.

Now we will prove (A.11). Define

$$\tilde\rho(x, y) := \begin{cases} D + \Delta & \text{if } x \in K \text{ and } y \in B(K, D + \Delta)^c, \\ 0 & \text{otherwise.}\end{cases}$$

For $\lambda \leq 0$ and $\theta \in A_\epsilon^c$

$$\begin{aligned}\log E_\theta e^{\lambda\rho(x,Y_1)} &\leq \log E_\theta e^{\lambda\tilde{\rho}(x,Y_1)} \\ &= I_K(x)\log\left[Q_\theta(B(K,D+\Delta)) + Q_\theta(B(K,D+\Delta)^c)e^{\lambda(D+\Delta)}\right] \\ &\leq I_K(x)\log\left[\epsilon + (1-\epsilon)e^{\lambda(D+\Delta)}\right].\end{aligned}$$

So the ergodic theorem gives

$$\begin{aligned}\liminf_{n\to\infty}\inf_{\theta\in A_\epsilon^c}&\left[\lambda D - \frac{1}{n}\sum_{k=1}^n \log E_\theta e^{\lambda\rho(X_k,Y_1)}\right] \\ &\geq \lambda D - \limsup_{n\to\infty}\frac{1}{n}\sum_{k=1}^n I_K(X_k)\log\left[\epsilon+(1-\epsilon)e^{\lambda(D+\Delta)}\right] \\ &\overset{\text{a.s.}}{=} \lambda D - \underbrace{P(K)\log\left[\epsilon+(1-\epsilon)e^{\lambda(D+\Delta)}\right]}_{\tilde{\Lambda}_\epsilon(\lambda)},\end{aligned}$$

where $\tilde{\Lambda}_\epsilon(\lambda)$ is defined as indicated. Defining

$$\tilde{\Lambda}_\epsilon^*(D) := \sup_{\lambda\leq 0}\left[\lambda D - \tilde{\Lambda}_\epsilon(\lambda)\right]$$

and taking the supremum over (rational) $\lambda \leq 0$ gives

$$\text{Prob}\left\{\sup_{\lambda\leq 0}\left[\liminf_{n\to\infty}\inf_{\theta\in A_\epsilon^c}\left[\lambda D - \frac{1}{n}\sum_{k=1}^n \log E_\theta e^{\lambda\rho(X_k,Y_1)}\right]\right] \geq \tilde{\Lambda}_\epsilon^*(D)\right\} = 1.$$

The reason we can restrict the supremum to rational λ is that both sides are concave [10]. (A.11) will be true if we can show that $\tilde{\Lambda}_\epsilon^*(D) \to \infty$ as $\epsilon \downarrow 0$.

Let λ^* satisfy

$$\frac{d}{d\lambda}\tilde{\Lambda}_\epsilon(\lambda^*) = D.$$

Some calculus shows that

$$\lambda^* = \frac{1}{D+\Delta}\log\frac{\alpha\epsilon}{(1-\alpha)(1-\epsilon)}, \quad \text{where} \quad \alpha := \frac{D}{D+\Delta}\frac{1}{P(K)} < 1.$$

Substitution and some algebra show that

$$\tilde{\Lambda}_\epsilon^*(D) = \lambda^* D - \tilde{\Lambda}_\epsilon(\lambda^*) = P(K)(\alpha-1)\log\epsilon + O(\epsilon) \to \infty \quad \text{as } \epsilon\downarrow 0.$$

Acknowledgments

I want to thank I. Kontoyiannis and M. Madiman for many useful comments. I. Kontoyiannis suggested the problem to me. The idea of lossy maximum likelihood estimation is his. A. Amarasingham, A. Hoffman and X. Zhang gave useful suggestions for the proofs.

References

[1] H. Attouch. *Variational Convergence for Functions and Operators.* Pitman, Boston, 1984.

[2] Hedy Attouch and Roger J-B Wets. Epigraphical analysis. In H. Attouch, J-P Aubin, F. Clarke, and I. Ekeland, editors, *Analyse Non Linèaire*, Annales de l'Institut Henri Poincaré, pages 73–100. Gauthier-Villars, Paris, 1989.

[3] Andrew R. Barron. *Logically Smooth Density Estimation.* PhD thesis, Standford University, Department of Electrical Engineering, 1985.

[4] Patrick Billingsley. *Convergence of Probability Measures.* Wiley, New York, second edition, 1999.

[5] Zhiyi Chi. The first-order asymptotic of waiting times with distortion between stationary processes. *IEEE Transactions on Information Theory*, 47(1):338–347, January 2001.

[6] I. Csiszár. On an extremum problem of information theory. *Studia Scientiarum Mathematicarum Hungarica*, 9:57–71, 1974.

[7] Amir Dembo and Ioannis Kontoyiannis. Source coding, large deviations, and approximate pattern matching. *IEEE Transactions on Information Theory*, 48(6):1590–1615, June 2002.

[8] Robert M. Gray. *Probability, Random Processes, and Ergodic Properties.* Springer-Verlag, New York, 1988.

[9] Matthew Harrison. Epi-convergence of lossy likelihoods. APPTS #03-4, Brown University, Division of Applied Mathematics, Providence, RI, April 2003.

[10] Matthew Harrison. The first order asymptotics of waiting times between stationary processes under nonstandard conditions. APPTS #03-3, Brown University, Division of Applied Mathematics, Providence, RI, April 2003.

[11] Matthew Harrison and Ioannis Kontoyiannis. Maximum likelihood estimation for lossy data compression. In *Proceedings of the Fortieth Annual Allerton Conference on Communication, Control and Computing*, pages 596–604, Allerton, IL, October 2002.

[12] Peter J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, Statistics*, pages 221–233, Berkeley and Los Angeles, 1967. University of California Press.

[13] Olav Kallenberg. *Foundations of Modern Probability.* Springer, New York, second edition, 2002.

[14] I. Kontoyiannis. Model selection via rate-disortion theory. In *34th Annual Conference on Information Sciences and Systems*, Princeton, NJ, March 2000.

[15] Ioannis Kontoyiannis and Junshan Zhang. Arbitrary source models and Bayesian codebooks in rate-distortion theory. *IEEE Transactions on Information Theory*, 48(8):2276–2290, August 2002.

[16] J. Pfanzagl. On the measurability and consistency of minimum contrast estimates. *Metrika*, 14:249–272, 1969.

[17] H. L. Royden. *Real Analysis*. Prentice Hall, Englewood Cliffs, NJ, third edition, 1988.

[18] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, third edition, 1976.

[19] Gabriella Salinetti. Consistency of statistical estimators: the epigraphical view. In S. Uryasev and P. M. Pardalos, editors, *Stochastic Optimization: Algorithms and Applications*, pages 365–383. Kluwer Academic Publishers, Dordrecht, 2001.

[20] A. N. Shiryaev. *Probability*. Springer, New York, second edition, 1996.

[21] En-hui Yang and Zhen Zhang. On the redundancy of lossy source coding with abstract alphabets. *IEEE Transactions on Information Theory*, 45(4):1092–1110, May 1999.