Epi-Convergence of Lossy Likelihoods

Matthew Harrison

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

April 2, 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 epi-convergence of

$$-\frac{1}{n}\log Q_\theta(B(X_1^n, D)).$$

Epi-convergence is useful for studying the asymptotic behavior of minima and minimizers.

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

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

We are often interested in the asymptotic behavior of maximizers (over Θ) of the likelihood (maximum likelihood estimators), or equivalently, of minimizers of

$$-\frac{1}{n}\log P_\theta(X_1^n).$$

When written in this form, we notice that the (per symbol) negative log-likelihood is exactly the (per symbol) ideal Shannon code length for the data X_1^n and the source P_θ. This is a fundamental quantity in information theory, particularly, *lossless* data compression. The corresponding quantity in *lossy* data compression is

$$-\frac{1}{n}\log Q_\theta(B(X_1^n, D)), \tag{1.1}$$

where $B(X_1^n, D)$ is the distortion ball around X_1^n of radius D and $\{Q_\theta\}_{\theta\in\Theta}$ are probability measures on the reproduction sequence space [11]. 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 (1.1). See Harrison and Kontoyiannis (2002) [9] and Kontoyiannis (2000) [10] for a more detailed discussion of the motivations and possible applications.

A useful tool for studying the asymptotics of minimums and minimizers is epi-convergence [1, 2, 13]. Let (Θ, ν) be a metric space and let $O(\theta, \epsilon)$ be the ϵ-neighborhood around θ. Given a sequence of functions $f_n : \Theta \to [-\infty, \infty]$, $n \geq 1$, we define

$$\begin{aligned}
\operatorname*{epi\text{-}lim\,inf}_{n\to\infty} f_n(\theta) &:= \lim_{\epsilon\downarrow 0} \liminf_{n\to\infty} \inf_{\theta'\in O(\theta,\epsilon)} f_n(\theta'), \\
\operatorname*{epi\text{-}lim\,sup}_{n\to\infty} f_n(\theta) &:= \lim_{\epsilon\downarrow 0} \limsup_{n\to\infty} \inf_{\theta'\in O(\theta,\epsilon)} f_n(\theta').
\end{aligned}$$

Notice that the limits as $\epsilon \downarrow 0$ are actually suprema over $\epsilon > 0$, so everything is well-defined. Also notice that it is more precise to write

$$\left(\operatorname*{epi\text{-}lim\,inf}_{n\to\infty} f_n\right)(\theta)$$

and similarly for epi-lim sup, because epi-$\liminf_n f_n$ is a limiting function which we evaluate at θ, instead of the limit of the sequence of numbers $f_n(\theta)$, $n \geq 1$. The functions epi-$\liminf_n f_n$ and epi-$\limsup_n f_n$ are lower semicontinuous (l.sc.) and we always have epi-$\liminf_n f_n(\theta) \leq$ epi-$\limsup_n f_n(\theta)$. If the reverse inequality is also true then we can define epi-$\lim_n f_n(\theta)$ as the common value. When we are in the situation that the epi-limit exists for each $\theta \in \Theta$, we can talk about the function epi-$\lim_n f_n$, which is l.sc.

Epi-convergence of the functions $(f_n)_{n\geq 1}$ to a function f is neither stronger nor weaker than pointwise convergence. It is not hard to see that epi-$\liminf_n f_n \geq f$ implies $\liminf_n f_n \geq f$ and that $\limsup_n f_n \leq f$ implies epi-$\limsup_n f_n \leq f$. Counter examples to each of the reverse implications are easy to construct. In this paper we are interested in the epi-convergence of (1.1) almost surely (a.s.), meaning that the collection of x_1^∞ for which the sequence of functions

$$\theta \mapsto -\frac{1}{n}\log Q_\theta(B(x_1^n, D)), \quad n \geq 1,$$

epi-converges has probability one. The pointwise convergence (a.s.) of (1.1) to a deterministic limit has been established in the literature under a variety of conditions [4, 5, 8]. Our work here will focus on the epi-lim inf lower bound and we will appeal to the pointwise convergence results in the literature for the epi-lim sup upper bound. The main result that we need is the following:

Proposition 1.1. *Let (Θ, ν) be a metric space and let $(f_n)_{n\geq 1}$ and g be extended real-valued functions on Θ. Define $f := \limsup_n f_n$. If $\{\theta : f(\theta) < r\}$ is dense in $\{\theta : g(\theta) < r\}$ for each $r \in \mathbb{R}$, then* epi-$\limsup_n f_n \leq g$.

The proof is only a few lines and is given in the Appendix.

In the next section we give the main results. We will always assume that the source $(X_n)_{n\geq 1}$ is stationary and ergodic. The reproduction distributions $\{Q_\theta\}_{\theta\in\Theta}$ are assumed to be stationary and to satisfy certain strong mixing assumptions. We also assume that

the distributional properties and the rates of mixing of the Q_θ are well behaved (locally) as θ varies over Θ. We only consider average single-letter distortion, that is

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

for some function ρ, called the single-letter distortion function. We can get the epi-lim inf lower bound using only one-sided mixing assumptions on the Q_θ, so these are the only assumptions that we mention. To get the epi-lim sup upper bound, we assume that the pointwise limits exist for many Q_θ in the spirit of Proposition 1.1. Using the pointwise results in the literature (at the time that this was written) will require additional two-sided mixing assumptions and perhaps further restrictions on ρ, such as boundedness.

After the main results, we consider the special case where Θ is convex and each Q_θ is independent and identically distributed (i.i.d.), or memoryless. In this case, we are able to obtain a complete characterization of the asymptotic behavior of the epi-limit, including necessary and sufficient conditions for epi-convergence (a.s.) analogous to the pointwise results of Harrison (2003) [8].

2 Main Results

We begin with the setup used throughout the remainder of the paper. $(S, \mathcal{S})$ and $(T, \mathcal{T})$ are standard measurable spaces.[1] $(X_n)_{n\ge1}$ is a stationary and ergodic random process on $(S^{\mathbb{N}}, \mathcal{S}^{\mathbb{N}})$ with distribution P which is assumed to be complete (that is, all subsets of sets with probability 0 are measurable). $\rho : S \times T \to [0, \infty)$ is an $\mathcal{S} \times \mathcal{T}$-measurable function ($\mathcal{S} \times \mathcal{T}$ denotes the smallest product σ-algebra).

Let Θ be a separable metric space with metric ν and let $O(\theta, \epsilon) := \{\theta' \in \Theta : \nu(\theta, \theta') < \epsilon\}$ denote the ϵ-neighborhood of θ. To each $\theta \in \Theta$ we associate a stationary probability measure Q_θ on $(T^{\mathbb{N}}, \mathcal{T}^{\mathbb{N}})$. We use $(Y_n)_{n\ge1}$ to denote a stationary random 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

$$\theta_m \to \theta \text{ implies } Q_{\theta_m,n} \xrightarrow{\tau} Q_{\theta,n} \text{ as } m \to \infty \text{ for each } n, \tag{2.1}$$

or

$$(T, \mathcal{T}) \text{ is a separable metric space with its Borel } \sigma\text{-algebra}, \tag{2.2a}$$

$$\rho(x, \cdot) \text{ is continuous for each } x \in S, \tag{2.2b}$$

$$\theta_m \to \theta \text{ implies } Q_{\theta_m,n} \xrightarrow{w} Q_{\theta,n} \text{ as } m \to \infty \text{ for each } n. \tag{2.2c}$$

τ-Convergence is setwise convergence of probability measures.[2] w-Convergence is weak convergence of probability measures.[3] When T is finite, assumptions (2.1) and (2.2) are

[1] Standard measurable spaces include Polish spaces and let us avoid uninteresting pathologies while working with random sequences [7].

[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 $Q(\partial A) = 0$.

equivalent. When each Q_θ is i.i.d., then (2.1) and (2.2c) will hold whenever they hold for $n = 1$.

We also need to control the mixing properties of the Q_θ. In particular, we assume that each Q_θ is ψ_+-mixing, that is, there exists finite $C(\theta) \geq 1$ and $d(\theta) \geq 1$ such that

$$Q_\theta(A \cap B) \leq C(\theta) Q_\theta(A) Q_\theta(B) \tag{2.3}$$

for all $A \in \sigma(Y_1^n)$ and $B \in \sigma(Y_{n+d(\theta)}^\infty)$ and any n (c.f. Chi, 2001 [4]). This includes all i.i.d. processes ($C, d \equiv 1$) and all finite state Markov chains. Finally, we need to relate these mixing constants to the topology on Θ. We assume that C and d are both locally, uniformly bounded, that is, for each $\theta \in \Theta$

$$\inf_{\epsilon>0} \sup_{\theta' \in O(\theta,\epsilon)} d(\theta') < \infty \quad \text{and} \quad \inf_{\epsilon>0} \sup_{\theta' \in O(\theta,\epsilon)} C(\theta') < \infty. \tag{2.4}$$

For the case when each Q_θ is i.i.d., this condition is valid because we can take $C, d \equiv 1$.

Fix $D \in \mathbb{R}$. We define the following standard quantities:

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

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

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

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

where log denotes the natural logarithm $\log_e$. Many properties of these quantities can be found in the literature (see, for example, the Appendix of Harrison, 2003 [8]). When $\lambda \leq 0$, then we actually have

$$\Lambda_\infty(\theta, \lambda) = \lim_{n \to \infty} \Lambda_n(\theta, \lambda). \tag{2.5}$$

When Q_θ is i.i.d., then $\Lambda_n(\theta, \cdot)$ and $\Lambda_n^*(\theta)$ do not depend on n. An important property is that Λ_n^* is lower semicontinuous (l.sc.) on Θ for all $1 \leq n \leq \infty$.

Several recent papers [4, 5, 8] give conditions for which

$$\lim_{n \to \infty} L_n(\theta, X_1^n) \overset{\text{a.s.}}{=} \Lambda_\infty^*(\theta), \tag{2.6}$$

which Dembo and Kontoyiannis (2002) [5] call the *generalized AEP (Asymptotic Equipartition Property)*. When Q_θ satisfies a strong two-sided mixing condition, Harrison (2003) [8] provides necessary and sufficient conditions for this property (see Section 2.1 below). Here we are interested in proving an analogous result for epi-convergence, which is neither stronger nor weaker than pointwise convergence. Defining

$$\Theta_{\lim} := \left\{ \theta \in \Theta : \limsup_{n \to \infty} L_n(\theta, X_1^n) \overset{\text{a.s.}}{\leq} \Lambda_\infty^*(\theta) \right\},$$

$$\Theta_r := \left\{ \theta \in \Theta : \Lambda_\infty^*(\theta) < r \right\}, \qquad r \leq \infty,$$

we are ready for our main result.

Theorem 2.1. *Λ^*_∞ is lower semicontinuous on Θ and*

$$\text{Prob}\left\{\operatorname*{epi-lim\,inf}_{n\to\infty} L_n(\theta, X_1^n) \geq \Lambda^*_\infty(\theta), \quad \forall\theta\in\Theta\right\} = 1. \tag{2.7}$$

Suppose that for each $r < \infty$, $\Theta_{\lim} \cap \Theta_r$ is dense in Θ_r. Then

$$\text{Prob}\left\{\operatorname*{epi-lim}_{n\to\infty} L_n(\theta, X_1^n) = \Lambda^*_\infty(\theta), \quad \forall\theta\in\Theta\right\} = 1. \tag{2.8}$$

The proof of Theorem 2.1 is relatively straightforward, although ensuring that everything is measurable and moving the $\forall\theta\in\Theta$ inside the probability requires a little care. The epi-lim inf lower bound, although stronger than the pointwise lower bound, follows from Chebyshev's inequality in the same manner that the lower bound of the pointwise limit (2.6) is typically established. Notice that (2.7) implies that $\liminf_n L_n(\theta, X_1^n) \overset{\text{a.s.}}{\geq} \Lambda^*_\infty(\theta)$, so we see that $\Theta_{\lim}$ is actually

$$\Theta_{\lim} = \left\{\theta\in\Theta : \lim_{n\to\infty} L_n(\theta, X_1^n) \overset{\text{a.s.}}{=} \Lambda^*_\infty(\theta)\right\}.$$

The epi-lim sup upper bound is weaker than the pointwise upper bound and we essentially get it for free from Proposition 1.1 by assuming that pointwise limits exist all over the parameter space, i.e., that $\Theta_{\lim} \cap \Theta_r$ is dense in Θ_r for each r. In some ways the result would be more general if we just assumed that the epi-lim sup behaved appropriately, since we do not use anything about the pointwise limits in deriving the epi-lim inf lower bound. However, epi-limits are much less familiar objects than pointwise limits and results about pointwise limits can be found in the literature. Chi (2001) [4], Dembo and Kontoyiannis (2002) [5] and Harrison (2003) [8] each contain further conditions that will allow us to establish pointwise limits for most parameter values and then infer (2.8). Before turning to examples, we consider a special case.

2.1 Convex, memoryless families

In this section only, we consider the special case when Θ is convex and each Q_θ is i.i.d. Convexity arises frequently in nonparametric settings and in mixture models. The nice thing about this setting is that we can completely characterize the behavior of the epi-limit in Theorem 2.1 in exactly the same manner that the pointwise limit was characterized in Harrison (2003)[8][Theorem 2.1].

We still need all of the assumptions detailed in the previous section, and we add some more. First, we assume that each Q_θ is i.i.d. This simplifies several of the assumptions from the previous section. (2.3) and (2.4) are trivially true (take $C(\theta) = d(\theta) = 1$). (2.1) and (2.2c) will be true whenever they are true for $n = 1$. Also, Λ_n and Λ^*_n do not depend on n. This turns out to be crucial here because Λ^*_1 is convex (under the following convexity assumptions), which means Λ^*_∞ is convex.

Second, we assume that Θ is convex set with the property that

$$\lambda\theta' + (1-\lambda)\theta \to \theta \text{ as } \lambda \downarrow 0, \text{ and} \tag{2.9a}$$

$$Q_{\lambda\theta'+(1-\lambda)\theta,1} = \lambda Q_{\theta',1} + (1-\lambda)Q_{\theta,1} \tag{2.9b}$$

for each $\theta, \theta' \in \Theta$ and $0 \leq \lambda \leq 1$. (2.9a) just says that the topology on Θ is continuous w.r.t. convex combinations. (2.9b) says that the convexity properties on Θ extend to the

first marginals of the Q_θ. The reason that we do not use the easier to state assumption that Θ is a convex set of probability measures on T and each Q_θ is i.i.d. with distribution θ is that the latter rules out the common situation (in mixture models, for example) where multiple θ correspond to the same probability measure.

Define

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

Here are the pointwise results (which are for a specific θ and do not need any of the assumptions on Θ):

Theorem 2.2. *[8][Theorem 2.1] Fix θ. We have*

$$\begin{aligned} &\operatorname{Prob}\{L_n(\theta, X_1^n) = \infty \text{ eventually}\} = 1 \quad \text{if } D_{\min}(\theta) > D\ , \\ &\operatorname{Prob}\{L_n(\theta, X_1^n) < \infty \text{ eventually}\} = 1 \quad \text{if } D_{\min}(\theta) < D\ . \end{aligned} \tag{2.10}$$

If either $D_{\min}(\theta) \neq D$ or $\Lambda_\infty^*(\theta) = \infty$ or $m(\theta, X_1)$ is a.s. constant, then

$$\lim_{n\to\infty} L_n(\theta, X_1^n) \stackrel{a.s.}{=} \Lambda_\infty^*(\theta). \tag{2.11}$$

Otherwise, $0 < D = D_{\min}(\theta) < \infty$, and

$$\operatorname{Prob}\{L_n(\theta, X_1^n) = \infty \text{ infinitely often}\} > 0, \tag{2.12a}$$

$$\operatorname{Prob}\{L_n(\theta, X_1^n) < \infty \text{ infinitely often}\} = 1, \tag{2.12b}$$

$$\lim_{m\to\infty} L_{n_m}(\theta, X_1^{n_m}) \stackrel{a.s.}{=} \Lambda_\infty^*(\theta) < \infty, \tag{2.12c}$$

where $(n_m)_{m\geq 1}$ is the (a.s.) infinite subsequence of $(n)_{n\geq 1}$ for which $L_n(\theta, X_1^n)$ is finite, or (a.s.) equivalently, the subsequence where $\sum_{k=1}^n m(\theta, X_k) \leq nD$.

The epi-limit results that we prove here are quite similar. Define

$$m(\Theta, x) := \inf_{\theta\in\Theta} m(\theta, x), \qquad D_{\min}(\Theta) := Em(\Theta, X_1), \qquad \Lambda_\infty^*(\Theta) := \inf_{\theta\in\Theta} \Lambda_\infty^*(\theta).$$

Theorem 2.3. *We have*

$$\begin{aligned} &\operatorname{Prob}\left\{\inf_{\theta\in\Theta} L_n(\theta, X_1^n) = \infty \text{ eventually}\right\} = 1 \quad \text{if } D_{\min}(\Theta) > D\ , \\ &\operatorname{Prob}\left\{\inf_{\theta\in\Theta} L_n(\theta, X_1^n) < \infty \text{ eventually}\right\} = 1 \quad \text{if } D_{\min}(\Theta) < D\ . \end{aligned} \tag{2.13}$$

If either $D_{\min}(\Theta) \neq D$ or $\Lambda_\infty^*(\Theta) = \infty$ or $m(\Theta, X_1)$ is a.s. constant, then

$$\operatorname{Prob}\left\{\operatorname*{epi-lim}_{n\to\infty} L_n(\theta, X_1^n) = \Lambda_\infty^*(\theta), \quad \forall\theta\in\Theta\right\} = 1. \tag{2.14}$$

Otherwise $0 < D = D_{\min}(\Theta) = \min_{\theta\in\Theta} D_{\min}(\theta) < \infty$, $\Theta_\infty \neq \emptyset$ whereas $\Theta_{\lim} \cap \Theta_\infty = \emptyset$, and

$$\operatorname{Prob}\left\{\inf_{\theta\in\Theta} L_n(\theta, X_1^n) = \infty \text{ infinitely often}\right\} > 0 \tag{2.15a}$$

$$\operatorname{Prob}\left\{\inf_{\theta\in\Theta} L_n(\theta, X_1^n) < \infty \text{ infinitely often}\right\} = 1 \tag{2.15b}$$

$$\operatorname{Prob}\left\{\operatorname*{epi-lim}_{m\to\infty} L_{n_m}(\theta, X_1^{n_m}) = \Lambda_\infty^*(\theta), \quad \forall\theta\in\Theta\right\} = 1, \tag{2.15c}$$

where $(n_m)_{m\geq 1}$ is the (a.s.) infinite subsequence of $(n)_{n\geq 1}$ for which $\inf_{\theta\in\Theta} L_n(\theta, X_1^n)$ is finite, or (a.s.) equivalently, the subsequence where $\sum_{k=1}^n m(\Theta, X_k) \leq nD$.

Proposition 2.4. *If* $\inf_{\theta\in\Theta} D_{\min}(\theta) < \infty$, *then* $\inf_{\theta\in\Theta} D_{\min}(\theta) = D_{\min}(\Theta)$.

Notice that (2.15) implies that the epi-limit in (2.14) does not exist for some θ with positive probability, so the conditions for (2.14) are both necessary and sufficient. Because of the conclusion that $\Theta_{\lim} \cap \Theta_\infty = \emptyset$ when (2.14) does not hold, we see that a sufficient condition for (2.14) is the conceptually simple requirement that the generalized AEP (2.11) holds with a finite limit for at least one point θ in the parameter space. The reason that we can extend a pointwise limit at a single point to an epi-limit over the entire space is the convexity assumption, which lets us infer properties of one parameter value from another. Another easy condition to verify for (2.14) is $m(\Theta, X_1)$ a.s. constant, since in many examples $m(\Theta, \cdot) \equiv 0$.

The proof of Theorem 2.3 is somewhat complicated because it has several parts. We use the strikingly similar pointwise results and the convexity assumption both to derive the conditions for Theorem 2.1 when (2.14) holds and to derive the pathological behavior when (2.15) holds. The pointwise results can be extended verbatim to the case where Q_θ satisfies certain strong two-sided mixing conditions [8]. For the epi-convergence results, however, it is not exactly clear how one should extend the convexity assumption to include cases when Q_θ is not a product measure. The main problem is that Λ^*_∞ might not be convex.

2.2 Examples

In this section we go through several examples that are covered by Theorems 2.1 or 2.3. 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.[4] The rest of the assumptions that are needed for the theorems are addressed on a case by case basis.

2.2.1 Example: memoryless families

Suppose that (Θ, ν) is a separable metric space, that each Q_θ is i.i.d. (memoryless) and that either (2.1) or (2.2) holds. Assumptions (2.3) and (2.4) are always true for i.i.d. families and we can apply Theorem 2.1 to get (2.7). Theorem 2.2 essentially describes $\Theta_{\lim}$. To verify that $\Theta_{\lim} \cap \Theta_r$ is dense in Θ_r for a particular example will require an investigation of $m(\theta, X_1)$ and perhaps an additional constraint on D to avoid any pathologies. A common assumption is that $D_{\min}(\theta) < D$ for all $\theta \in \Theta$. In this case, we see that $\Theta_{\lim} = \Theta$ and we can conclude (2.8). Another situation that is rare in practice but that comes up in theory (when thinking about lossless data compression as a special case of lossy data compression) is when $D = 0$. Again, we see that $\Theta_{\lim} = \Theta$ and we can conclude (2.8).

We are often in the situation where the statistics of the source $(X_n)_{n\geq 1}$ are unknown. In this case it might be difficult to verify that every point of Θ satisfies the generalized AEP (that is, $\Theta = \Theta_{\lim}$). In Example 2.2.2 we describe a common situation where we always have (2.8) (even though $\Theta_{\lim} \neq \Theta$), regardless of the source statistics. The proof takes advantage of the special structure of the example to calculate Λ^*_∞ explicitly and show that $\Theta_{\lim} \cap \Theta_r$ is dense in Θ_r.

[4] 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.

Another way to show that (2.8) holds regardless of the source statistics is to use the methods of Section 2.1. Examples 2.2.3–2.2.5 are convex families. We can use Theorem 2.3 to show that (2.8) holds regardless of the source statistics.

2.2.2 Example: memoryless Gaussian families, squared-error distortion

Take $S = T := \mathbb{R}$ and $\rho(x, y) := |x-y|^2$ to be squared-error distortion. Let $\Theta := M \times \Sigma := \mathbb{R} \times [0, \infty)$ 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. Any metric ν equivalent to the Euclidean metric on Θ will give (2.2c), so (2.2) holds and we are in the situation described by Example 2.2.1. We are about to show that $\Theta_{\lim} \cap \Theta_r$ is dense in Θ_r for each r, so we can also conclude (2.8).

First of all, if $\sigma > 0$, then $m((\mu, \sigma), x) = 0$ and Theorem 2.2 shows that $(\mu, \sigma) \in \Theta_{\lim}$. If $D_{\min}((\mu, 0)) \neq D$ or $D = 0$, then $(\mu, 0) \in \Theta_{\lim}$ as well. So we need only analyze the situation where $0 < D = D_{\min}((\mu, 0)) < \infty$. We have $m((\mu, 0), x) = (x - \mu)^2$ and $D_{\min}((\mu, 0)) = E[(X_1 - \mu)^2]$. When $D = D_{\min}((\mu, 0)) < \infty$ we see that X_1 has finite variance $v^2 \leq D$.

If $v^2 < D$, then in any neighborhood of a μ with $D_{\min}((\mu, 0)) = D$, there are μ' with $D_{\text{ave}}((\mu', 0)) := E_P E_{(\mu',0)}[(X_1 - Y_1)^2] < D$. This implies that $(\mu', 0) \in \Theta_{\lim}$ and $\Lambda^*_\infty((\mu', 0)) = 0$ [8]. So in this case, any problem point $(\mu, 0) \notin \Theta_{\lim}$ is a limit point of $\Theta_{\lim} \cap \Theta_r$ for any $r > 0$.

If $0 < D_{\min}((\mu, 0)) = D = v^2 < \infty$, then we see that $\mu = EX_1$. We have [5]

$$\Lambda^*_\infty((EX_1, \sigma)) = \frac{1}{2} \log \frac{a}{D} - \frac{(a - D)^2}{2a\sigma^2}, \qquad a := \frac{1}{2}\left(\sigma^2 + \sqrt{\sigma^4 + 4D^2}\right), \qquad \sigma > 0.$$

Letting $\sigma \downarrow 0$ shows that $\Lambda^*_\infty((EX_1, \sigma)) \to 0$ as $\sigma \downarrow 0$. Since $(EX_1, \sigma) \in \Theta_{\lim}$ for each $\sigma > 0$, we have shown that the problem point $(EX_1, 0)$ is a limit point of $\Theta_{\lim} \cap \Theta_r$ for any $r > 0$. In both cases $\Theta_{\lim} \cap \Theta_r$ is dense in Θ_r for each r. (Note that Θ_r is always empty for $r \leq 0$, so there is nothing to verify.)

2.2.3 Example: memoryless, nonparametric, weak topology

Let $(T, \mathcal{T})$ be a separable metric space with its Borel σ-algebra. Let Θ be the set of all probability measures on $(T, \mathcal{T})$ with a metric ν that metrizes weak convergence of probability measures, for example, the Prohorov metric. Take Q_θ to be i.i.d. θ and suppose that $\rho(x, \cdot)$ is continuous for each $x \in S$. Billingsley (1999) [3] shows that Θ is separable and that (2.2c) holds, so we are in situation (2.2). It is easy to see that (2.9) is also valid.

In this case, we can apply Theorem 2.3. Suppose that for every $x \in S$, there exists a $y \in T$ with $\rho(x, y) \leq D$. For each y, Θ contains the distribution that puts unit mass on $\{y\}$, so $m(\Theta, x) \leq D$ for each $x \in S$. This means that either $D_{\min}(\Theta) < D$ or $m(\Theta, X_1) \stackrel{\text{a.s.}}{=} D$, a constant, so (2.14) is valid. Of course, there are many other situations where (2.14) is valid as detailed in Theorem 2.3.

2.2.4 Example: memoryless, discrete alphabet, strong topology

Let T be countable and let Θ be the set of all probability measures on T (or, equivalently, the probability simplex on $\mathbb{R}^{\mathbb{N}}$). Take the metric ν on Θ to be the total variation distance (or, equivalently, the supremum metric on $\mathbb{R}^{\mathbb{N}}$). Then Θ is separable. Defining Q_θ to

be i.i.d. θ makes situations (2.1) and (2.9) valid, so Theorem 2.3 holds. Under the same conditions on ρ and D in the previous example (Section 2.2.3), we can infer (2.14).

If $T = \{1, \ldots, N\}$ is finite, then the same arguments hold. If we are conceptualizing Θ as the probability simplex on $\mathbb{R}^N$, then we can take ν to be the Euclidean metric.

2.2.5 Example: memoryless, discrete mixture proportions

Let $\{q_n\}_{n\geq 1}$ be a countable family of probability measures on $(T, \mathcal{T})$. Let Θ be the probability simplex on $\mathbb{R}^{\mathbb{N}}$ with the supremum metric. Θ is separable. Defining Q_θ to be i.i.d. with first marginal

$$Q_{\theta,1} := \sum_{n\geq 1} \theta_n q_n, \qquad \theta := (\theta_1, \theta_2, \ldots), \quad \sum_{n\geq 1} \theta_n = 1, \quad \theta_n \geq 0,$$

gives (2.1) and (2.9), so Theorem 2.3 holds. There are a variety of conditions that ensure that (2.14) holds. A common situation is that $m(\Theta, x) = 0$ for all $x \in S$. Note that this does not depend on the source statistics.

If $\{q_n\}_{1\leq n\leq N}$ is a finite family and Θ is the probability simplex on $\mathbb{R}^N$ with the Euclidean metric, then the same arguments hold.

2.2.6 Example: finite state, irreducible Markov chains

Let T be a finite set and let $\{Q_\theta\}_{\theta\in\Theta}$ be the class of stationary, first-order, irreducible Markov chains on T. Let Θ be the corresponding set of probability transition matrices,[5] which we can think about as a subset of $\mathbb{R}^{T\times T}$, and let ν be a metric on Θ that is equivalent to the Euclidean metric when Θ is viewed as a subset of $\mathbb{R}^{T\times T}$. We will show that if $E[\min_{y\in T} \rho(X_1, y)] \neq D$ or $D = 0$, then (2.8) is true. In the special case where $S = T$ and $\rho(x, x) = 0$ (such as Hamming distortion), then we get (2.8) without any restriction on the source statistics.

Let $p_\theta := p_\theta(i, j)$, $i, j \in T$, be the transition probability matrix for Q_θ. Each Q_θ has a unique stationary distribution, which we denote $\pi_\theta := \pi_\theta(i)$, $i \in T$. Since p_θ is irreducible, each $\pi_\theta(i) > 0$. It is also not hard to see that we can take $d(\theta) = 1$ and $C(\theta) = 1/\left[\min_{i\in T} \pi_\theta(i)\right] < \infty$ for each θ. This gives (2.3). By assumption $\theta_n \to \theta$ implies $p_{\theta_n}(i, j) \to p_\theta(i, j)$ for each i, j, and it is not hard to see that this implies that $\pi_{\theta_n}(i) \to \pi_\theta(i)$ for each i. From this we get both (2.1) and (2.4). Theorem 2.1 gives (2.7).

Consider the subset $\Theta' \subset \Theta$ of probability transition matrices with all positive elements. For $\theta \in \Theta'$ we can apply the results in Harrison (2003) [8][Theorem 2.1] to see whether or not $\theta \in \Theta_{\text{lim}}$. As long as $E[\min_{y\in T} \rho(X_1, y)] \neq D$ or $D = 0$, we see that $\theta \in \Theta_{\text{lim}}$ and we have $\Theta' \subset \Theta_{\text{lim}}$.

Now consider a point $\theta \notin \Theta'$. We want to construct a point in Θ_{lim} that is (arbitrarily) close to θ with (arbitrarily) similar Λ_∞. This will show that $\Theta_{\text{lim}} \cap \Theta_r$ is dense in Θ_r. Pick any point $\theta' \in \Theta'$. For any $0 \leq \epsilon \leq 1$, consider the point θ_ϵ corresponding to the transition probability matrix $p_{\theta_\epsilon} := (1 - \epsilon)p_\theta + \epsilon p_{\theta'}$. If $\epsilon > 0$, p_{θ_ϵ} has all positive entries because $p_{\theta'}$ does. Also, $p_{\theta_\epsilon} \to p_\theta$ as $\epsilon \downarrow 0$, so $\theta_\epsilon \to \theta$ as well.

[5] Every (time homogenous) finite state, irreducible Markov chain has a unique stationary distribution. The transition probabilities thus determine the whole process, because we assume that each Q_θ is stationary.

Fix $0 < \alpha < \min_{i\in T} \pi_\theta(i)$. We can choose $\delta > 0$ small enough so that $0 \le \epsilon < \delta$ implies $\min_{i\in T} \pi_{\theta_\epsilon}(i) > \alpha$ and $-\log(1-\epsilon) < \alpha$. For $\lambda < 0$ we have

$$\begin{aligned}\Lambda_n(\theta_\epsilon, \lambda) &= \frac{1}{n} E_P \left[\log \sum_{y_1^n \in T^n} e^{\lambda \rho_n(X_1^n, y_1^n)} \pi_{\theta_\epsilon}(y_1) \prod_{k=2}^{n} p_{\theta_\epsilon}(y_{k-1}, y_k) \right] \\ &\ge \frac{1}{n} E_P \left[\log \sum_{y_1^n \in T^n} e^{\lambda \rho_n(X_1^n, y_1^n)} \pi_{\theta}(y_1) \alpha \prod_{k=2}^{n} (1-\epsilon) p_{\theta}(y_{k-1}, y_k) \right] \\ &= \frac{\log \alpha + (n-1)\log(1-\epsilon)}{n} + \Lambda_n(\theta, \lambda).\end{aligned}$$

So

$$\Lambda_\infty(\theta_\epsilon, \lambda) \ge \Lambda_\infty(\theta, \lambda) + \log(1-\epsilon) \ge \Lambda_\infty(\theta, \lambda) - \alpha$$

and we have

$$\Lambda_\infty^*(\theta_\epsilon) \le \Lambda_\infty^*(\theta) + \alpha$$

for each $0 \le \epsilon < \delta$. Since $\theta_\epsilon \in \Theta_{\lim}$ for each $\epsilon > 0$, $\theta_\epsilon \to \theta$ as $\epsilon \downarrow 0$ and α is arbitrary, we see that $\Theta_{\lim} \cap \Theta_r$ is dense in Θ_r for each r. Theorem 2.1 gives (2.8) as claimed.

2.2.7 Example: ψ-mixing, parametric family

Suppose that all the assumptions of Section 2 are valid so that we can apply Theorem 2.1. Suppose also that S and T are Polish spaces, that ρ is bounded and that each Q_θ is also ψ-mixing with

$$\sup_{n\ge 1} E\left[\operatorname*{ess\,inf}_{Q_\theta} \rho_n(X_1^n, Y_1^n) \right] \ne D, \qquad \forall \theta \in \Theta.$$

Chi (2001) [4] shows that $\Theta = \Theta_{\lim}$, so (2.8) holds.

2.2.8 Example: penalized lossy likelihoods

In the next two examples we point out certain interesting extensions of Theorem 2.1 that are part corollary and part simple modifications of the proof. We always assume that all of the assumptions of Section 2 are valid so that we can apply Theorem 2.1.

Let $F_n : \Theta \times S^n \to (-\infty, \infty]$, $n \ge 1$, be a sequence of functions with $F_n(\theta, \cdot)$ measurable for each $\theta \in \Theta$. We are interested in establishing the epi-convergence of $L_n + F_n$. F_n is thought of as a penalty. Assume that

$$\text{Prob}\left\{ \operatorname*{epi\text{-}lim\,inf}_{n\to\infty} F_n(\theta, X_1^n) \ge 0, \quad \forall \theta \in \Theta \right\} = 1. \tag{2.16}$$

Then (2.7) holds with L_n replaced by $L_n + F_n$. This is easy to see since $\text{epi-lim inf}_n (L_n + F_n) \ge \text{epi-lim inf}_n L_n + \text{epi-lim inf}_n F_n$.

Define

$$\Theta_{\lim}^F := \left\{ \theta : \limsup_{n\to\infty} \left[L_n(\theta, X_1^n) + F_n(\theta, X_1^n) \right] \overset{\text{a.s.}}{\le} \Lambda_\infty^*(\theta) \right\}.$$

If $\Theta_{\lim}^F \cap \Theta_r$ is dense in Θ_r for each $r < \infty$, then (2.8) also holds with L_n replaced by $L_n + F_n$. The proof is exactly the same as the proof of (2.8) in Section 3.2. Notice that because of (2.7) and (2.16), $\Theta_{\lim}^F$ is actually

$$\Theta_{\lim}^F = \left\{ \theta : \lim_{n\to\infty} \left[L_n(\theta, X_1^n) + F_n(\theta, X_1^n) \right] \overset{\text{a.s.}}{=} \Lambda_\infty^*(\theta) \right\}.$$

Typically, we will have $F_n \geq 0$, so that (2.16) is trivially true. Also, typically F_n does not depend on X_1^n, so measurability is not an issue and

$$\Theta_{\lim}^F = \Theta_{\lim} \cap \left\{ \theta : \lim_{n\to\infty} F_n(\theta) = 0 \right\}.$$

In many examples, we always have $F_n(\theta) \to 0$ as $n \to \infty$, so that $\Theta_{\lim}^F = \Theta_{\lim}$. In this special case, epi-convergence of L_n implies epi-convergence of $L_n + F_n$.

2.2.9 Example: approximate lossy likelihoods

Assume that all of the assumptions of Section 2 are valid. 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 [15] and can sometimes be simpler to compute than L_n in applications [M. Madiman, personal communication]. An inspection of the proof of (2.7) in Section 3.1 shows that we can replace L_n with R_n in (2.7). Indeed, the only fact that we use about L_n in the proof is that it satisfies the following Chebyshev inequality for any $\lambda \leq 0$:

$$L_n(\theta, x_1^n) \geq \lambda D - \frac{1}{n} \log E_\theta e^{\lambda n \rho_n(x_1^n, Y_1^n)}.$$

Trivially, R_n satisfies this same inequality. Furthermore, the same inequality shows that $L_n \geq R_n$. In particular, if (2.8) holds, then it also holds with L_n replaced by R_n.

2.3 Applications

In this section we describe two applications of the results. The first is an extension of the Markov chain example. This demonstrates how to extend the results to parameter spaces that might violate some of the assumptions. The second is a result about the convergence of maximizers and maxima of the lossy likelihood for compact Θ.

2.3.1 Finite state Markov chains

Let T be a finite set of size $|T|$. We want to extend Example 2.2.6 to the class of all first order (homogenous) Markov chains on T. We cannot use Theorem 2.1 directly. Some of the problems include: we do not know if Λ_∞^* is l.sc.; the uniform mixing condition (2.4) does not hold; we cannot simultaneously make each Q_θ stationary and satisfy (2.1) (which is equivalent to (2.2) in this case). We continue to assume that $(S, \mathcal{S})$ is a standard space, that $(X_n)_{n\geq 1}$ is stationary and ergodic taking values in S and that $\rho(\cdot, y)$ is measurable for each $y \in T$ (which gives product measurability for discrete T).

Let Θ be the set of all possible probability transition matrices for first order (homogenous) Markov chains on T with a metric equivalent to the Euclidean metric on $\mathbb{R}^{T\times T}$. For each $\theta \in \Theta$, we use p_θ to denote the transition matrix so that $p_\theta(i,j) = \text{Prob}\{Y_{k+1} = j | Y_k = i\}$. For each θ, let Q_θ be a Markov chain on T with uniform initial distribution and with transition probability matrix p_θ. Notice that Q_θ need not be stationary.

The definition Λ^*_∞ still makes sense. We first show that if Q_θ is irreducible, then $\Lambda^*_\infty(\theta)$ does not change if we use the uniform initial distribution instead of the (unique) stationary initial distribution. This lets us use the results from Example 2.2.6.

Then we show that

$$\text{Prob}\left\{\operatorname*{epi-lim\,inf}_{n\to\infty} L_n(\theta, X_1^n) \geq \text{lsc } \Lambda^*_\infty(\theta), \quad \forall\theta\in\Theta\right\} = 1, \tag{2.17}$$

where

$$\text{lsc } \Lambda^*_\infty(\theta) := \sup_{\epsilon>0} \inf_{\theta'\in O(\theta,\epsilon)} \Lambda^*_\infty(\theta')$$

is the lower semicontinuous envelope of Λ^*_∞. Example 2.2.6 shows that Λ^*_∞ is l.sc. on the open subset $\Theta_{\text{irr}} \subset \Theta$ of irreducible Markov chains. So Λ^*_∞ is equal to its l.sc. envelope on Θ_{irr} and (2.17) is a generalization of (2.7) when restricted to irreducible Markov chains.

Finally, if $E\left[\min_{y\in T}\rho(X_1, y)\right] \neq D$ or $D = 0$, then we show that

$$\text{Prob}\left\{\operatorname*{epi-lim}_{n\to\infty} L_n(\theta, X_1^n) = \text{lsc } \Lambda^*_\infty(\theta), \quad \forall\theta\in\Theta\right\} = 1. \tag{2.18}$$

Notice that Θ is compact, so epi-convergence immediately implies the convergence of minima and minimizers (see Section 2.3.2).

Suppose that Q_θ is irreducible and let π_θ be its unique stationary distribution (which has all positive elements). We use $\tilde{Q}_\theta$ to denote the unique stationary Markov chain with the same probability transition matrix as Q_θ and similarly for the other notation. For any set $A \in \mathcal{T}^n$

$$Q_\theta(A) = \sum_{y_1^n\in A} \frac{1}{|T|} \prod_{k=2}^n p_\theta(y_{k-1}, y_k) \leq K(\theta) \sum_{y_1^n\in A} \pi_\theta(y_1) \prod_{k=2}^n p_\theta(y_{k-1}, y_k) = K(\theta)\tilde{Q}_\theta(A),$$

where $K(\theta) := \left[|T| \min_{y\in T}\pi_\theta(y)\right]^{-1} < \infty$. Similarly,

$$Q_\theta(A) \geq \sum_{y_1^n\in A} \frac{\pi_\theta(y_1)}{|T|} \prod_{k=2}^n p_\theta(y_{k-1}, y_k) = |T|^{-1}\tilde{Q}_\theta(A).$$

Combining these gives

$$\begin{aligned}
\tilde{\Lambda}_n(\theta,\lambda) - n^{-1}\log|T| \leq \Lambda_n(\theta,\lambda) \leq \tilde{\Lambda}_n(\theta,\lambda) + n^{-1}\log K(\theta) \\
\tilde{L}_n(\theta, x_1^n) - n^{-1}\log K(\theta) \leq L_n(\theta, x_1^n) \leq \tilde{L}_n(\theta, x_1^n) + n^{-1}\log|T|
\end{aligned}$$

which implies

$$\begin{gathered}
\Lambda_\infty(\theta,\lambda) = \tilde{\Lambda}_\infty(\theta,\lambda), \\
\liminf_{n\to\infty} L_n(\theta, x_1^n) = \liminf_{n\to\infty} \tilde{L}_n(\theta, x_1^n) \quad \text{and} \quad \limsup_{n\to\infty} L_n(\theta, x_1^n) = \limsup_{n\to\infty} \tilde{L}_n(\theta, x_1^n).
\end{gathered} \tag{2.19}$$

So for irreducible Q_θ, $\Lambda^*_\infty(\theta)$ and asymptotic limits of $L_n(\theta, X_1^n)$ do not change if we use the uniform initial distribution in place of the stationary initial distribution.

Now we will verify (2.17). Fix $\alpha > 0$ and for each $\theta \in \Theta$ choose $\delta := \delta(\theta, \alpha) < \alpha/2$ such that

$$c(\theta,\delta) := \max_{y\in T} \sum_{y'\in T} \sup_{\theta'\in O(\theta,\delta)} p_{\theta'}(y, y') < 1 + \alpha$$

and such that $\theta^* := \theta^*(\theta, \alpha) \in O(\theta, \alpha/2)$ where θ^* has

$$p_{\theta^*}(y, y') := \frac{\sup_{\theta' \in O(\theta,\delta)} p_{\theta'}(y, y')}{\sum_{y'' \in T} \sup_{\theta' \in O(\theta,\delta)} p_{\theta'}(y, y'')}.$$

Notice that p_{θ^*} has all positive entries and is thus irreducible. Choose $\theta_1, \ldots, \theta_m$ with corresponding $\delta_1, \ldots, \delta_m$ and $\theta_1^*, \ldots, \theta_m^*$ such that the $\{O(\theta_j, \delta_j)\}_{j=1}^m$ cover Θ.

Fix $\theta \in \Theta$ and choose $1 \le j \le m$ such that $\theta \in O(\theta_j, \delta_j)$. For $\epsilon > 0$ sufficiently small so that $O(\theta, \epsilon) \subset O(\theta_j, \delta_j)$ we have

$$\begin{aligned}
\inf_{\theta' \in O(\theta,\epsilon)} L_n(\theta', x_1^n) &\ge \inf_{\theta' \in O(\theta_j,\delta_j)} L_n(\theta', x_1^n) \\
&\ge -\frac{1}{n} \log \sum_{y_1^n \in B(x_1^n, D)} \frac{1}{|T|} \prod_{k=2}^n \sup_{\theta' \in O(\theta_j,\delta_j)} p_{\theta'}(y_{k-1}, y_k) \\
&\ge -\frac{1}{n} \log \sum_{y_1^n \in B(x_1^n, D)} \frac{1}{|T|} \prod_{k=2}^n [p_{\theta_j^*}(y_{k-1}, y_k) c(\theta_j, \delta_j)] \\
&\ge L_n(\theta_j^*, x_1^n) - \log c(\theta_j, \delta_j) \ge \tilde{L}_n(\theta_j^*, x_1^n) - \log(1+\alpha) - n^{-1} \log K(\theta_j^*).
\end{aligned}$$

Taking limits and using Example 2.2.6 gives

$$\text{Prob}\left\{ \sup_{\epsilon > 0} \liminf_{n \to \infty} \inf_{\theta' \in O(\theta,\epsilon)} L_n(\theta', X_1^n) \ge \Lambda_\infty^*(\theta_j^*) - \log(1+\alpha) \right\} = 1. \tag{2.20}$$

$\nu(\theta_j^*, \theta) < \alpha$ and the exceptional sets in (2.20) only depend on θ_j^*, of which there are only finitely many. So we have

$$\text{Prob}\left\{ \operatorname*{epi-lim\,inf}_{n \to \infty} L_n(\theta, X_1^n) \ge \inf_{\theta' \in O(\theta,\alpha)} \Lambda_\infty^*(\theta') - \log(1+\alpha), \quad \forall \theta \in \Theta \right\} = 1.$$

Letting $\alpha \to 0$ gives (2.17).

Let Θ' denote the set of θ with all positive transition probabilities. We will show that $\Theta' \cap \Theta_r$ is dense in Θ_r for each r. This proceeds in a similar manner as Example 2.2.6. Fix $\theta \in \Theta_r$ and $\theta' \in \Theta'$. For each $0 < \epsilon < 1$, let $\theta_\epsilon \in \Theta'$ correspond to the transition probability matrix $p_{\theta_\epsilon} := (1-\epsilon) p_\theta + \epsilon p_{\theta'}$. We have

$$\Lambda_n(\theta_\epsilon, \lambda) \ge \frac{1}{n} E_P \left[\log \sum_{y_1^n \in T^n} e^{\lambda n \rho_n(X_1^n, y_1^n)} \frac{1}{|T|} \prod_{k=2}^n (1-\epsilon) p_\theta(y_{k-1}, y_k) \right] \ge \Lambda_n(\theta, \lambda) + \log(1-\epsilon)$$

and this gives $\Lambda_\infty^*(\theta_\epsilon) \le \Lambda_\infty^*(\theta) + \log(1-\epsilon)$. Taking ϵ small enough shows that $\Theta' \cap \Theta_r$ is dense in Θ_r for each r.

This implies that the l.sc. envelope of Λ_∞^* is completely specified by the value of Λ_∞^* on Θ'

$$\text{lsc } \Lambda_\infty^*(\theta) = \sup_{\epsilon > 0} \inf_{\theta' \in O(\theta,\epsilon) \cap \Theta'} \Lambda_\infty^*(\theta').$$

So we see that $\Theta' \cap \Theta_r^{\text{lsc}}$ is dense in Θ_r^{lsc} for each r where $\Theta_r^{\text{lsc}} := \{\theta : \text{lsc } \Lambda_\infty^*(\theta) < r\}$.

Suppose that $E[\min_{y \in T} \rho(X_1, y)] \ne D$ or $D = 0$, so that $\Theta' \subset \Theta_{\text{lim}}$ as shown in Example 2.2.6. Then we must have $\Theta_{\text{lim}} \cap \Theta_r^{\text{lsc}}$ is dense in Θ_r^{lsc} for each r and (2.18) is true.

2.3.2 Maximizers of the lossy likelihood

Assume everything in from Section 2 and suppose that Θ is compact with metric ν. Choose

$$\hat{\theta}_n(x_1^n) \in \arg\max_{\theta\in\Theta} Q_\theta(B(x_1^n, D))$$

to be a maximizer of the lossy likelihood, or equivalently, choose

$$\hat{\theta}_n(x_1^n) \in \arg\min_{\theta\in\Theta} L_n(\theta, x_1^n)$$

to be a minimizer of L_n. We will show that $L_n(\cdot, x_1^n)$ is l.sc. for each x_1^n, so a minimizer always exists. If there are multiple minimizers, just choose one.

Let $\Theta^* := \arg\min_{\theta\in\Theta} \Lambda_\infty^*(\theta)$ be the set of minimizers of Λ_∞^*, which is not empty because Λ_∞^* is l.sc. and Θ is compact. If (2.8) holds, then the fact that Θ is compact immediately gives [1]

$$\text{Prob}\left\{\lim_{n\to\infty} \nu(\hat{\theta}_n(X_1^n), \Theta^*) = 0\right\} = 1,$$
$$\text{Prob}\left\{\lim_{n\to\infty} L_n(\hat{\theta}_n(X_1^n), X_1^n) = \inf_{\theta\in\Theta} \Lambda_\infty^*(\theta)\right\} = 1.$$

Notice that if we are in the setting of Example 2.2.8 and the penalties $F_n(\cdot, x_1^n)$ are l.sc. for each x_1^n, then all of this holds with L_n replaced by $L_n + F_n$. Similarly, for the setting of Example 2.2.9. The approximation $R_n(\cdot, x_1^n)$ is l.sc. because it is a supremum of continuous functions (see (A.2) in the Appendix).

We need only show that $L_n(\cdot, x_1^n)$ is l.sc., or equivalently, that $\theta \mapsto Q_\theta(B(x_1^n, D))$ is u.sc. If (2.1) holds, then the latter is continuous by definition. If (2.2) holds, then $B(x_1^n, D)$ is closed and u.sc. follows from a well known property of weak convergence of probability measures [14][pp. 311].

The l.sc. of L_n follows from (2.1) or (2.2). Suppose we are in a setting where either of these hold and where (2.8) holds with Λ_∞^* possibly replaced by another (necessarily l.sc.) function Γ. Then the results of this section hold with Λ_∞^* replaced by Γ. An example of this is given in Section 2.3.1 where many of the assumptions of Section 2 are not valid, but (2.1) holds and (2.8) holds with Λ_∞^* replaced by lsc Λ_∞^*.

3 Proof of Theorem 2.1

We deal with most measurability issues in the Appendix. That the distribution of $(X_n)_{n\geq 1}$ is assumed to be complete clears up many problems. If we assume that Θ is locally compact, then (with a lot of extra work) we can relax this assumption using the results of Pfanzagl (1969) [12]. It is worth noting that we do not assume that each Q_θ is complete, because then $(T, \mathcal{T})$ might vary with θ (unless we assumed some dominating measure). In the case where we are dealing with w-convergence (2.2), the assumption that $\rho(x, \cdot)$ is continuous is important for establishing measurability. Most of the arguments actually extend to the case where $\rho(x, \cdot)$ is l.sc. and where all of the quantities that we need are measurable.

The Appendix also contains a listing of several nice properties of Λ and Λ^* as functions of θ. These are useful for the proofs. We begin with the lower bound.

3.1 Proof: Lower bound

Fix $\theta \in \Theta$ and use (2.4) to choose finite $C, d \geq 1$ and $\delta := \delta_\theta > 0$, so that

$$\sup_{\theta' \in O(\theta,\delta)} C(\theta') \leq C \qquad \text{and} \qquad \sup_{\theta' \in O(\theta,\delta)} d(\theta') \leq d.$$

Define

$$h_n(x_1^n, y_1^n) := \begin{cases} \sum_{k=1}^{n-d+1} \rho(x_k, y_k) & \text{if } d \leq n, \\ 0 & \text{otherwise.} \end{cases}$$

For each $\lambda \leq 0$ and any $0 < \epsilon \leq \delta$, we have

$$\begin{aligned}
&\sup_{\theta' \in O(\theta,\epsilon)} \log E_{\theta'} e^{\lambda h_{n+m}(x_1^{n+m}, Y_1^{n+m})} + \log C \\
&\quad \leq \sup_{\theta' \in O(\theta,\epsilon)} \log E_{\theta'} e^{\lambda h_n(x_1^n, Y_1^n)} e^{\lambda h_m(x_{n+1}^{n+m}, Y_{n+1}^{n+m})} + \log C \\
&\quad \leq \sup_{\theta' \in O(\theta,\epsilon)} \left[\log E_{\theta'} e^{\lambda h_n(x_1^n, Y_1^n)} + \log C + \log E_{\theta'} e^{\lambda h_m(x_{n+1}^{n+m}, Y_1^m)} \right] + \log C \\
&\quad \leq \sup_{\theta' \in O(\theta,\epsilon)} \log E_{\theta'} e^{\lambda h_n(x_1^n, Y_1^n)} + \log C + \sup_{\theta'' \in O(\theta,\epsilon)} \log E_{\theta''} e^{\lambda h_m(x_{n+1}^{n+m}, Y_1^m)} + \log C,
\end{aligned}$$

where the first inequality follows since ρ is nonnegative and $\lambda \leq 0$ and the second from the stationarity and mixing properties of each $Q_{\theta'}$. Since each of these terms is bounded above by $\log C$, the pointwise subadditive ergodic theorem gives

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \sup_{\theta' \in O(\theta,\epsilon)} \log E_{\theta'} e^{\lambda h_n(X_1^n, Y_1^n)} &\overset{\text{a.s.}}{=} \lim_{n\to\infty} \frac{1}{n} E_P \left[\sup_{\theta' \in O(\theta,\epsilon)} \log E_{\theta'} e^{\lambda h_n(X_1^n, Y_1^n)} \right] \\
&= \inf_{n \geq N} \frac{1}{n} E_P \left[\sup_{\theta' \in O(\theta,\epsilon)} \log E_{\theta'} e^{\lambda h_n(X_1^n, Y_1^n)} + \log C \right] \qquad (3.1)
\end{aligned}$$

for each N, where we have removed the $(\log C)/n$ in the first two expressions because it becomes negligible in the limit.

Notice that

$$h_n(x_1^n, y_1^n) \leq n\rho_n(x_1^n, y_1^n) = h_{n+d-1}(x_1^{n+d-1}, y_1^{n+d-1})$$

so that for $\lambda \leq 0$

$$\log E_{\theta'} e^{\lambda h_n(x_1^n, Y_1^n)} \geq \log E_{\theta'} e^{\lambda n \rho_n(x_1^n, Y_1^n)} = \log E_{\theta'} e^{\lambda h_{n+d-1}(x_1^{n+d-1}, y_1^{n+d-1})}.$$

Combining this with (3.1) gives

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \sup_{\theta' \in O(\theta,\epsilon)} \log E_{\theta'} e^{\lambda n \rho_n(X_1^n, Y_1^n)} &\overset{\text{a.s.}}{=} \lim_{n\to\infty} \frac{1}{n} E_P \left[\sup_{\theta' \in O(\theta,\epsilon)} \log E_{\theta'} e^{\lambda n \rho_n(X_1^n, Y_1^n)} \right] \\
&= \inf_{n \geq N} \frac{1}{n+d-1} E_P \left[\sup_{\theta' \in O(\theta,\epsilon)} \log E_{\theta'} e^{\lambda n \rho_n(X_1^n, Y_1^n)} + \log C \right] \qquad (3.2)
\end{aligned}$$

for any N. Notice that we can repeat each step of the proof so far without the $\sup_{\theta' \in O(\theta,\epsilon)}$ to get that $\lim_{n\to\infty} \Lambda_n(\theta, \lambda)$ exists. So we have

$$\Lambda_\infty(\theta, \lambda) = \lim_{n\to\infty} \Lambda_n(\theta, \lambda), \qquad \lambda \leq 0,\ \theta \in \Theta. \qquad (3.3)$$

This proves (2.5).

Finally, letting $\epsilon \downarrow 0$ (take ϵ rational to control the exceptional sets) in (3.2) gives

$$\begin{aligned}
&\lim_{\epsilon\downarrow 0}\lim_{n\to\infty}\frac{1}{n}\sup_{\theta'\in O(\theta,\epsilon)}\log E_{\theta'}e^{\lambda n\rho_n(X_1^n,Y_1^n)} \overset{\text{a.s.}}{=} \lim_{\epsilon\downarrow 0}\lim_{n\to\infty}\frac{1}{n}E_P\left[\sup_{\theta'\in O(\theta,\epsilon)}\log E_{\theta'}e^{\lambda n\rho_n(X_1^n,Y_1^n)}\right]\\
&\quad=\inf_{\epsilon>0}\inf_{n\geq N}\frac{1}{n+d+1}E_P\left[\sup_{\theta'\in O(\theta,\epsilon)}\log E_{\theta'}e^{\lambda n\rho_n(X_1^n,Y_1^n)}+\log C\right]\\
&\quad=\inf_{n\geq N}\left[\frac{1}{n+d+1}\inf_{\epsilon>0}E_P\left[\sup_{\theta'\in O(\theta,\epsilon)}\log E_{\theta'}e^{\lambda n\rho_n(X_1^n,Y_1^n)}\right]+\frac{\log C}{n+d-1}\right]\\
&\quad=\inf_{n\geq N}\left[\frac{1}{n+d+1}E_P\left[\inf_{\epsilon>0}\sup_{\theta'\in O(\theta,\epsilon)}\log E_{\theta'}e^{\lambda n\rho_n(X_1^n,Y_1^n)}\right]+\frac{\log C}{n+d-1}\right]\\
&\quad=\inf_{n\geq N}\left[\frac{n}{n+d-1}\Lambda_n(\theta,\lambda)+\frac{\log C}{n+d-1}\right]=\Lambda_\infty(\theta,\lambda),
\end{aligned}\tag{3.4}$$

where the fourth equality follows from the monotone convergence theorem, the fifth equality from the fact that $\log E_\theta e^{\lambda n\rho_n(x_1^n,Y_1^n)}$ is continuous in θ for each x_1^n as shown in the Appendix, and the final equality from (3.3) and the fact that N is arbitrary.

When $\lambda\leq 0$, Chebyshev's inequality gives

$$-\frac{1}{n}\log Q_{\theta'}(B(x_1^n,D))\geq -\frac{1}{n}\log E_{\theta'}e^{\lambda n(\rho_n(X_1^n,Y_1^n)-D)}=\lambda D-\frac{1}{n}\log E_{\theta'}e^{\lambda n\rho_n(x_1^n,Y_1^n)}.\tag{3.5}$$

Combining this with (3.4) gives

$$\begin{aligned}
&\lim_{\epsilon\downarrow 0}\liminf_{n\to\infty}\inf_{\theta'\in O(\theta,\epsilon)}L_n(\theta',X_1^n)\geq \lambda D-\lim_{\epsilon\downarrow 0}\limsup_{n\to\infty}\frac{1}{n}\sup_{\theta'\in O(\theta,\epsilon)}\log E_{\theta'}e^{\lambda n\rho_n(X_1^n,Y_1^n)}\\
&\quad\overset{\text{a.s.}}{=}\lambda D-\Lambda_\infty(\theta,\lambda).
\end{aligned}$$

Optimizing over λ (take λ rational to control the exceptional sets) gives

$$\lim_{\epsilon\downarrow 0}\liminf_{n\to\infty}\inf_{\theta'\in O(\theta,\epsilon)}L_n(\theta',X_1^n)\overset{\text{a.s.}}{\geq}\Lambda_\infty^*(\theta)$$

and we have proved that

$$\text{Prob}\left\{\operatorname*{epi-lim\,inf}_{n\to\infty}L_n(\theta,X_1^n)\geq\Lambda_\infty^*(\theta)\right\}=1,\quad\forall\theta\in\Theta.$$

The reason we can restrict the supremum to rational $\lambda\leq 0$, is that $\lambda D-\Lambda_\infty(\theta,\lambda)$ is concave in λ [8].

Now we want to move the $\forall\theta\in\Theta$ inside the probability. We begin with (3.4) to compute that

$$\begin{aligned}
&\lim_{\epsilon\downarrow 0}\sup_{\lambda\leq 0}\left[\lambda D-\lim_{n\to\infty}\frac{1}{n}E_P\left[\sup_{\theta'\in O(\theta,\epsilon)}\log E_{\theta'}e^{\lambda n\rho_n(X_1^n,Y_1^n)}\right]\right]\\
&\quad=\sup_{\lambda\leq 0}\sup_{\epsilon>0}\left[\lambda D-\lim_{n\to\infty}\frac{1}{n}E_P\left[\sup_{\theta'\in O(\theta,\epsilon)}\log E_{\theta'}e^{\lambda n\rho_n(X_1^n,Y_1^n)}\right]\right]\\
&\quad=\sup_{\lambda\leq 0}\left[\lambda D-\inf_{\epsilon>0}\lim_{n\to\infty}\frac{1}{n}E_P\left[\sup_{\theta'\in O(\theta,\epsilon)}\log E_{\theta'}e^{\lambda n\rho_n(X_1^n,Y_1^n)}\right]\right]\\
&\quad=\sup_{\lambda\leq 0}\left[\lambda D-\Lambda_\infty(\theta,\lambda)\right]=\Lambda_\infty^*(\theta).
\end{aligned}\tag{3.6}$$

Fix $\alpha > 0$. Using (3.6), for each θ choose $0 < r = r_{\theta,\alpha} < \delta_\theta \wedge \alpha$ small enough that

$$\sup_{\lambda \le 0} \left[\lambda D - \lim_{n\to\infty} \frac{1}{n} E_P \left[\sup_{\theta' \in O(\theta,r)} \log E_{\theta'} e^{\lambda n \rho_n(X_1^n, Y_1^n)} \right] \right] > \alpha^{-1} \wedge \Lambda_\infty^*(\theta) - \alpha. \tag{3.7}$$

Using Chebyshev's inequality (3.5) and (3.2) gives

$$\begin{aligned} \liminf_{n\to\infty} \inf_{\theta' \in O(\theta,r)} L_n(\theta', X_1^n) &\ge \lambda D - \limsup_{n\to\infty} \frac{1}{n} \sup_{\theta' \in O(\theta,r)} \log E_{\theta'} e^{\lambda n \rho_n(X_1^n, Y_1^n)} \\ &\overset{\text{a.s.}}{=} \lambda D - \lim_{n\to\infty} \frac{1}{n} E_P \left[\sup_{\theta' \in O(\theta,r)} \log E_{\theta'} e^{\lambda n \rho_n(X_1^n, Y_1^n)} \right]. \end{aligned}$$

The final expression is shown to be concave in λ in the Appendix. Optimizing over λ (λ rational as before) and using (3.7) gives

$$\liminf_{n\to\infty} \inf_{\theta' \in O(\theta,r)} L_n(\theta', X_1^n) \overset{\text{a.s.}}{>} \alpha^{-1} \wedge \Lambda_\infty^*(\theta) - \alpha. \tag{3.8}$$

The collection of neighborhoods $\{O(\theta, r)\}_{\theta\in\Theta}$ is an open cover for Θ. Since Θ is a separable metric space, we can choose a countable subcover $\{O(\theta, r)\}_{\theta\in G}$ for some $G \subset \Theta$. Since G is discrete, we can use (3.8) to get

$$\text{Prob}\left\{ \liminf_{n\to\infty} \inf_{\theta' \in O(\theta,r)} L_n(\theta', X_1^n) > \alpha^{-1} \wedge \Lambda_\infty^*(\theta) - \alpha, \quad \forall \theta \in G \right\} = 1. \tag{3.9}$$

For each $\theta \in \Theta$, there is a $\tilde\theta := \tilde\theta_{\theta,\alpha} \in G$ so that $\theta \in O(\tilde\theta, \tilde r)$, where $\tilde r := r_{\tilde\theta,\alpha}$. For all ϵ small enough, $O(\theta, \epsilon) \subset O(\tilde\theta, \tilde r)$ and we have

$$\lim_{\epsilon\downarrow 0} \liminf_{n\to\infty} \inf_{\theta' \in O(\theta,\epsilon)} L_n(\theta', x_1^n) \ge \liminf_{n\to\infty} \inf_{\theta' \in O(\tilde\theta,\tilde r)} L_n(\theta', x_1^n).$$

So

$$\begin{aligned} &\left\{ \lim_{\epsilon\downarrow 0} \liminf_{n\to\infty} \inf_{\theta' \in O(\theta,\epsilon)} L_n(\theta', X_1^n) > \alpha^{-1} \wedge \inf_{\theta' \in O(\theta,\alpha)} \Lambda_\infty^*(\theta') - \alpha, \quad \forall \theta \in \Theta \right\} \\ &\quad \supset \left\{ \liminf_{n\to\infty} \inf_{\theta' \in O(\tilde\theta,\tilde r)} L_n(\theta', X_1^n) > \alpha^{-1} \wedge \Lambda_\infty^*(\tilde\theta) - \alpha, \quad \forall \tilde\theta \in G \right\}, \end{aligned} \tag{3.10}$$

where we have used the fact that $\tilde r < \alpha$ to get $\tilde\theta \subset O(\theta, \alpha)$. (3.9) shows that the last expression in (3.10) has probability 1, so we must have

$$\text{Prob}\left\{ \lim_{\epsilon\downarrow 0} \liminf_{n\to\infty} \inf_{\theta' \in O(\theta,\epsilon)} L_n(\theta', X_1^n) > \alpha^{-1} \wedge \inf_{\theta' \in O(\theta,\alpha)} \Lambda_\infty^*(\theta') - \alpha, \quad \forall \theta \in \Theta \right\} = 1$$

as well. Letting $\alpha \downarrow 0$ (α rational) will complete the proof of (2.7) as long as we can show that Λ_∞^* is l.sc. on Θ.

The Appendix shows that $\Lambda_n(\cdot, \lambda)$ is u.sc. for each $\lambda \le 0$ and $n < \infty$. (3.4) thus shows that $\Lambda_\infty(\cdot, \lambda)$ is an infimum of u.sc. functions and so it is also u.sc. for each $\lambda \le 0$. This shows that Λ_∞^* is a supremum of l.sc. functions, which is l.sc.

3.2 Proof: Upper bound

For each $r < \infty$, let $G_r \subset \Theta_{\lim}$ be a countable, dense subset of Θ_r. If Θ_r is empty (as it will always be for $r \leq 0$), then we take $G_r = \emptyset$. By assumption, we can always find such a collection of G_r. Define

$$G := \bigcup_{r \in \mathbb{Q}} G_r$$

to be the union of all G_r for rational r. Then $G \subset \Theta_{\lim}$ is at most countable and we have

$$\text{Prob}\left\{\limsup_{n\to\infty} L_n(\theta, X_1^n) \leq \Lambda_\infty^*(\theta), \quad \forall \theta \in G\right\} = 1.$$

If $\theta \in \Theta_r$ for some $r \in \mathbb{R}$, then $\theta \in \Theta_q \subset \Theta_r$ for some $q \in \mathbb{Q}$ with $\Lambda_\infty^*(\theta) < q < r$. Since $G \cap \Theta_q$ is dense in Θ_q for each $q \in \mathbb{Q}$, $G \cap \Theta_r$ is also dense in Θ_r for each $r \in \mathbb{R}$. Proposition 1.1 gives

$$\left\{\limsup_{n\to\infty} L_n(\theta, X_1^n) \leq \Lambda_\infty^*(\theta), \quad \forall \theta \in G\right\} \subset \left\{\text{epi-}\limsup_{n\to\infty} L_n(\theta, X_1^n) \leq \Lambda_\infty^*(\theta), \quad \forall \theta \in \Theta\right\}$$

and we have

$$\text{Prob}\left\{\text{epi-}\limsup_{n\to\infty} L_n(\theta, X_1^n) \leq \Lambda_\infty^*(\theta), \quad \forall \theta \in \Theta\right\} = 1.$$

This is the epi-lim sup upper bound. Combining it with (2.7) gives (2.8) and completes the proof of Theorem 2.1.

4 Proof of Theorem 2.3

As before, measurability issues are addressed in the Appendix. Since Λ_n^* does not depend on n, we will drop the subscript. The Appendix shows that Λ^* is convex. Proposition 2.4 and (2.13) are proved in Section 4.1, although we refer to the results here.

We first note that Theorem 2.1 immediately gives us

$$\text{Prob}\left\{\text{epi-}\liminf_{n\to\infty} L_n(\theta, X_1^n) \geq \Lambda^*(\theta), \quad \forall \theta \in \Theta\right\} = 1, \tag{4.1}$$

so to establish (2.14) and (2.15c) we need only establish the epi-lim sup upper bound.

Theorem 2.2 implies that $\Theta_\infty^c \subset \Theta_{\lim}$, so there are three mutually exclusive possibilities:

$$\Theta_\infty = \emptyset, \tag{4.2}$$

$$\Theta_{\lim} \cap \Theta_\infty \neq \emptyset, \tag{4.3}$$

$$\Theta_{\lim}^c = \Theta_\infty \neq \emptyset. \tag{4.4}$$

In words, (4.2) says that $\Lambda^*(\theta) = \infty$ for all θ. (4.3) says that there is at least one θ with a finite, pointwise limit (2.11). And (4.4) says that every θ with a finite $\Lambda^*(\theta)$ does not have an a.s. pointwise limit (2.12) and there is at least one such "bad" θ. We will first show that (4.2) or (4.3) implies (2.14) and that (4.4) implies $D_{\min}(\Theta) = D$.

Suppose (4.2) is true. Then $\Lambda^* \equiv \infty$ and (4.1) gives (2.14).

Suppose (4.3) is true and choose $\theta_0 \in \Theta_{\lim} \cap \Theta_\infty$. Pick r so that Θ_r is not empty and choose $\theta \in \Theta_r$ and $\epsilon > 0$. In the next paragraph, we will construct a point $\theta' \in \Theta_{\lim} \cap \Theta_r \cap O(\theta, \epsilon)$ by taking a convex combination of θ and θ_0. This shows that $\Theta_{\lim} \cap \Theta_r$ is dense in Θ_r for each r and we can apply Theorem 2.1 to get (2.14).

Λ^* is convex and $\Lambda^*(\theta_0) < \infty$, so we can choose $0 < \lambda_0 < 1$ such that $0 < \lambda < \lambda_0$ implies

$$\Lambda^*(\lambda\theta_0 + (1-\lambda)\theta) \le (1-\lambda)\Lambda^*(\theta) + \lambda\Lambda^*(\theta_0) < r.$$

Furthermore, since $\lambda\theta_0 + (1-\lambda)\theta \to \theta$ as $\lambda \downarrow 0$, we can choose $0 < \lambda < \lambda_0$ such that $\theta' := \lambda\theta_0 + (1-\lambda)\theta \in \Theta_r \cap O(\theta, \epsilon)$. Recalling that we can write $Q_{\theta',n} = [Q_{\theta',1}]^n$, we have

$$\begin{aligned}
\limsup_{n\to\infty} L_n(\theta', X_1^n) &= \limsup_{n\to\infty} -\frac{1}{n}\log[\lambda Q_{\theta_0,1} + (1-\lambda)Q_{\theta,1}]^n(B(X_1^n, D)) \\
&\le \limsup_{n\to\infty} -\frac{1}{n}\log[\lambda Q_{\theta_0,1}]^n(B(X_1^n, D)) = -\log\lambda + \limsup_{n\to\infty} L_n(\theta_0, X_1^n) \\
&\overset{\text{a.s.}}{=} -\log\lambda + \Lambda^*(\theta_0) < \infty,
\end{aligned}$$

so (2.12a) cannot be true for θ' and Theorem 2.2 implies that $\theta' \in \Theta_{\lim}$. So $\theta' \in \Theta_{\lim} \cap \Theta_r \cap O(\theta, \epsilon)$ as desired and we have finished case (4.3).

Suppose (4.4) is true. Fix $\theta_0 \in \Theta_\infty = \Theta_{\lim}^c$. Theorem 2.2 shows that $0 < D_{\min}(\theta_0) = D < \infty$. This shows that $\inf_{\theta\in\Theta} D_{\min}(\theta) \le D$. Suppose there exists a θ with $D_{\min}(\theta) < D$. Let $\theta' = \lambda\theta_0 + (1-\lambda)\theta$ for some $0 < \lambda < 1$. The Appendix shows that $\Lambda^*(\theta') \le \Lambda^*(\theta_0) - \log\lambda < \infty$, so $\theta' \in \Theta_\infty$. But it is easy to see that $D_{\min}(\theta') \le D_{\min}(\theta) < D$, so Theorem 2.2 implies that $\theta' \in \Theta_{\lim}$. We have shown that $\theta' \in \Theta_\infty \cap \Theta_{\lim}$ which contradicts (4.4) and there cannot be any θ with $D_{\min}(\theta) < D$. Thus, $\inf_{\theta\in\Theta} D_{\min}(\theta) = D < \infty$ and Proposition 2.4 shows that $D_{\min}(\Theta) = D$, which is what we wanted to prove. In fact, we have proved

$$0 < D_{\min}(\Theta) = \min_{\theta\in\Theta} D_{\min}(\theta) = D < \infty \quad \text{and} \quad D_{\min}(\theta) = D \text{ for all } \theta \in \Theta_\infty. \tag{4.5}$$

Since we must have $m(\Theta, x) \le m(\theta, x)$ for any θ, (4.5) shows that

$$m(\Theta, X_1) \overset{\text{a.s.}}{=} m(\theta, X_1), \text{ for all } \theta \in \Theta_\infty. \tag{4.6}$$

Now we will prove (2.14). If $D \ne D_{\min}(\Theta)$, then we cannot be in situation (4.4). We must be in either situation (4.2) or situation (4.3), so (2.14) holds. If $\Lambda_\infty^*(\Theta) = \infty$, then we are in situation (4.2) and (2.14) holds. If $m(\Theta, X_1)$ is a.s. constant, then we are either in one of (4.2) or (4.3), in which case (2.14) holds, or we are in (4.4). In the latter case, (4.6) shows that $m(\theta, X_1)$ is a.s. constant for each $\theta \in \Theta_\infty$. For any such θ, Theorem 2.2 shows that $\theta \in \Theta_{\lim}$, which contradicts (4.4). So the latter case is impossible and we have proved (2.14).

Now we will prove (2.15). Suppose $D = D_{\min}(\Theta)$, $\Lambda_\infty^*(\Theta) < \infty$ and $m(\Theta, X_1)$ is not a.s. constant. We must have $0 < D_{\min}(\Theta) < \infty$. Theorem 2.2 can be used to infer that $D_{\min}(\theta) \le D$ for each $\theta \in \Theta_\infty$, which is not empty by assumption. Using Proposition 2.4 gives both parts of (4.5). As before, this leads to (4.6). Since $m(\Theta, X_1)$ is not a.s. constant, neither is $m(\theta, X_1)$ for each $\theta \in \Theta_\infty$ and Theorem 2.2 shows that $\theta \notin \Theta_{\lim}$. So $\Theta_\infty = \Theta_{\lim}^c \ne \emptyset$ and we must be in situation (4.4).

Fix $\theta' \in \Theta_\infty$. Theorem 2.2 shows that

$$\text{Prob}\left\{\sum_{k=1}^n m(\theta', X_k) > nD \text{ i.o.}\right\} > 0 \quad \text{and} \quad \text{Prob}\left\{\sum_{k=1}^n m(\theta', X_k) \le nD \text{ i.o.}\right\} = 1,$$

so (4.6) gives

$$\text{Prob}\left\{\sum_{k=1}^{n} m(\Theta, X_k) > nD \text{ i.o.}\right\} > 0, \tag{4.7}$$

$$\text{Prob}\left\{\sum_{k=1}^{n} m(\Theta, X_k) \le nD \text{ i.o.}\right\} = 1. \tag{4.8}$$

For any x_1^n we have

$$\begin{aligned}\inf_{\theta\in\Theta} L_n(\theta, x_1^n) &\ge \inf_{\theta\in\Theta} -\frac{1}{n}\log Q_\theta\left\{y_1^n : \sum_{k=1}^{n} m(\theta, x_k) \le nD\right\} \\ &\ge \inf_{\theta\in\Theta} -\frac{1}{n}\log Q_\theta\left\{y_1^n : \sum_{k=1}^{n} m(\Theta, x_k) \le nD\right\} = \infty \text{ if } \sum_{k=1}^{n} m(\Theta, x_k) > nD. \end{aligned} \tag{4.9}$$

Recalling that $\theta' \in \Theta_\infty$, we have the following series of implications

$$\begin{aligned}\inf_{\theta\in\Theta} L_n(\theta, X_1^n) < \infty \implies \sum_{k=1}^{n} m(\Theta, X_k) \le nD \overset{\text{a.s.}}{\implies} \sum_{k=1}^{n} m(\theta', X_k) \le nD \\ \overset{\text{a.s.}}{\implies} L_n(\theta', X_1^n) < \infty \implies \inf_{\theta\in\Theta} L_n(\theta, X_1^n) < \infty, \end{aligned} \tag{4.10}$$

where the first implication follows from (4.9), the second from (4.6) and the third from Theorem 2.2. This shows that

$$\left\{\inf_{\theta\in\Theta} L_n(\theta, X_1^n) < \infty\right\} \quad \text{and} \quad \left\{\sum_{k=1}^{n} m(\Theta, X_k) \le nD\right\} \quad \text{differ by a null set.} \tag{4.11}$$

Combining (4.7) and (4.11) proves (2.15a). Combining (4.8) and (4.11) proves (2.15b) and shows that $(n_m)_{m\ge1}$ satisfies that claims of the theorem.

The last thing to prove is (2.15c). The proof closely follows the steps in Section 3.2. Choose an at most countable, dense subset $G_r \subset \Theta_r$ and let $G := \bigcup_{r\in\mathbb{Q}} G_r$, so that $G \subset \Theta_\infty$ is at most countable and $G \cap \Theta_r$ is dense in Θ_r for each r. (4.6) implies

$$\left\{\sum_{k=1}^{n} m(\Theta, X_k) \le nD\right\} \quad \text{and} \quad \left\{\sum_{k=1}^{n} m(\theta, X_k) \le nD, \ \forall\theta \in G\right\} \quad \text{differ by a null set.}$$

So we can take $(n_m)_{m\ge1}$ to be the (a.s.) infinite subsequence where the latter event occurs. Theorem 2.2 (2.12c) gives

$$\text{Prob}\left\{\limsup_{m\to\infty} L_{n_m}(\theta, X_1^{n_m}) \le \Lambda^*(\theta), \ \forall\theta \in G\right\} = 1$$

and (2.15c) follows from Proposition 1.1 and (4.1) just like in Section 3.2. This completes the proof of all parts of Theorem 2.3.

4.1 Proof of Proposition 2.4

Here we prove Proposition 2.4 and (2.13). We always have

$$\inf_{\theta\in\Theta} D_{\min}(\theta) = \inf_{\theta\in\Theta} Em(\theta, X_1) \geq Em(\Theta, X_1) = D_{\min}(\Theta). \tag{4.12}$$

Choose a countable dense subset $G \subset \Theta$. The Appendix shows that $m(\cdot, x)$ is u.sc. for each x, so $m(\Theta, x) = \inf_{\theta\in G} m(\theta, x)$ for each x. Let $(\theta_n)_{n\geq 1}$ be an enumeration of G. Define $\hat{\theta}_n := \sum_{k=1}^n \theta_k/n$. Recalling that $Q_{\hat{\theta}_n,1} = \sum_{k=1}^n Q_{\theta_k,1}/n$, it is easy to see that

$$0 \leq m(\Theta, x) \leq m(\hat{\theta}_n, x) \leq \min_{1\leq k\leq n} m(\theta_k, x) \downarrow \inf_{\theta\in G} m(\theta, x) = m(\Theta, x) \tag{4.13}$$

as $n \to \infty$.

If $\inf_{\theta\in\Theta} D_{\min}(\theta) < \infty$, then without loss of generality we can assume that θ_1 has $Em(\theta_1, X_1) := D_{\min}(\theta_1) < \infty$. In this case, the dominated convergence theorem applied to (4.13) will give

$$D_{\min}(\hat{\theta}_n) := Em(\hat{\theta}_n, X_1) \downarrow Em(\Theta, X_1) = D_{\min}(\Theta).$$

This gives $\inf_{\theta\in\Theta} D_{\min}(\theta) \leq D_{\min}(\Theta)$ and completes the proof of Proposition 2.4. We no longer assume that $\inf_{\theta\in\Theta} D_{\min}(\theta) < \infty$.

Suppose $D_{\min}(\Theta) > D$. Then ergodic theorem gives

$$\frac{1}{n}\sum_{k=1}^n m(\Theta, X_k) \stackrel{\text{a.s.}}{\to} D_{\min}(\Theta) > D$$

So

$$\text{Prob}\left\{\sum_{k=1}^n m(\Theta, X_k) > nD, \text{ eventually}\right\} = 1. \tag{4.14}$$

For any θ, we have the following trivial implications

$$\sum_{k=1}^n m(\Theta, x_k) > nD \implies \sum_{k=1}^n m(\theta, x_k) > nD \implies Q_\theta(B(x_1^n, D)) = 0$$
$$\implies L_n(\theta, x_1^n) = \infty.$$

Combining this with (4.14) gives the first part of (2.13).

Now suppose $D_{\min}(\Theta) < D$. Again the ergodic theorem gives

$$\text{Prob}\left\{\sum_{k=1}^n m(\Theta, X_k) < nD, \text{ eventually}\right\} = 1. \tag{4.15}$$

Suppose that for some n and some sequence x_1^n we have

$$\sum_{k=1}^n m(\Theta, x_k) < nD. \tag{4.16}$$

Then using (4.13) we can take N large enough so that

$$\sum_{k=1}^n m(\hat{\theta}_N, x_k) < nD$$

which implies that $Q_{\hat{\theta}_N}(B(x_1^n, D)) > 0$ and thus $L_n(\hat{\theta}_N, x_1^n) < \infty$. In particular, (4.16) implies that $\inf_{\theta\in\Theta} L_n(\theta, x_1^n) < \infty$. (4.15) completes the proof of (2.13).

A Appendix

We have ignored some important measurability issues in the text. We address them here. We also discuss some continuity and convexity properties that are used in the proofs. Unlike the main text, some of the arguments here will differ depending on whether we are in situation (2.1) or (2.2). We begin with the proof of Proposition 1.1.

Proof of Proposition 1.1. Fix θ. If $g(\theta) = \infty$, there is nothing to prove, so choose finite $r > g(\theta)$. Fix $\epsilon > 0$. By hypothesis, there exists a $\theta' \in O(\theta, \epsilon)$ with $f(\theta') < r$, so

$$\limsup_{n\to\infty} \inf_{\theta'' \in O(\theta,\epsilon)} f_n(\theta'') \leq \limsup_{n\to\infty} f_n(\theta') = f(\theta') < r.$$

Letting $\epsilon \downarrow 0$ gives epi-$\limsup_n f_n(\theta) \leq r$. Since $r > g(\theta)$ was arbitrary, the proof is complete. □

A.1 Measurability and continuity

Define

$$m(\theta, x) := \operatorname*{ess\,inf}_{Q_\theta} \rho(x, Y_1), \qquad \Lambda_n(\theta, \lambda, x_1^n) := \frac{1}{n} \log E_\theta e^{\lambda n \rho_n(x_1^n, Y_1^n)}.$$

The measurability of these functions in x or x_1^n for fixed θ and $\lambda \leq 0$ is established in Harrison (2003) [8]. We will show that

$$m(\cdot, x) \text{ is u.sc. for each } x \in S, \tag{A.1}$$

$$\Lambda_n(\cdot, \lambda, x_1^n) \text{ is continuous for each } x_1^n \in S^n \text{ and } \lambda \leq 0. \tag{A.2}$$

From (A.1), we see that

$$m(\Theta, x) := \inf_{\theta\in\Theta} m(\theta, x) = \inf_{\theta\in G} m(\theta, x)$$

for any countable dense $G \subset \Theta$. Since $m(\theta, \cdot)$ is measurable, $m(\Theta, \cdot)$ is measurable as well. All measurability issues in Section 4 are taken care of by this fact and the completeness of P.

Similarly, from (A.2), we see that

$$\Lambda_n(U, \lambda, x_1^n) := \sup_{\theta\in U} \Lambda_n(\theta, \lambda, x_1^n) = \sup_{\theta\in G_U} \Lambda_n(\theta, \lambda, x_1^n), \qquad \lambda \leq 0,$$

for any countable dense $G_U \subset U$ and any $U \subset \Theta$. Since $\Lambda_n(\theta, \lambda, \cdot)$ is measurable, $\Lambda_n(U, \lambda, \cdot)$ is also. All measurability issues in Section 3 are taken care of by this fact and the completeness of P.

Fix $\lambda \leq 0$. Notice that $\Lambda_n(\theta, \lambda, x_1^n) \leq 0$. Let $\theta_m \to \theta$. (A.2) and Fatou's Lemma give

$$\limsup_{m\to\infty} \Lambda_n(\theta_m, \lambda) = \limsup_{m\to\infty} E_P \Lambda_n(\theta_m, \lambda, X_1^n) \leq E_P \Lambda_n(\theta, \lambda, X_1^n) = \Lambda_n(\theta, \lambda).$$

So $\Lambda_n(\cdot, \lambda)$ is u.sc. for $n < \infty$.

This implies that Λ_n^* is l.sc. for $n < \infty$ because it is a supremum of l.sc. functions (namely $\lambda D - \Lambda_n(\cdot, \lambda)$).

Proof of (A.1). Fix $x \in S$, $\epsilon > 0$ and $\theta, (\theta_n)_{n\geq 1} \in \Theta$ such that $\theta_n \to \theta$. Define $A := \{y \in T : \rho(x,y) < m(\theta,x) + \epsilon\}$. By definition, $Q_{\theta,1}(A) > 0$ and we have

$$\liminf_{n\to\infty} Q_{\theta_n,1}(A) \geq Q_{\theta,1}(A) > 0. \tag{A.3}$$

The first inequality is actually an equality in situation (2.1). In situation (2.2) $\rho(x,\cdot)$ is continuous, which means A is open and the first inequality is a well known property of weak convergence of probability measures [14][p.311].

(A.3) implies that

$$\limsup_{n\to\infty} m(\theta_n, x) \leq m(\theta, x) + \epsilon.$$

Since ϵ was arbitrary, this establishes the u.sc. of $m(\cdot, x)$. □

Proof of (A.2). Fix $x_1^n \in S_n$, $\lambda \leq 0$ and $\theta, (\theta_m)_{m\geq 1} \in \Theta$ such that $\theta_m \to \theta$. $e^{\lambda n \rho_n(x_1^n,\cdot)}$ is bounded and in situation (2.2) it is also continuous, so

$$E_{\theta_m} e^{\lambda n \rho_n(x_1^n, Y_1^n)} \to E_\theta e^{\lambda n \rho_n(x_1^n, Y_1^n)}$$

and $\Lambda_n(\cdot, \lambda, x_1^n)$ is continuous. □

A.2 Convexity

We first prove a convexity result used as justification for (3.8). In the notation of the last section, $\Lambda_n(\theta, \cdot, x_1^n)$ is a log moment generating function and is convex [6]. So $\sup_{\theta' \in O(\theta,r)} \Lambda_n(\theta', \cdot, x_1^n)$ is convex and we know from the previous section that it is measurable in x_1^n for fixed $\lambda \leq 0$. Expectations and limits preserve convexity. This shows that

$$\lambda D - \lim_{n\to\infty} \frac{1}{n} E_P \left[\sup_{\theta' \in O(\theta,r)} \log E_{\theta'} e^{\lambda n \rho_n(X_1^n, Y_1^n)} \right]$$

is concave in λ for $\lambda \leq 0$ as claimed.

Now we prove some convexity results in θ when Θ is convex. Suppose we have the added assumptions of Section 2.1, namely (2.9). Then

$$\Lambda_1^* \text{ is convex and } \Lambda_1^*(\epsilon\theta + (1-\epsilon)\theta') \leq \Lambda_1^*(\theta) - \log \epsilon \tag{A.4}$$

for any $0 \leq \epsilon \leq 1$ and $\theta, \theta' \in \Theta$.

Proof of (A.4). For $0 \leq \epsilon \leq 1$, the concavity of the logarithm gives

$$\begin{aligned}\Lambda_1(\epsilon\theta + (1-\epsilon)\theta', \lambda) &= E_P\left[\log\left[\epsilon E_\theta e^{\lambda\rho(X_1,Y_1)} + (1-\epsilon) E_{\theta'} e^{\lambda\rho(X_1,Y_1)}\right]\right] \\ &\geq E_P\left[\epsilon \log E_\theta e^{\lambda\rho(X_1,Y_1)} + (1-\epsilon)\log E_{\theta'} e^{\lambda\rho(X_1,Y_1)}\right] = \epsilon\Lambda_1(\theta,\lambda) + (1-\epsilon)\Lambda_1(\theta',\lambda)\end{aligned}$$

and establishes the concavity of $\Lambda_1(\cdot,\lambda)$. This means that $\lambda D - \Lambda_1(\cdot,\lambda)$ is convex. So Λ_1^* is a supremum of convex functions which is convex.

We also have

$$\begin{aligned}\Lambda_1(\epsilon\theta + (1-\epsilon)\theta', \lambda) &= E_P\left[\log\left[\epsilon E_\theta e^{\lambda\rho(X_1,Y_1)} + (1-\epsilon) E_{\theta'} e^{\lambda\rho(X_1,Y_1)}\right]\right] \\ &\geq E_P\left[\log\left[\epsilon E_\theta e^{\lambda\rho(X_1,Y_1)}\right]\right] = \Lambda_1(\theta,\lambda) + \log\epsilon.\end{aligned}$$

This gives

$$\Lambda_1^*(\epsilon\theta + (1-\epsilon)\theta') \leq \sup_{\lambda\leq 0}\left[\lambda D - \Lambda_1(\theta,\lambda) - \log\epsilon\right] = \Lambda_1^*(\theta) - \log\epsilon.$$ □

Acknowledgments

I want to thank I. Kontoyiannis and M. Madiman for many useful comments. I. Kontoyiannis, especially, for invaluable advice and for suggesting the problems that led to this paper. M. Madiman suggested Example 2.2.9.

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

[4] 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.

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

[6] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer, New York, second edition, 1998.

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

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

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

[15] 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.