Assumptions of Physics
Technical Brief
More technical briefs at https://assumptionsofphysics.org/resources/briefs

Entropy is not continuous in standard probability spaces

Gabriele Carcassi for the AoP collaboration

Category: Ensemble spaces - Tags: Entropy, total variation, L^1 topology, Jensen-Shannon divergence, probability measures

Abstract

We show that the topologies coming from L^1, total variation and Jensen–Shannon divergence are all equivalent but do not guarantee the continuity of the entropy. However, a topology that guarantees that mixing (i.e. convex combination) and the entropy are continuous must contain the L^1/total variation/JSD topology.

1 Introduction

Unlike in the finite-dimensional case, multiple topologies are possible in infinite-dimensional linear/convex spaces. This raises the question as to what is the appropriate one for probability spaces, particularly in the context of ensembles with an entropy defined. The L^1 topology for probability densities, the total variation topology for probability measures and the topology generated by open balls of the Jensen–Shannon divergence can be shown to be all equivalent, but they are not enough to guarantee the continuity of the entropy. That is, a sequence of distributions may converge in total variation while its entropy fails to converge. Conversely, we show that a topology that guarantees the continuity of statistical mixing and the entropy guarantees the convergence in L^1/total variation/JSD.

2 Definitions and basic properties

Let (X, Σ, μ) be a measure space. Let $S(\rho) = -\int_X \rho \log \rho \, d\mu$ be the entropy of a probability measure (i.e. ρ is the Radon–Nikodym derivative of a probability measure absolutely continuous with respect to μ). Let $P_\mu = \left\{\rho \in L^1(\mu) \,\middle|\, \rho \geq 0,\ \int_X \rho \, d\mu = 1,\ S(\rho) \in \mathbb{R}\right\}$ be the space of μ-densities with finite entropy.[1]

We now want to compare three different convergence criteria that give rise to the same topology. Considering the probability densities as functions, the L^1 distance between densities is

$$d_1(\rho, \sigma) = \|\rho - \sigma\|_{L^1(\mu)} = \int_X |\rho - \sigma| \, d\mu. \tag{1}$$

[1] Logarithms are taken in base 2 unless otherwise stated.

If p_ρ and p_σ are the corresponding probability measures, then the total variation between the probability measures is given by

$$\|p_\rho - p_\sigma\|_{\mathrm{TV}} = \frac{1}{2}\|\rho - \sigma\|_{L^1(\mu)}. \tag{2}$$

Thus L^1 convergence of densities is equivalent to convergence in total variation of the corresponding measures. The Jensen–Shannon divergence between the two densities is given by

$$\mathrm{JSD}(\rho, \sigma) = S\left(\frac{\rho + \sigma}{2}\right) - \frac{1}{2}S(\rho) - \frac{1}{2}S(\sigma). \tag{3}$$

We can show that this too generates the same topology.

Lemma 4. *For $u \in [-1, 1]$, define*

$$f(u) = \frac{1}{2}\left[(1+u)\log(1+u) + (1-u)\log(1-u)\right], \tag{5}$$

with the convention $0 \log 0 = 0$. Then

$$\frac{u^2}{2\ln 2} \le f(u) \le |u|. \tag{6}$$

Proof. The function f is even, with $f(0) = 0$ and $f(1) = 1$. For $u \in (-1, 1)$,

$$f''(u) = \frac{1}{(1-u^2)\ln 2}. \tag{7}$$

Therefore $f''(u) \ge \frac{1}{\ln 2}$. Since $f(0) = f'(0) = 0$, this gives

$$f(u) \ge \frac{u^2}{2\ln 2}. \tag{8}$$

Moreover, $f''(u) \ge 0$, so f is convex. Hence on $[0, 1]$, f lies below the chord joining $(0, 0)$ and $(1, 1)$:

$$f(u) = f((1-u)0 + u1) \le (1-u)f(0) + uf(1) = u. \tag{9}$$

By evenness, this gives $f(u) \le |u|$ on all of $[-1, 1]$. □

Proposition 10 (Jensen–Shannon and total variation generate the same topology)**.** *Let $\rho, \sigma \in P_\mu$. Then*

$$\frac{1}{8\ln 2}\|\rho - \sigma\|^2_{L^1(\mu)} \le \mathrm{JSD}(\rho, \sigma) \le \frac{1}{2}\|\rho - \sigma\|_{L^1(\mu)}. \tag{11}$$

Consequently,

$$\mathrm{JSD}(\rho_n, \rho) \to 0 \quad \Longleftrightarrow \quad \|\rho_n - \rho\|_{L^1(\mu)} \to 0. \tag{12}$$

Equivalently, the topology generated by the Jensen–Shannon divergence is the total variation topology on the associated probability measures.

Proof. Let

$$m = \frac{\rho + \sigma}{2}. \tag{13}$$

Where $\rho + \sigma > 0$, define

$$u = \frac{\rho - \sigma}{\rho + \sigma}, \tag{14}$$

and set $u = 0$ where $\rho + \sigma = 0$. Then $u \in [-1, 1]$ and

$$\rho = m(1 + u), \qquad \sigma = m(1 - u). \tag{15}$$

Substituting into the definition of Jensen–Shannon divergence gives

$$\mathrm{JSD}(\rho, \sigma) = \int_X m f(u)\, d\mu. \tag{16}$$

By the scalar bounds on f,

$$\frac{1}{2\ln 2} \int_X m u^2\, d\mu \le \mathrm{JSD}(\rho, \sigma) \le \int_X m|u|\, d\mu. \tag{17}$$

Since $m\, d\mu$ is a probability measure, Cauchy–Schwarz gives

$$\begin{aligned} \left(\int_X m|u|\, d\mu \right)^2 &= \left(\int_X \sqrt{m} \left(\sqrt{m}|u| \right)\, d\mu \right)^2 \le \left(\int_X \sqrt{m}^2\, d\mu \right) \left(\int_X \left(\sqrt{m}|u| \right)^2\, d\mu \right) \\ &= \left(\int_X m\, d\mu \right) \left(\int_X m u^2\, d\mu \right) = \int_X m u^2\, d\mu, \end{aligned} \tag{18}$$

Moreover,

$$\int_X m|u|\, d\mu = \int_X \frac{\rho + \sigma}{2} \frac{|\rho - \sigma|}{\rho + \sigma}\, d\mu = \frac{1}{2} \|\rho - \sigma\|_{L^1(\mu)}. \tag{19}$$

Putting it all together, we have

$$\begin{aligned} \frac{1}{2\ln 2} \left(\int_X m|u|\, d\mu \right)^2 \le \mathrm{JSD}(\rho, \sigma) \le \int_X m|u|\, d\mu \\ \frac{1}{8\ln 2} \|\rho - \sigma\|^2_{L^1(\mu)} \le \mathrm{JSD}(\rho, \sigma) \le \frac{1}{2} \|\rho - \sigma\|_{L^1(\mu)}. \end{aligned} \tag{20}$$

The two inequalities imply the equivalence of Jensen–Shannon convergence and L^1 convergence. Since L^1 convergence of densities is equivalent to total variation convergence of the associated probability measures, the Jensen–Shannon topology is the total variation topology. □

We want to stress that, given the measure-theoretic construction, all these results apply uniformly to both countable discrete spaces with counting measure and continuous dominated measure spaces.

3 Failure of entropy continuity

We now show that, whenever the measure space contains finite-measure sets of arbitrarily large measure, entropy is not continuous with respect to the L^1 topology.

Theorem 21 (Entropy is not L^1-continuous on spaces with arbitrarily large finite-measure sets)**.** *Let (X, Σ, μ) be a measure space. Suppose there exists a measurable set $B \in \Sigma$ such that*

$$0 < \mu(B) < \infty, \tag{22}$$

and a sequence of measurable sets $A_n \in \Sigma$ such that

$$A_n \cap B = \varnothing, \qquad 1 < \mu(A_n) < \infty, \qquad \mu(A_n) \to \infty. \tag{23}$$

Then the entropy

$$S(\rho) = -\int_X \rho \log \rho \, d\mu \tag{24}$$

is not continuous in the $L^1(\mu)$ topology on finite-entropy probability densities.

Proof. For a measurable set U with $0 < \mu(U) < \infty$, define the uniform probability density on U by

$$u_U = \frac{\mathbf{1}_U}{\mu(U)}. \tag{25}$$

Its entropy is

$$S(u_U) = -\int_U \frac{1}{\mu(U)} \log\left(\frac{1}{\mu(U)}\right) d\mu = \log \mu(U). \tag{26}$$

Let

$$\rho = u_B. \tag{27}$$

Then

$$S(\rho) = \log \mu(B). \tag{28}$$

Define

$$\varepsilon_n = \frac{1}{\log \mu(A_n)}. \tag{29}$$

Since $\mu(A_n) \to \infty$, we have $\varepsilon_n \to 0$. Passing to a tail of the sequence if necessary, we may assume $\varepsilon_n \in (0, 1)$ for all n.

Now define

$$\rho_n = (1 - \varepsilon_n) u_B + \varepsilon_n u_{A_n}. \tag{30}$$

Since A_n and B are disjoint,

$$\|\rho_n - \rho\|_{L^1} = \varepsilon_n \|u_{A_n}\|_{L^1} + \varepsilon_n \|u_B\|_{L^1} = 2\varepsilon_n \to 0. \tag{31}$$

Thus $\rho_n \to \rho$ in $L^1(\mu)$.

On the other hand, since the supports are disjoint, entropy satisfies the exact mixture formula

$$S(\rho_n) = h_2(\varepsilon_n) + (1 - \varepsilon_n) S(u_B) + \varepsilon_n S(u_{A_n}), \tag{32}$$

where

$$h_2(\varepsilon) = -\varepsilon \log \varepsilon - (1 - \varepsilon) \log(1 - \varepsilon). \tag{33}$$

Therefore

$$S(\rho_n) = h_2(\varepsilon_n) + (1 - \varepsilon_n) \log \mu(B) + \varepsilon_n \log \mu(A_n). \tag{34}$$

By construction,

$$\varepsilon_n \log \mu(A_n) = 1. \tag{35}$$

Moreover,

$$h_2(\varepsilon_n) \to 0 \tag{36}$$

and

$$(1 - \varepsilon_n) \log \mu(B) \to \log \mu(B). \tag{37}$$

Hence

$$S(\rho_n) \to \log \mu(B) + 1. \tag{38}$$

But

$$S(\rho) = \log \mu(B). \tag{39}$$

Thus $\rho_n \to \rho$ in $L^1(\mu)$, while $S(\rho_n) \not\to S(\rho)$. Hence entropy is not L^1-continuous. □

Remark. The proof only uses the existence of finite-measure sets of arbitrarily large measure, disjoint from a fixed finite-measure set. In the countable discrete case with counting measure, one may take $B = \{1\}$ and A_n to be finite sets of cardinality tending to infinity, disjoint from B. Then the construction reduces to the standard example where a probability mass ε_n is spread uniformly over many points.

Remark. Choosing instead

$$\varepsilon_n = \frac{1}{\sqrt{\log \mu(A_n)}}, \tag{40}$$

we still have

$$\varepsilon_n \to 0, \tag{41}$$

and therefore

$$\rho_n \to \rho \qquad \text{in } L^1(\mu). \tag{42}$$

However,

$$\varepsilon_n \log \mu(A_n) = \sqrt{\log \mu(A_n)} \to \infty. \tag{43}$$

Therefore

$$S(\rho_n) \to \infty \tag{44}$$

while still

$$\rho_n \to \rho \qquad \text{in } L^1(\mu). \tag{45}$$

Thus arbitrarily small L^1 perturbations can carry arbitrarily large entropy.

4 Entropy and mixing force total variation

We now show that requiring entropy and statistical mixing to be continuous forces the topology to contain the L^1/total variation/JSD topology.

Theorem 46 (Entropy continuity and mixing continuity force total variation)**.** *Let τ be a topology on P_μ such that:*

- *statistical mixing (e.g. $(\lambda, \rho, \sigma) \mapsto \lambda\rho + (1-\lambda)\sigma$) is continuous;*
- *entropy (i.e. $S(\rho) = -\int_X \rho \log \rho \, d\mu$) is continuous.*

Then τ contains the topology generated by the Jensen–Shannon divergence. Consequently, τ contains the total variation topology, equivalently the $L^1(\mu)$ topology on densities.

Proof. For $\rho, \sigma \in P_\mu$, the Jensen–Shannon divergence is

$$\mathrm{JSD}(\rho, \sigma) = S\left(\frac{\rho + \sigma}{2}\right) - \frac{1}{2}S(\rho) - \frac{1}{2}S(\sigma). \tag{47}$$

Since statistical mixing is continuous and entropy is continuous, the map

$$(\rho, \sigma) \mapsto \mathrm{JSD}(\rho, \sigma) \tag{48}$$

is continuous. Therefore τ must contain the topology generated by the Jensen–Shannon divergence which, as shown in Prop. 10, is the same as the total variation topology, equivalently the $L^1(\mu)$ topology on densities. □

5 Conclusion

The L^1/total variation topology is the natural strong topology for dominated probability measures. It reduces to the ℓ^1 topology in the countable discrete case and to the $L^1(\mu)$ topology for densities in the general dominated case. The Jensen–Shannon divergence generates this same topology. However, entropy is not continuous in this topology in typical infinite-dimensional spaces. Therefore any ensemble topology that requires both continuous statistical mixing and continuous entropy must refine total variation: Jensen–Shannon or L^1 convergence alone is not sufficient.