On information gain, Kullback-Leibler divergence, entropy production and the involution kernel

A. O. Lopes and J. K. Mengue

March 30, 2022

Abstract

It is well known that in Information Theory and Machine Learning the Kullback-Leibler divergence, which extends the concept of Shannon entropy, plays a fundamental role. Given an *a priori* probability kernel $\hat{\nu}$ and a probability π on the measurable space $X \times Y$ we consider an appropriate definition of entropy of π relative to $\hat{\nu}$, which is based on previous works. Using this concept of entropy we obtain a natural definition of information gain for general measurable spaces which coincides with the mutual information given from the K-L divergence in the case $\hat{\nu}$ is identified with a probability ν on X. This will be used to extend the meaning of specific information gain and dynamical entropy production to the model of thermodynamic formalism for symbolic dynamics over a compact alphabet (TFCA model). Via the concepts of involution kernel and dual potential, one can ask if a given potential is symmetric - the relevant information is available in the potential. In the affirmative case, its corresponding equilibrium state has zero entropy production.

Key words: information gain, Kullback-Leibler divergence, entropy production, Thermodynamic Formalism, symbolic spaces.

AMS Subject Classification: 37D35; 62B10; 60G10.

1 Introduction

The main goal of this paper is to introduce and study the concepts of information gain and entropy production to equilibrium measures in symbolic dynamics over a compact space (rather than finite) alphabet. In this way, the first part of this work lies in the frontier between Information Theory and

Ergodic Theory. In the second part of the paper, we consider such model of thermodynamic formalism for symbolic dynamics over a compact alphabet (which we will abbreviate by TFCA model) in Ergodic Theory (see [25]).

We start by introducing some elements of Information Theory. In Data Compression the Shannon entropy[1], $S(P) = -\sum_{i=1}^{d} p_i \log(p_i)$, of a probability vector $P = (p_1, ..., p_d)$ plays an important role (see [40] and [12] chap. 5). For the benefit of the reader we exhibit introductory examples concerning $S(P)$ in Section 2.

Related to this, in the study of Decision Trees in Machine Learning it is also considered another important concept, called **information gain**. Following [38] (see p. 89-90), for a probability π on $X \times Y = \{1, ..., d\} \times \{1, ..., r\}$ with $x-$marginal $P = (p_1, ..., p_d)$, we define the information gain of π with respect to P as

$$IG(\pi, P) = \underbrace{-\sum_{x=1}^{d} p_x \log(p_x)}_{S(P)} - \underbrace{\sum_{y=1}^{r} q_y \left[-\sum_{x=1}^{d} \frac{\pi_{x,y}}{q_y} \log\left(\frac{\pi_{x,y}}{q_y}\right)\right]}_{H(\pi)}, \tag{1}$$

where $q_y = \sum_x \pi_{x,y}$, that is, $Q = (q_1, ..., q_r)$ is the y-marginal of π. In this expression the number

$$-\left[\sum_{x=1}^{d} \frac{\pi_{x,y}}{q_y} \log(\frac{\pi_{x,y}}{q_y})\right]$$

is the Shannon entropy of the probability obtained from the distribution of π on the line $X \times \{y\}$ and, therefore, $H(\pi)$ is just the weighted mean of these entropies according to Q. Example 7 (in Section 2) will exhibit a concrete interpretation of $IG(\pi, P)$.

Denoting by P and Q the marginals of π, we get that the number $IG(\pi, P)$ can be rewritten as

$$\sum_{x=1}^{d} \sum_{y=1}^{r} \pi_{x,y} \log\left(\frac{\pi_{x,y}}{p_x q_y}\right),$$

which, in Information Theory, is called of **mutual information** (see [12]).

In Ergodic Theory, for the case of the symbolic space

$$\Omega = \Omega^+ = \{1, 2, ..., d\}^{\mathbb{N}} = \{|x_1, x_2, x_3, ...) \,|\, x_i \in \{1, 2, ..., d\}, \forall i \in \mathbb{N}\}, \tag{2}$$

it is considered the Kolmogorov-Sinai entropy for stationary probabilities, that means, probabilities μ on Ω which are invariant by the shift map σ :

[1]we will consider here $\log(x) = \ln(x)$, but any basis could be also used. Furthermore, $0 \log(0) = 0$, by convention.

$\Omega \to \Omega$, $\sigma(|x_1, x_2, x_3, ...)) = |x_2, x_3, x_4, ...)$. The set Ω is a compact metric space, when equipped with the metric $d(|x_1, x_2, x_3, ...), |y_1, y_2, y_3, ...)) = 2^{-n}$, where $n = \min\{i \,|\, x_i \neq y_i\}$, if $x \neq y$. It is a measurable space when equipped with the Borel $\sigma-$algebra $\mathcal{B}$. For any $n \geq 1$, and any fixed symbols $b_1, ..., b_n$ in $\{1, 2, ..., d\}$, we define the cylinder set $|b_1, b_2, ..., b_n] = \{|x_1, x_2, x_3, ...) \in X \,|\, x_1 = b_1, ..., x_n = b_n\}$. A Borel probability μ on Ω is called shift-invariant if it satisfies $\mu(|b_1, b_2, ..., b_n]) = \sum_{i=1}^d \mu(|i, b_1, ..., b_n])$ for any cylinder set. Finally, the Kolmogorov-Sinai entropy of a shift-invariant Borel probability μ is given by

$$h(\mu) = \lim_{n \to \infty} -\frac{1}{n} \sum_{i_1, ..., i_n} \mu(|i_1, ..., i_n]) \log(\mu(|i_1, ..., i_n])). \tag{3}$$

General references for Ergodic Theory are [41] and [43].

In Thermodynamic Formalism (see [37], [43]) it is quite common to consider the concept of pressure for a Lipschitz potential $\phi : \Omega \to \mathbb{R}$, where $\Omega = \{1, 2, ..., d\}^{\mathbb{N}}$. We say that a shift-invariant probability μ_ϕ is the equilibrium probability for the Lipschitz function $\phi : \Omega \to \mathbb{R}$, if

$$P(\phi) := \sup_{\mu \text{ shift-invariant}} [\int \phi \, d\mu + h(\mu)] = \int \phi \, d\mu_\phi + h(\mu_\phi).$$

The number $P(\phi)$ is called the pressure of the potential ϕ.

In ergodic theory for symbolic dynamics appears also the concept of specific information gain which can be used to introduce the entropy production for equilibrium probabilities (see [22] for an introduction to these concepts in a setting compatible with the present work). If μ_ϕ is the equilibrium probability for the Lipschitz function ϕ and if μ is shift-invariant, then, the **specific information gain** of μ with respect to μ_ϕ is given by

$$h(\mu, \mu_\phi) := \lim_n \frac{1}{n} \sum_{|i_1, ..., i_n]} \mu(|i_1, ..., i_n]) \log\left(\frac{\mu(|i_1, ..., i_n])}{\mu_\phi(|i_1, ..., i_n])}\right). \tag{4}$$

Furthermore, from Proposition 1 in [22] (see also [6]), we get

$$h(\mu, \mu_\phi) = \underbrace{[\int \phi \, d\mu_\phi + h(\mu_\phi)]}_{P(\phi)} - [\int \phi \, d\mu + h(\mu)]. \tag{5}$$

In Section 3 we exhibit some analogies between equations (5) and (1). Comparing the equations (3) and (4) it is natural to interpret the specific information gain as a relative entropy. Furthermore, in [6] the value $h(\mu, \mu_\phi)$

is characterized by a variant of the Shannon-McMillan-Breiman Theorem. Indeed, from a result on Section 3.2 of [6] we get the following: consider an ergodic probability μ on Ω, and for a given Lipschitz function $\phi : \Omega \to \mathbb{R}$, consider the corresponding equilibrium probability μ_ϕ. Then, for μ almost every point $x = |x_1, x_2, x_3...) \in \Omega$,

$$\lim_{n\to\infty} \frac{1}{n} \log \left(\frac{\mu\,(|x_1, x_2, ..., x_n])}{\mu_\phi\,(\,|x_1, x_2, ..., x_n]\,)} \right) = h(\mu, \mu_\phi). \tag{6}$$

An interpretation of this expression in the sense of the Statistical Mechanics of non equilibrium is the following: the observed system μ_ϕ is the equilibrium probability for the Lipschitz function ϕ, then, given a random point $x \in \Omega$, its time $n-1$ orbit $\{x, \sigma(x), ..., \sigma^{n-1}(x)\,\}$ describes the dynamical evolution of the system under consideration. For each $n \in \mathbb{N}$, let $\nu_n^x = \dfrac{1}{n}\,(\,\delta_x + \delta_{\sigma(x)} + ... + \delta_{\sigma^{n-1}(x)}\,)$ the associated probability to x at time n (the empirical measure). Then, from Birkhoff's ergodic Theorem, for μ_ϕ a.e. x, we get that $\nu_n^x \to \mu_\phi$, as $n \to \infty$. Denote by μ another ergodic probability (which is not the equilibrium for ϕ). Then, for μ a.e. x, as $n \to \infty$, we get (in the sense of (6))

$$\frac{\mu\,(\,|x_1, x_2, ..., x_n]\,)}{\mu_\phi\,(\,|x_1, x_2, ..., x_n]\,)} \sim e^{n\,h(\mu,\mu_\phi)}.$$

Therefore, the value $h(\mu, \mu_\phi)$ quantifies the asymptotic exponential rate which describes how the dynamical time evolution of the system discriminates between μ_ϕ and μ, when $n \to \infty$.

We will present now the concept of entropy production for equilibrium probabilities on $\{1, ..., d\}^{\mathbb{N}}$ and its relations with the specific information gain. Remember that - for the sake of notation see (2) - we denote Ω by Ω^+. The elements of Ω^+ are denoted by $x = |\,x_1, x_2, ...)$. Consider the space $\Omega^- = \{1, 2, ..., d\}^{\mathbb{N}}$ where any point in the space Ω^- will be written in the form $y = (..., y_3, y_2, y_1|$. In this way any point in $\hat{\Omega} := \Omega^- \times \Omega^+$ will be written in the form $(..., y_3, y_2, y_1|x_1, x_2, x_3, ...) = (y\,|\,x)$.

We consider on $\hat{\Omega}$ the shift map $\hat{\sigma}$ given by

$$\hat{\sigma}((..., y_3, y_2, y_1|x_1, x_2, x_3, ...)) = (..., y_3, y_2, y_1, x_1|x_2, x_3, ...). \tag{7}$$

$\hat{\sigma}$ is a bijection and its natural projection over $\Omega = \Omega^+$ is the shift map σ. The natural projection of the inverse map $\hat{\sigma}^{-1}$ over Ω^- is denoted by σ^-, that is $\sigma^-((...y_4, y_3, y_2, y_1|) = (...y_4, y_3, y_2|$. Observe that (Ω^-, σ^-) can be identified with (Ω^+, σ), via the conjugation $\theta : \Omega^- \to \Omega^+ = \Omega$, which is given by

$$\theta((..., z_3, z_2, z_1|) = |z_1, z_2, z_3, ...). \tag{8}$$

Any σ-invariant probability μ on Ω^+ can be extended (uniquely) to a $\hat{\sigma}$-invariant probability $\hat{\mu}$ on $\Omega^- \times \Omega$. The restriction of $\hat{\mu}$ to Ω^-, denoted by μ^-, is σ^--invariant. By identifying (Ω^-, σ^-) with (Ω, σ), via the conjugation θ and denoting by $\theta_*\mu^-$ the push forward of μ^-, we get

$$\theta_*\mu^-(|a_1, a_2....a_m]) = \mu(|a_m, ..., a_2, a_1]). \tag{9}$$

Finally, the **entropy production** of an equilibrium probability μ on Ω is defined by the specific information gain $e_p(\mu) := h(\mu\,,\,\theta_*\mu^-)$, that is,

$$e_p(\mu) = h(\mu\,,\,\theta_*\mu^-) = \lim_n \frac{1}{n} \sum_{|a_1,...,a_n]} \mu(|a_1, ..., a_n]) \log\left(\frac{\mu(|a_1, ..., a_n])}{\mu(|a_n, ..., a_2, a_1])}\right). \tag{10}$$

Expression (10) can be considered for equilibrium probabilities of potentials in the Walters' family of functions in $\{0,1\}^{\mathbb{N}}$ (see [42]). Indeed, it follows from the symmetry claimed by Corollary 2.3 in [42] that the equilibrium probabilities of potentials in this family have entropy production zero. This family contains probabilities with infinite range dependence and is a natural extension of the set of Markov measures in Thermodynamic Formalism.

In the same way the probabilities on $\{1,2\}^{\mathbb{N}}$ described in [29] has entropy production zero (see Section 2 in [29]).

From now on we consider more general spaces (measurable spaces or compact metric spaces) instead of finite sets or finite alphabet.

In Information Theory, for a measurable space X, the Shannon entropy is extended by the Kullback-Leibler divergence (see [23]) given by

$$D_{KL}(P|\nu) = \begin{cases} \int \log(\frac{dP}{d\nu})\, dP & \text{if } P \ll \nu \\ \\ +\infty, & \text{otherwise} \end{cases},$$

where ν can be interpreted as an a priori probability on X, P is another probability on X and $P \ll \nu$ means that P is absolutely continuous with respect to ν.

The K-L divergence is also used to extend for measurable spaces the information gain or mutual information. If π is absolutely continuous with respect to $P \times Q$, then, denoting by $\frac{d\pi}{dPdQ}$ the Radon-Nikodyn derivative, the mutual information can be expressed in terms of

$$D_{KL}(\pi \,|\, P \times Q) = \int \log\left(\frac{d\pi}{dPdQ}(x,y)\right) d\pi(x,y). \tag{11}$$

From another point of view, in the TFCA model studied in [25], which considers a symbolic dynamic over an alphabet given by a compact metric space M (instead of a finite or enumerable set), it was proposed to consider a relative entropy given by

$$h^\nu(\mu) := -\sup\{\int c(|x_1, x_2, ...))\, d\mu(|x_1, x_2, ...)) \mid \int e^{c(|a,w))}\, d\nu(a) = 1\, \forall w \in M^{\mathbb{N}}\},$$

where ν is an a priori probability on M, μ is a shift-invariant (stationary) probability on $\Omega := M^{\mathbb{N}} = \{|x_1, x_2, x_3, ...)| x_i \in M\ \forall i \in \mathbb{N}\}$ and the functions $c : M^{\mathbb{N}} \to \mathbb{R}$ are necessarily Lipschitz. Variations of this expression appear in [26], [35] and more recently in [24].

In [1] it was proved that h^ν coincides with the specific entropy in Statistical Mechanics, which is related to the D_{KL}. In the present work we propose to rewrite h^ν in terms of a variational characterization of (11) which assures that $h^\nu(\mu)$ is related with D_{KL} in a more direct way than [1]. Precisely, if P is a probability on X and π is a probability on $X \times Y$ with $y-$marginal Q, then from Theorem 13 we obtain that

$$D_{KL}(\pi \mid \nu \times Q) = \sup\left\{\int c(x, y)\, d\pi(x, y) \mid \int e^{c(x,y)} d\nu(x) = 1\, \forall y,\, c \in \mathcal{F}(\pi)\right\}, \quad (12)$$

where $c \in \mathcal{F}(\pi)$ if c is a measurable function such that $\int c\, d\pi$ is well defined. Furthermore, in Theorem 28 we prove that for compact metric spaces X and Y the above supremum can be taken over Lipschitz functions.

We notice that taking $X := M$, $Y := M^{\{2,3,4,5,...\}}$, and identifying Ω with $X \times Y$ by the rule

$$\Omega \ni |x_1, x_2, x_3, x_4, ...) \to (x_1,\ |x_2, x_3, x_4, ...)) \in X \times Y, \quad (13)$$

then, the entropy h^ν can be rewritten as

$$h^\nu(\mu) = -\sup\{\int c(x, y)\, d\mu(x, y) \mid \int e^{c(x,y)}\, d\nu(x) = 1\, \forall y \in Y\},$$

where the supremum is taken over Lipschitz functions. It follows from (12) that the entropy proposed in [25] (and [35]) can be rewritten in terms of D_{KL}. We elaborate more about this issue for the case of the TFCA model in Section 6.

Explicit expressions for the entropy of shift invariant probabilities in the case of a certain family of potentials (called of product type) defined on symbolic spaces where the alphabet is the interval $[0, 1]$ is presented in [36] (for the case of the finite alphabet $\{-1, 1\}$ see [9]).

In [30] the authors analyze the change of the relative entropy (KL-divergence) for Gibbs probabilities under the action of the dual of the Ruelle operator and its relation with the Second Law of Thermodynamics.

In Section 4 we introduce the concept of information gain with respect to a probability kernel.

Definition 1. *Let X and Y be measurable spaces. We will call of a **probability kernel** any family $\hat{\nu} = \{\hat{\nu}^y \,|\, y \in Y\}$ of probabilities on $X \times Y$, such that,*

1) $\forall\, y \in Y$, we have $\hat{\nu}^y(X_y) = 1$, where $X_y = \{(x, y) \,|\, x \in X\}$,

2) $\forall A \subset X \times Y$ measurable, we have that $y \to \nu^y(A)$ is measurable.

If $\hat{\nu}$ is a probability kernel and Q is a probability on Y, then we can define a probability $\pi = \hat{\nu}\, dQ$ on $X \times Y$ by $\pi(A) = \int \hat{\nu}^y(A)\, dQ(y)$. It means

$$\int f(x, y)\, d\pi(x, y) := \int f(x, y)\hat{\nu}^y(dx)dQ(y). \tag{14}$$

The right-hand side of the above expression can be seen as a Rokhlin's disintegration of π.

Following [24] we consider for the present setting the definition of entropy described below.

Definition 2. *Let X and Y be measurable spaces. We define the entropy of any probability π on $X \times Y$ relative to the probability kernel $\hat{\nu}$ as*

$$H^{\hat{\nu}}(\pi) = -\sup\{\int c(x, y)\, d\pi(x, y) \,|\, \int e^{c(x,y)}\hat{\nu}^y(dx) = 1\,\forall y,\, c \in \mathcal{F}(\pi)\}.$$

Finally, we will introduce and study the following meaning of information gain, which is able to extend all the different notions of information gain considered in this paper.

Definition 3. *Let X and Y be measurable spaces. We define the **information gain** of a probability π on $X \times Y$ relative to the probability kernel $\hat{\nu}$, by*

$$IG(\pi, \hat{\nu}) = -H^{\hat{\nu}}(\pi).$$

If π has a y-marginal Q then from Theorem 13 of Section 4 we have that

$$IG(\pi, \hat{\nu}) = D_{KL}(\pi \,|\, \hat{\nu}\, dQ).$$

Following [24], it is possible to remark that there are natural extensions of the above concepts if we replace $X \times Y$ by a measurable space M with a measurable partition (which induces an equivalence relation) and probability kernels by general transverse functions. On the other hand, the above information gain is related with the generalized conditional relative entropy (see chap. 5 in [18]) in the following sense: If π has $y-$marginal Q and π_0 has a disintegration $\pi_0 = \hat{\nu}\, d\tilde{Q}$ then the conditional relative entropy of π with respect to π_0 is given by $D_{KL}(\pi|\hat{\nu}\, dQ)$ and therefore its value coincides with $IG(\pi, \hat{\nu})$ above defined.

In addition to being connected with the previous work [24], we remark that there are at least two natural reasons for our preference of the above approach using probability kernels instead of a probability π_0. The first one is because the conditional relative entropy, as above defined, does not consider π_0 totally, but only $\hat{\nu}$, while the $y-$marginal $\tilde{Q}$ of π_0 is replaced by Q. So it is not necessary to compute a disintegration (or a regular conditional probability measure) $\hat{\nu}$ for π_0, but just to consider a priori such probability kernel $\hat{\nu}$ instead of π_0. In this case, it is not necessary to impose more restrictions on the spaces which would be necessary in order to get a disintegration. The second one is that for a fixed probability π_0 the regular conditional probability measure $\hat{\nu}$ is in general not unique. If $\hat{\nu}$ and $\hat{\mu}$ are different probability kernels satisfying

$$\pi_0 = \hat{\nu}\, d\tilde{Q} = \hat{\mu}\, d\tilde{Q}$$

then the conditional relative entropy may not be well defined and more assumptions are required, as for example $Q \ll \tilde{Q}$. In Section 5 we consider compact metrical spaces X and Y and show that, under some assumptions on π_0, an information gain (or, conditional relative entropy) $IG(\pi, \pi_0)$ can be naturally introduced.

The above generalized information gain will be used in Section 6 to introduces the concept of information gain in the TFCA model. Finally, we will be able to introduce the definition of entropy production in the TFCA model (see Section 7). In our reasoning, it will be natural to use as a tool the concept of involution kernel (for references about the involution kernel with setting compatible with the present paper see [2], [25] and [28]).

In Section 7 we define the concepts of involution kernel and dual potential (see Definition 44) and in Theorem 54 we present a formula for the entropy production of an equilibrium probability μ_A in terms of A and a dual potential A^-, that is $e_p(\mu_A) = \int (A) - (A^- \circ \theta^{-1})\, d\mu_A$. It follows that if the potential is symmetric, then the corresponding equilibrium probability has entropy production zero (Theorem 55).

An important issue that we would like to emphasize is the following: given a potential A, the natural way to obtain its equilibrium measure is to find eigenvalues, eigenfunctions and eigenmeasures for the Ruelle operator. Given a potential A, to find an involution kernel and its dual potential is a much simpler matter in several cases (and there is no need to find an eigenvalue). In this way, if we manage to show that the potential is symmetric, we are guaranteed that its equilibrium measure is symmetric. This is a great simplification of the problem in these cases.

For example, in Section 5 in [8] the authors present examples (product type potentials in Example 2 and Ising type potentials in Example 3) of potentials in $\{-1,1\}^{\mathbb{N}}$ that are symmetric by exhibiting the explicit expression of the involution kernel. From Theorem 55 it follows that the equilibrium probabilities for such potentials have zero entropy production. Furthermore, we can extend the reasoning of Example 2 in Section 5 in [8] (which considers the alphabet $\{-1,1\}$) in order to get an explicit involution kernel for a Lipschitz potential $A : [0,1]^{\mathbb{N}} \to \mathbb{R}$ (in this case the alphabet is the compact set $[0,1]$). Indeed, consider a sequence $a_n > 0$, $n \geq 1$, such that $\sum_{i\geq 1}\sum_{j>i} a_j < \infty$, and for $x = |x_1, x_2, x_3...) \in [0,1]^{\mathbb{N}}$, we define $A(x) = \sum_{n=1}^{\infty} a_n x_n$ (it is a product type potential). Consider $W : [0,1]^{\mathbb{N}\times\mathbb{N}} \to \mathbb{R}$, given by

$$W(y|x) = \sum_{i=1}^{\infty}[(x_i + y_i)(\sum_{j>i} a_j)] = \sum_{i=1}^{\infty}(x_i + y_i)(a_{i+1} + a_{i+2} + ...).$$

Then, one can show that W is an involution kernel and A is symmetric. A particular example is when $a_n = 2^{-n}$, $n \geq 1$, in which case A is of Lipchitz class and its equilibrium measure has entropy production zero. Thermodynamic Formalism and equilibrium measures for product type potentials on $[0,1]^{\mathbb{N}}$ are studied in [36].

As another example, in Section 10.4.1 in [19] explicit expressions for the involution kernel of potentials in the Walters' family are calculated (see also [20]).

Results related to the role of the entropy production (the fluctuation Theorem and the detailed balance condition) in problems in Physics and Dynamics can be found in [16], [22], [32], [39] and [4]. A concrete example of a system where the entropy production plays an important role is presented in [13]: a classical gas confined in a cylinder by a movable piston (see the first page of [13]).

We would like to thank L. Cioletti for helpful comments during the writing of this paper.

2 Simple examples in information theory

The reader interested in the results of a more theoretical nature can skip this section without prejudice to what follows.

In this section we consider the Shannon entropy $S(P)$ and information gain $IG(\pi, P)$ via worked examples. We believe that this short presentation will be helpful for mathematicians that do not have much familiarity with these concepts. Nice general references on Information Theory and information gain are [12] and [31]. We start by considering the Shannon entropy which is sometimes alternatively called mean information.

The number $S(P)$ can be interpreted (taking basis 2 for the logarithm) as a lower bound for the average of questions of type "yes or no" which are necessary in order to analyze the statistics of a symbol picked at random - according to the probability distribution $P = (p_1, ..., p_d)$ - on the finite alphabet $\{1, ..., d\}$. From the sequence of answers to successive questions - of a certain type - one can introduce a binary code on the set $\{1, ..., d\}$, where 0 corresponds to "yes" and 1 to "no" (see [12] chap. 5).

Example 4. *Suppose that a box has balls of 4 possible different colors. Two people will play a game with the following rules: one ball is picked off the box by one of them and the other person must discover the color of this ball by making questions of the type "yes or not".*

If this game is repeated several times, the balls are picked randomly according with the probability $P = (p_1, p_2, p_3, p_4) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ *and the strategy used for the questions is optimal, what is the mean value of the number of questions which are necessary?*

We will replace the colors with symbols of the set $\{1, 2, 3, 4\}$. *One can consider the following strategy of questions:*
Q1: is the picked symbol 1 or 2?
- with the answer "yes" it can be considered the question Q2: is the symbol 1?
- with the answer "no" it can be considered the question Q2': is the symbol 3?
Using this strategy it is necessary exactly two questions in order to discover the symbol (color) which was taken. It coincides with the Shannon entropy (the mean information)

$$S(P) = -\sum_{i=1}^{4} \frac{1}{4} \log_2(\frac{1}{4}) = 2.$$

Observe that the set of symbols $\{1, 2, 3, 4\}$ *can be encoded as the answers* (yy, yn, ny, nn). *Replacing* y *by* 0 *and* n *by* 1 *we can encode* $\{1, 2, 3, 4\}$ *as* $(00, 01, 10, 11)$ *in binary expansion, which is optimal.*

Example 5. *Proceeding as in above example, but now assuming that the colors of the balls are picked randomly according to the probability* $P = (p_1, p_2, p_3, p_4) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$*, one can use the following strategy of questions:*

Q1: is the symbol (color) 1? (with probability (frequency) $\frac{1}{2}$ this unique question solves the problem)
- with the answer "yes" we finish.
- with the answer "no" we consider the question Q2: is the symbol 2?
- with the answer "yes" we finish.
- with the answer "no" again, we then consider the question Q3: is the symbol 3?
If this game is repeated several times, using this strategy the mean number of questions is:

$$(1\,question)\frac{1}{2} + (2\,questions)\frac{1}{4} + (3\,questions)\frac{1}{4} = \frac{7}{4}.$$

It coincides with the Shannon entropy (mean information)

$$S(P) = -[\frac{1}{2}\log_2(\frac{1}{2}) + \frac{1}{4}\log_2(\frac{1}{4}) + \frac{1}{8}\log_2(\frac{1}{8}) + \frac{1}{8}\log_2(\frac{1}{8})] = \frac{7}{4}.$$

In this case $\{1, 2, 3, 4\}$ *can be encoded as* $\{0, 10, 110, 111\}$ *in binary expansion, being this one optimal.*

Example 6. *Proceeding as above and supposing that there are only two colors of balls which are picked randomly according with the probability* $P = (p_1, p_2) = (\frac{2}{3}, \frac{1}{3})$ *one can consider the following question:*
Q1: is the color (symbol) 1?
With this strategy, the mean number of questions is exactly 1 which is bigger than the Shannon entropy $S(P) \approx 0,918$. *In this case* $\{1, 2\}$ *can be encoded as* $\{0, 1\}$ *in binary expansion.*

We refer to [12] chap. 5 for a more complete discussion of the topic. Our intention above was just to illustrate - with introductory and simple examples - the fact that the Shannon entropy is as a lower bound for the average number of questions and how one can introduce a binary code for a set of symbols $\{1, ..., d\}$.

From now on we will discuss an example concerning the Information Gain $IG(\pi, P)$ (or mutual information). We refer to [38] (see p. 89-90) for a more detailed discussion of this topic in the context of decision trees in Machine Learning.

Example 7. *Consider - in a similar way as before - a box with a collection of 100 objects, being 30 of them of the color blue and 70 of them of the color red. It's also known that:*

a. 10 of the blue objects are balls and 20 of them are cubes
b. 45 of the red objects are balls and 25 of them are cubes.

Considering all this set of information we can construct probabilities P and π in the following way:

$$\begin{array}{c} 30\ blue \\ 70\ red \end{array} \rightarrow P = \begin{pmatrix} 0.3 \\ 0.7 \end{pmatrix} \qquad \begin{array}{ccc} & balls & cubes \\ blue & 10 & 20 \\ red & 45 & 25 \end{array} \rightarrow \pi = \begin{pmatrix} 0.10 & 0.20 \\ 0.45 & 0.25 \end{pmatrix}.$$

We consider that π is defined in a Cartesian product $X \times Y$ and has $x-$marginal $P = (\frac{30}{100}, \frac{70}{100})$ (adding in the lines of π) and $y-$marginal $Q = (\frac{55}{100}, \frac{45}{100})$ (adding in the rows of π).

We will consider two kinds of different games.

Game one: *One object is randomly picked of the box and we shall discover its color by asking questions of the type yes or no. In this case the Shannon's entropy, or mean information, is equal to*

$$S(P) = -\left[\frac{30}{100}\log(\frac{30}{100}) + \frac{70}{100}\log(\frac{70}{100})\right].$$

Game two: *In this game - in a similar way as in game one - we have the same goal. However, in the present game, after the object was picked we receive a partial information about the result, which is: "it is a cube" or "it is a ball".*

In this game, with probability (or, frequency) $\frac{55}{100}$, the information to be received it will be that it was picked a ball. Using this information we must concentrate our attention for such class of objects and so the colors are distributed according to the probability $(\frac{10}{55}, \frac{45}{55})$. Similarly, with probability (frequency) $\frac{45}{100}$, the information received will be that a cube was picked. In this case, we consider the colors distributed according to the probability $(\frac{20}{45}, \frac{25}{45})$. Therefore, the mean information in this game is given by a weighted mean of two Shannon's entropies, that is,

$$H(\pi) = -\frac{55}{100}\left[\frac{10}{55}\log(\frac{10}{55}) + \frac{45}{55}\log(\frac{45}{55})\right] - \frac{45}{100}\left[\frac{20}{45}\log(\frac{20}{45}) + \frac{25}{45}\log(\frac{25}{45})\right].$$

Finally, we observe that the information gain $IG(\pi, P)$ given in (1) is the difference between the mean information in game one and the mean information in game two,

$$IG(\pi, P) = S(P) - H(\pi).$$

3 Relations between the different concepts of information gain

In this section we propose to explain a relation between the information gain given by (1) and the specific information gain given by (4) and (5). In Thermodynamic Formalism a Lipschitz potential $\phi : \Omega \to \mathbb{R}$ is called **normalized** if $\sum_{x_1} e^{\phi(|x_1,x_2,x_3,...))} = 1, \forall\, x_2, x_3, ... \in \{1, ..., d\}$. In this case $P(\phi) = 0$ and furthermore

$$e^{\phi(|x_1,x_2,x_3,...))} = \lim_{n\to\infty} \frac{\mu_\phi(|x_1, x_2, ..., x_n])}{\sum_i \mu_\phi(|i, x_2, ..., x_n])}$$

(see [37], cor. 3.2.2). We will call, for any shift-invariant probability μ, **Jacobian** of μ the function

$$J^\mu(|x_1, x_2, ...)) := \lim_{n\to\infty} \frac{\mu(|x_1, x_2, ..., x_n])}{\sum_i \mu(|i, x_2, ..., x_n])} = \lim_{n\to\infty} \frac{\mu(|x_1, x_2, ..., x_n])}{\mu(|x_2, ..., x_n])},$$

which is defined $\mu-$a.e[2] . In this way, for a normalized potential ϕ, we have that $\log(J^{\mu_\phi}) = \phi$, and (5) can be rewritten as

$$h(\mu, \mu_\phi) = -[\int \log(J^{\mu_\phi})\, d\mu + h(\mu)]. \tag{15}$$

We also remark that from Lemma 7 in [26] the Kolmogorov-Sinai entropy satisfies

$$h(\mu) = -\sup \left\{ \int c\, d\mu \,|\, \begin{array}{c} c \text{ is Lipschitz and} \\ \sum_{x_1} e^{c(|x_1,x_2,x_3...))} = 1 \\ \forall\, x_2, x_3, ... \in \{1, ..., d\} \end{array} \right\}. \tag{16}$$

Furthermore, for a normalized function ϕ we have $h(\mu_\phi) = -\int \phi\, d\mu_\phi$. Therefore, if $\mu = \mu_\psi$ is the equilibrium probability for the normalized function ψ, then (15) can be rewritten as

$$h(\mu_\psi, \mu_\phi) = -[\int \log(J^{\mu_\phi})\, d\mu_\psi - \int \log(J^{\mu_\psi}) d\mu_\psi] = -[\int \phi\, d\mu_\psi - \int \psi d\mu_\psi].$$

In order to explain the relations between $h(\mu, \mu_\phi)$ and the information gain given by (1) we need also to extend (1). For a probability π on the finite

[2]our abstract definition corresponds to the inverse of the usual Jacobian T' for the action of a locally invertible map T and the Lebesgue measure.

set $X \times Y$, we will call $J^\pi(x,y) := \frac{\pi_{x,y}}{\sum_x \pi_{x,y}}$ the **Jacobian** of the probability π (which is defined π-a.e.). Then, we have that $H(\pi)$ given in (1) satisfies

$$H(\pi) = -\sum_{x=1}^{d}\sum_{y=1}^{r} \pi_{x,y} \log(J^\pi(x,y)) = -\int \log(J^\pi)\, d\pi.$$

In Proposition 62 of Appendix Section 8 we will prove (in a similar way as in chap. 3 in [34]) that

$$H(\pi) = -\sup\{\sum_{x,y} f(x,y)\pi_{x,y} \mid \sum_{x\in X} e^{f(x,y)} = 1,\ \forall y\}. \tag{17}$$

For any given probability P on $X = \{1,...,d\}$ and any given probability $\tilde{Q} = (\tilde{q}_1,...,\tilde{q}_r)$ on $Y = \{1,...,r\}$, with $\tilde{q}_i > 0$, $\forall i$, consider the product measure $\pi_0 = P \times \tilde{Q}$ on $X \times Y$. Then,

1. $J^{\pi_0}(x,y) = \dfrac{p_x\tilde{q}_y}{\sum_x p_x\tilde{q}_y} = \dfrac{p_x\tilde{q}_y}{\tilde{q}_y} = p_x$,
2. $S(P) = -\sum_{x,y} (\pi_0)_{x,y} \log(p_x) = -\sum_{x,y} (\pi_0)_{x,y} \log(J^{\pi_0}(x,y)) = H(\pi_0)$,
3. If π is any probability on $X \times Y$ with x-marginal P, then,

$$\begin{aligned} IG(\pi,P) &= S(P) - H(\pi) \overset{2.}{=} H(\pi_0) - H(\pi) \\ &= -[\int \log(J^{\pi_0})\, d\pi_0 + H(\pi)] \\ &= -[\int \log(J^{\pi_0})\, d\pi + H(\pi)], \end{aligned}$$

where the last equality is satisfied because $J^{\pi_0}(x,y) \overset{1.}{=} p_x$ depends only on the first coordinate, and the x-marginal of both probabilities π and π_0 is the probability P.

This allows us to extend the definition of information gain (1) in the following way:

Definition 8. *Let π_0, π be probabilities on $X \times Y$, such that $(\pi_0)_{x,y} > 0, \forall\, (x,y) \in X \times Y$. We define the information gain of π with respect to π_0 by*

$$IG(\pi,\pi_0) = -[\int \log(J^{\pi_0})\, d\pi + H(\pi)]. \tag{18}$$

The expression of the information gain $IG(\pi,\pi_0)$ and the expression of the specific information gain $h(\mu,\mu_\phi)$ given in (15) are similar. Furthermore, the Jacobians and both variational characterizations of $h(\mu)$ and $H(\pi)$ given in (16) and (17) are alike.

We believe that the next remark can help the reader in understanding why the introduction of probability kernels is natural to replace finite sets by measurable sets (in the study of Information gain).

Remark 9. *In the right hand side of above expression* (18) *does not appear* π_0 *but only* J^{π_0}*. If* $\tilde{Q}$ *is the* y *marginal of* π_0 *then, by definition of* J^{π_0}*, for any function* f*,*

$$\sum_{x,y} f(x,y)\pi_0(x,y) = \sum_{x,y} f(x,y)J^{\pi_0}(x,y)\tilde{Q}(y).$$

Furthermore, for each fixed y *we have that* $\sum_x J^{\pi_0}(x,y) = 1$*. Therefore, for each fixed* y*, we can interpret* J^{π_0} *as a probability in* $X \times \{y\}$*. In this way* J^{π_0} *is a probability kernel in the sense of Definition 1. A similar remark is true for* (15)*.*

4 Information Gain and probability kernels

Our purpose in this section is to extend the definition of information gain $IG(\pi, \pi_0)$, given by (18), for the case when X and Y are measurable spaces. As we will see, a natural way of to extend (18) is by considering probability kernels and the notion of entropy given in [24]. This entropy is an extension of that previously introduced in [25] and [35] for compact spaces using an a priori probability. In [35] an entropy has been introduced for holonomic probabilities associated with iterated function systems (IFS), but we point out that the expression of the entropy in [35] does not use such structures. It may seem surprising, but it is related, by a variational principle, with the spectral radius of a transfer operator which is defined from the IFS. As we will see below the entropy considered in this section does not consider any dynamics.

From now on we consider $\sigma-$algebras $\mathcal{A}$ on X and $\mathcal{B}$ on Y and the product $\sigma-$algebra on $X \times Y$. If $c : X \times Y \to \mathbb{R}$ is measurable then, for any fixed $y \in Y$, the function $c_y(x) := c(x,y)$ defined on X is measurable (see [5] Theorem 6.7).

In order to make an identification with the setting of [24] we consider in the space $X \times Y$ the equivalence relation $(x_1, y_1) \sim (x_2, y_2)$, if and only if, $y_1 = y_2$. So the equivalence classes are the horizontal lines of $X \times Y$. The so called transverse functions in [24] corresponds to probability kernels in the present setting (see Definition 1).

If X and Y are metric spaces, as considered in [24], then condition 1) in Definition 1 is equivalent to say that the probability ν^y has support on X_y. Another equivalent way of defining a probability kernel is as a family of probabilities $\hat{\nu}^y$ on X such that for any measurable set $B \subset X$ we have that $y \to \nu^y(B)$ is measurable (to prove this statement, just adapt the reasoning of Theorem 6.4 in [5] to the current setting).

The next definition was taken from the reasoning of [24].

Definition 10. *We define the entropy of any probability π on $X \times Y$ relative to the probability kernel $\hat{\nu}$ as*

$$H^{\hat{\nu}}(\pi) = -\sup\{\int c(x,y)\, d\pi(x,y) \mid \int e^{c(x,y)} \hat{\nu}^y(dx) = 1\, \forall y,\, c \in \mathcal{F}(\pi)\},$$

where $\mathcal{F}(\pi)$ is the set of measurable functions with a well-defined integral with respect to π.

It follows from Lemma 16 below that we can take the above supremum over functions c which are bounded below. Such functions belongs to $\mathcal{F}(\pi)$, even though we may have $\int c\, d\pi = +\infty$.

Usually, we also fix a probability ν on X satisfying $\text{supp}(\nu) = X$, which we call an *a priori* probability on X. Given an a priori probability ν on X and considering the identification of X and X_y, we can consider the a priori probability kernel $\hat{\nu}$ as given by $\hat{\nu}^y(dx) = \nu(dx)$ (for condition 2. see Theorem 6.4 in [5]). In this case we write $\hat{\nu} \equiv \nu$, and we denote $H^{\hat{\nu}}(\pi)$ also by $H^{\nu}(\pi)$, which will be given by

$$H^{\nu}(\pi) = -\sup\{\int c(x,y)\, d\pi(x,y) \mid \int e^{c(x,y)} d\nu(x) = 1\, \forall y,\, c \in \mathcal{F}(\pi)\}.$$

Definition 11. *We say that a measurable function $c : X \times Y \to \mathbb{R}$ is $\hat{\nu}$-normalized if*

$$\int e^{c(x,y)} \hat{\nu}^y(dx) = 1\,, \forall y \in Y.$$

If ν is an a priori probability on X, we say that $c : X \to \mathbb{R}$ is $\nu-$normalized, if it is measurable and $\int e^c d\nu = 1$.

Example 12. *If ν is an a priori probability on X and $c : X \times Y \to \mathbb{R}$ is ν-normalized, that is,*

$$\int e^{c(x,y)} d\nu(x) = 1\,, \forall y \in Y,$$

then defining, for each y, the probability $\hat{\nu}^y$ on X_y by $\hat{\nu}^y(dx) := e^{c(x,y)} d\nu(x)$, we get that $\hat{\nu}$ is a probability kernel. It corresponds to the case where all *the probabilities $\hat{\nu}^y$ are densities for the same probability ν. More generally, if $\hat{\nu}$ is a probability kernel and c is $\hat{\nu}-$normalized, then $e^{c(x,y)}\hat{\nu}^y(dx)$ is a probability kernel too. If Q is a probability on Y we get also a probability $\pi := \hat{\nu}\, dQ$ on $X \times Y$ by* (14) *and the right hand side is a disintegration of π with respect to the horizontal lines of $X \times Y$. If $\hat{\nu} \equiv \nu$ we get $\hat{\nu} dQ = d\nu dQ$ is a product measure.*

The function $c = 0$ is $\hat{\nu}-$normalized and therefore $H^{\hat{\nu}}(\pi) \leq 0$. If $\tilde{\nu}$ is not a probability, but just is a finite measure on X, satisfying $\tilde{\nu}(X) = d$ and we define $d\hat{\nu} := \frac{1}{d} d\tilde{\nu}$, then $H^{\tilde{\nu}}(\pi) = H^{\hat{\nu}}(\pi) + \log(d)$, where $H^{\tilde{\nu}}$ is defined in a similar way. Now, taking X and Y as finite sets, $\tilde{\nu}$ as the counting measure on X and applying equation (17), we came to the conclusion that such definition of entropy is a natural extension of the definition of $H(\pi)$.

If P is a probability on X we also define

$$S^{\nu}(P) := -\sup \left\{ \int c(x)\, dP(x) \mid \int e^{c(x)}\, d\nu(x) = 1, \text{ where } c \in \mathcal{F}(P) \right\}.$$

We start by proving the next theorem which shows that the above definitions provide variational characterizations of the Kullback-Leibler divergence. It also shows that $-H^{\hat{\nu}}$ is equivalent to generalized conditional relative entropy (see chap. 5 in [18]) as explained in the Introduction Section.

Theorem 13. *Let P and ν be probabilities on X, π be a probability on $X \times Y$ with $y-$marginal Q and $\hat{\nu}$ be a probability kernel. Then,*

$$S^{\nu}(P) = -D_{KL}(P \mid \nu) \quad \textit{and} \quad H^{\hat{\nu}}(\pi) = -D_{KL}(\pi \mid \hat{\nu}\, dQ).$$

Consequently, if $\hat{\nu} \equiv \nu$, we have $H^{\nu}(\pi) = -D_{KL}(\pi \mid \nu \times Q)$.

The proof will be divided into several lemmas (the last one is Lemma 21 which will finish the proof).

Remark 14. *It is known that (see for example chap.5 of [18])*

$$D_{KL}(P \mid \nu) = \sup \left\{ \int c\, dP - \log(\int e^{c}\, d\nu), \ \textit{where}\, c \in \mathcal{F}(P)\, \textit{and} \int e^{c}\, d\nu < \infty \right\}.$$

It follows that

$$D_{KL}(P \mid \nu) = -S^{\nu}(P) \ \textit{and}\ D_{KL}(\pi \mid \hat{\nu}\, dQ) \geq -H^{\hat{\nu}}(\pi).$$

Anyway we provide a complete proof.

Definition 15. *We say that a measurable function* $J : X \times Y \to [0, +\infty)$ *is a* $\hat{\nu}$ – *Jacobian, if* $\int J(x, y)\, \hat{\nu}^y(dx) = 1$, $\forall y \in Y$.
Given an a priori probability ν *on* X*, we say that a measurable function* $J : X \to [0, +\infty)$ *is a* $\nu-$*Jacobian if* $\int J d\nu = 1$.

If c is $\hat{\nu}-$normalized, then $J = e^c$ is a $\hat{\nu}$-Jacobian. On the other hand, if J is a $\hat{\nu}-$Jacobian and it does not assume the value zero, then $c = \log(J)$ is $\hat{\nu}-$normalized.

Lemma 16.

$$H^{\hat{\nu}}(\pi) = -\sup\left\{\int \log(J(x, y))\, d\pi(x, y) \,|\, J \textit{ is a } \hat{\nu}\textit{-Jacobian}\right\}.$$

Furthermore, we can also take the supremum over positive Jacobians J *satisfying* $\inf\{J(x, y)|x \in X, y \in Y\} > 0$.

Proof. If c is $\hat{\nu}-$normalized then $J = e^c$ is a $\hat{\nu}$-Jacobian. It follows that

$$\sup\left\{\int \log(J(x, y))\, d\pi(x, y) \,|\, J \text{ is a } \hat{\nu}\text{-Jacobian}\right\}$$

$$\geq \sup\left\{\int c(x, y)\, d\pi(x, y) \,|\, c \text{ is } \hat{\nu}\text{-normalized}\right\}.$$

On the other hand, for any fixed $\hat{\nu}$-Jacobian J, let $J_n = \frac{J+\frac{1}{n}}{1+\frac{1}{n}}$. As $J \geq 0$ we have $J_n \geq \frac{1}{n+1}$. The function J_n is also a $\hat{\nu}-$Jacobian. Furthermore,

$$\int \log(J)\, d\pi \leq \liminf_n \int \log(J + \frac{1}{n}) d\pi = \liminf_n \int \log(J_n)\, d\pi.$$

As the function $c_n = \log(J_n)$ is $\hat{\nu}-$normalized and bounded below, this ends the proof. □

Lemma 17. *Let* P *be a probability on* X *and* π *be a probability on* $X \times Y$*, with* $x-$*marginal* P*. If* P *is not absolutely continuous with respect to the a priori probability* ν*, then* $S^\nu(P) = -\infty$ *and* $H^\nu(\pi) = -\infty$.

Proof. If P is not absolutely continuous with respect to ν then there exists a measurable set A, such that, $\nu(A) = 0$ and $P(A) > 0$. For each $\beta > 0$, let $c_\beta : X \to \mathbb{R}$ be the measurable function defined as

$$c_\beta(x) = \begin{cases} 0 & \text{if } x \in X - A \\ \beta & \text{if } x \in A \end{cases}.$$

Then, we have $\int e^{c_\beta(x)}\, d\nu(x) = 1$ and $\int c_\beta(x)\, d\pi(x, y) = \int c_\beta(x)\, dP(x) = \beta P(A)$. As β is arbitrary, we can take $\beta \to +\infty$, and then we get that $S^\nu(P) = -\infty$ and also that $H^\nu(\pi) = -\infty$. □

Observe that $J_0 := \frac{dP}{d\nu}$ is a $\nu-$Jacobian. Let $X_0 = \{x \in X \,|\, J_0(x) > 0\}$. Given any measurable and bounded function $f : X \to \mathbb{R}$, we have

$$\int f\, dP = \int f \cdot J_0\, d\nu = \int f \cdot I_{X_0} \cdot J_0\, d\nu = \int_{X_0} f\, dP.$$

It follows that $P(X_0) = 1$ and $\int f(x)\, dP(x) = \int_{X_0} f(x)\, dP(x)$, for any measurable function f.

Furthermore the integral $\int \log(J_0) dP$ is well defined (it can be $+\infty$) because $\int \log(J_0) dP = \int \log(\frac{dP}{d\nu}) \frac{dP}{d\nu} d\nu$ and the function $x \log(x)$ is bounded below.

The next result shows that $-S^\nu(P)$ is the Kullback-Leibler divergence of P with respect to ν.

Lemma 18. *Let P be a probability on X, such that, $P \ll \nu$. Then,*

$$S^\nu(P) = -D_{KL}(P \,|\, \nu) = -\int \log(\frac{dP}{d\nu})\, dP.$$

Proof. Let $J_0 := \frac{dP}{d\nu}$ and $X_0 := \{x \in X \,|\, J_0(x) > 0\}$. We claim that

$$\int_{X_0} \log(J_0)\, dP = \sup \left\{ \int_{X_0} \log(J(x))\, dP(x) \,|\, J \text{ is a } \nu-\text{Jacobian} \right\}.$$

Indeed, from Lemma 16 we can consider a $\nu-$Jacobian $J : X \to (0, +\infty)$ such that $\inf_{x,y} J(x, y) > 0$. By applying the Jensen's inequality we have

$$\int_{X_0} \log\left(\frac{J}{J_0}\right) dP \leq \log \int_{X_0} \frac{J}{J_0}\, dP = \log \int_{X_0} J\, d\nu \leq \log \int J\, d\nu = 0.$$

This shows that

$$\int_{X_0} \log(J)\, dP \leq \int_{X_0} \log(J_0)\, dP.$$

□

A similar estimate for π will be described by the next result.

Lemma 19. *Assume that there exists a $\hat{\nu}-$Jacobian J on $X \times Y$ satisfying*

$$\iint f(x, y) J_0(x, y)\, \hat{\nu}^y(dx) d\pi(x, y) = \int f(x, y)\, d\pi(x, y), \tag{19}$$

for any measurable function $f : X \times Y \to \mathbb{R}$. Then,

$$H^{\hat{\nu}}(\pi) = -\int \log(J_0)\, d\pi.$$

Proof. The reasoning is similar to the previous case. The set $A_0 = \{(x, y) \in X \times Y | J_0(x, y) > 0\}$ satisfies $\pi(A_0) = 1$. For any $\hat{\nu}-$Jacobian $J : X \times Y \to (0, +\infty)$, satisfying $\inf_{x,y} J(x, y) > 0$ we have

$$\int_{A_0} \log\left(\frac{J}{J_0}\right) d\pi \le \log \int_{A_0} \frac{J}{J_0} d\pi = \log \int \frac{J}{J_0} \cdot I_{A_0} \cdot J_0 \; \hat{\nu}^y(dx) d\pi(x, y)$$

$$= \log \int J \cdot I_{A_0} \; \hat{\nu}^y(dx) d\pi(x, y) \le \log \int J \; \hat{\nu}^y(dx) d\pi(x, y) = 0.$$

□

We will say that a function J_0 satisfying (19) is **a** $\hat{\nu}-$**Jacobian of** π.

Denoting by Q the $y-$marginal of π, the equation (19) can be rewritten as

$$\iint f(x, y) J_0(x, y) \, \hat{\nu}^y(dx) dQ(y) = \int f(x, y) \, d\pi(x, y)$$

and so $J_0(x, y) \, \hat{\nu}^y(dx) dQ(y)$ is a disintegration of π with respect to the horizontal lines of $X \times Y$. Supposing also that ν is an a priori probability on X, and $\hat{\nu} \equiv \nu$, we get that $\pi \ll \nu \times Q$, with $\frac{d\pi}{d\nu dq} = J_0$. Then, under the hypotheses of the above proposition and assuming that $\hat{\nu} \equiv \nu$, we get (see also (11))

$$H^{\nu}(\pi) = -D_{KL}(\pi \,|\, \nu \times Q).$$

Lemma 20. *Let $\hat{\nu}$ be a probability kernel and π be a probability on $X \times Y$ with $y-$marginal Q. Suppose that $\pi \ll \hat{\nu} \, dQ$. Then*

$$H^{\hat{\nu}}(\pi) = -D_{KL}(\pi \,|\, \hat{\nu} \, dQ).$$

Proof. We suppose that $\pi \ll \hat{\nu} \, dQ$ and we denote by J its Radon-Nikodyn derivative. Then, for any measurable function $g : X \times Y \to \mathbb{R}$ we have

$$\iint g(x, y) J(x, y) \, \hat{\nu}^y(dx) dQ(y) = \int g(x, y) \, d\pi(x, y).$$

As Q is the y-marginal of π, taking functions g depending just of the second coordinate, we get

$$\int g(y) [\int J(x, y) \, \hat{\nu}^y(dx)] dQ(y) = \int g(y) \, dQ(y).$$

Then $\int J(x, y) \, \hat{\nu}^y(dx) = \frac{dQ}{dQ} = 1$ for $Q-$a.e. y. Replacing J by 1 in a subset of $X \times Y$ having zero measure with respect to π, we get a measurable Jacobian $\tilde{J}$ satisfying, for any measurable function $g : X \times Y \to \mathbb{R}$,

$$\iint g(x, y) \tilde{J}(x, y) \, \hat{\nu}^y(dx) dQ(y) = \int g(x, y) \, d\pi(x, y).$$

It follows from Lemma 19 that

$$H^{\hat{\nu}}(\pi) = -\int \log(\tilde{J})\, d\pi = -D_{KL}(\pi \,|\, \hat{\nu}\, dQ).$$

□

Lemma 21. *Let $\hat{\nu}$ be a probability kernel and π be a probability on $X \times Y$, with $y-$marginal Q. If π is not absolutely continuous with respect to $\hat{\nu}\, dQ$, then $H^{\hat{\nu}}(\pi) = -\infty$.*

Proof. If π is not absolutely continuous with respect to $\hat{\nu}\, dQ$, then, there exists a measurable set $A \subset X \times Y$, such that, $\int \hat{\nu}^y(A)\, dQ(y) = 0$ and $\pi(A) > 0$. It follows that $\{y \,|\, \hat{\nu}^y(A) \neq 0\}$ is a measurable set on Y satisfying $Q(\{y \,|\, \hat{\nu}^y(A) \neq 0\}) = 0$. The set $X \times \{y \,|\, \hat{\nu}^y(A) \neq 0\}$ is measurable in $X \times Y$ and, as the $y-$marginal of π is Q, we get $\pi(X \times \{y \,|\, \hat{\nu}^y(A) \neq 0\}) = Q(\{y \,|\, \hat{\nu}^y(A) \neq 0\}) = 0$. Let $B = A - (X \times \{y \,|\, \hat{\nu}^y(A)) \neq 0\})$. The set B is measurable and $\pi(B) = \pi(A) > 0$, while $\hat{\nu}^y(B) = 0\, \forall y \in Y$.

For each $\beta > 0$, let $c_\beta : X \times Y \to \mathbb{R}$ be the measurable function defined as

$$c_\beta(x, y) = \begin{cases} 0 & \text{if}\, (x, y) \in B^C \\ \beta & \text{if}\, (x, y) \in B \end{cases}.$$

Then, for each fixed y we have

$$\int e^{c_\beta(x,y)}\, d\hat{\nu}^y(dx) = e^{\beta} \hat{\nu}^y(B) + e^0 \hat{\nu}^y(B^C) = 1.$$

This shows that c_β is $\hat{\nu}-$normalized. The rest of the proof is similar to the reasoning of Proposition 17.

□

The above results end the proof of Theorem 13.

Proposition 22. *Let P be a probability on X satisfying $P \ll \nu$. Consider any probability Q on Y and any probability π on $X \times Y$, with $x-$marginal P. Then, we have:*
1. $H^\nu(\pi) \leq S^\nu(P)$
2. $S^\nu(P) = H^\nu(P \times Q)$.

Proof. The proof of item 1. is a direct consequence of the definitions of S^ν and H^ν because we can consider any measurable function $c : X \to \mathbb{R}$, as a measurable function defined on $X \times Y$ which depends just on the first coordinate.

In order to prove item 2. we consider the function $J(x, y) = \frac{dP}{d\nu}(x)$. This function is a $\nu-$Jacobian of $P \times Q$, then, applying Propositions 18 and 19, we conclude the proof. □

The proof of the next result follows the same reasoning which was used in [25] and [24].

Proposition 23. *The entropy $H^{\hat{\nu}}(\cdot)$ has the following properties:*
1. $H^{\hat{\nu}}$ is concave
2. $H^{\hat{\nu}}$ is upper semi continuous. More precisely, if $\int f d\pi_n \to \int f\, d\pi$, for any measurable and bounded function f on $X \times Y$, then, $\limsup_n H^{\hat{\nu}}(\pi_n) \leq H^{\hat{\nu}}(\pi)$.

Definition 24. *We define the **information gain** of a probability π on $X \times Y$, with respect to the a priori probability kernel $\hat{\nu}$, as*

$$IG(\pi, \hat{\nu}) = -H^{\hat{\nu}}(\pi).$$

If π has marginals P and Q and $\pi \ll P \times Q$, then choosing $\hat{\nu} \equiv P$ we get

$$IG(\pi, P) = -H^P(\pi) = D_{KL}(\pi \,|\, P \times Q),$$

which corresponds to the mutual information given by (11). As $c = 0$ is $\hat{\nu}-$normalized we get $IG(\pi, \hat{\nu}) \geq 0$. Furthermore, $IG(\pi, \hat{\nu}) = 0$, if $d\pi = \hat{\nu}^y(dx)dQ(y)$, for some Q.

The information gain $IG(\pi, P)$ above defined can be computed from $S^\nu(P)$ and $H^\nu(\pi)$, and it "does not depend" on the choice of the *a priori* probability ν on X as the following result shows.

Proposition 25. *Let P be a probability on X, π be a probability on $X \times Y$, with $x-$marginal P and let ν be an a priori probability on X. Assume that $S^\nu(P)$ and $H^\nu(\pi)$ are finite. Then,*

$$IG(\pi, P) = S^\nu(P) - H^\nu(\pi).$$

Proof. By hypothesis there exists $\phi : X \to [0, +\infty)$, a $\nu-$Jacobian of P and $J : X \times Y \to [0, +\infty)$, which is a $\nu-$Jacobian of π. Denoting by Q the $y-$marginal of π, we have $d\pi(x, y) = J(x, y)d\nu(x)dQ(y)$ and $dP(x) = \phi(x)d\nu(x)$. The set $A = \{x \in X \,|\, \phi(x) > 0\}$ satisfies $P(A) = 1$ and, as π has $x-$marginal P, we finally get $\pi(A \times Y) = 1$. So we can write $d\pi(x, y) = \frac{J(x,y)}{\phi(x)} dP(x)dQ(y)$ and therefore

$$S^\nu(P) - H^\nu(\pi) = -\int \log(\phi)dP + \int \log(J)d\pi = -H^P(\pi) = IG(\pi, P).$$

□

Proposition 26. *Let $\hat{\nu}$ be an a priori probability kernel and π be a probability on $X \times Y$. Given a bounded and $\hat{\nu}-$normalized function $\phi_0 : X \times Y \to \mathbb{R}$, consider the a priori probability kernel $\hat{\mu}^y(dx) = e^{\phi_0(x,y)}\hat{\nu}^y(dx)$. Then,*

$$IG(\pi, \hat{\mu}) = -\int \phi_0 \, d\pi + IG(\pi, \hat{\nu}).$$

Proof.

$$IG(\pi, \hat{\mu}) = -H^{\hat{\mu}}(\pi) = \sup\{\int c \, d\pi \mid \int e^{c(x,y)} \, \hat{\mu}^y(dx) = 1 \, \forall y\}$$

$$= \sup\{\int c \, d\pi \mid \int e^{c+\phi_0} \, \hat{\nu}^y(dx) = 1 \, \forall y\}$$

$$= \sup\{\int c - \phi_0 \, d\pi \mid \int e^{c} \, \hat{\nu}^y(dx) = 1 \, \forall y\}$$

$$= -\int \phi_0 \, d\pi - H^{\hat{\nu}}(\pi).$$

□

Corollary 27. *Let ν be an a priori probability on X and π be a probability on $X \times Y$. Given a bounded and $\nu-$normalized function $\phi_0 : X \times Y \to \mathbb{R}$, consider the a priori probability kernel $\hat{\mu}^y(dx) = e^{\phi_0(x,y)}d\nu(x)$. Then,*

$$IG(\pi, \hat{\mu}) = -\int \phi_0 \, d\pi - H^{\nu}(\pi).$$

This last result shows that the above definition of information gain, using probability kernels, is a natural extension of (18) (see also Remark 9). Given a probability Q_0 on Y we can associate a probability π_0 on $X \times Y$ given by $d\pi_0 = e^{\phi_0(x,y)}d\nu(x)dQ_0(y)$. A natural generalization of (18), in principle, could be given by

$$IG(\pi, \pi_0) = -\int \phi_0 \, d\pi - H^{\nu}(\pi), \tag{20}$$

but we remark that in some cases it is not even well defined. Indeed, if $Y = \{0, 1\}$ and $\pi_0 = \nu \times \delta_0$, the functions ϕ_0 and ψ_0 given by $\phi_0(x, y) = 0$ and $\psi_0(x, y) = y \cdot f(x)$, where $f \neq 0$ is a ν-normalized function, provide two different disintegrations of π_0, which are

$$d\pi_0(x, y) = e^{\phi_0(x,y)}d\nu(x)d\delta_0(y) \quad \text{and} \quad d\pi_0(x, y) = e^{\psi_0(x,y)}d\nu(x)d\delta_0(y).$$

If $\pi = \nu \times \delta_1$, then $H^\nu(\pi) = H^\nu(\nu \times \delta_1) = S^\nu(\nu) = 0$, and so

$$-\int \psi_0 \, d\pi - H^\nu(\pi) = -\int \psi_0 \, d\pi = \int f(x) \, d\nu(x)$$

while

$$-\int \phi_0 \, d\pi - H^\nu(\pi) = 0.$$

The problem concerning the extension of (18) for measurable spaces can be solved by using probability kernels. We observe that the right-hand side of (20) contains just π, ϕ_0 and ν. Therefore, we realize that we are not using Q_0 and the associated π_0, but only the probability kernel $\hat{\mu} = e^{\phi_0} d\nu$, which is part of a disintegration of π_0. In this sense, it is natural to define information gain by using probability kernels and to consider $IG(\pi, \hat{\mu})$ instead of trying to define $IG(\pi, \pi_0)$.

5 Entropy and information gain for compact metric spaces

In this section, we consider compact metric spaces X and Y equipped with their respective Borel σ-algebras. We will prove that, in this case, H^ν coincides with the entropy considered in [35], which is defined from a supremum taken over Lipschitz functions instead of measurable functions. This shows that the concept of entropy, as defined in [24] for metric spaces, which is also considered here for measurable spaces, extends the concept of entropy as described in [25] and [35]. This also shows that such entropies are given by simple expressions concerning D_{KL}. In the second part of this section, we also propose to rewrite the information gain by using "special" probabilities π_0 instead of probability kernels $\hat{\nu}$ (see the end of the previous section). This will be coherent with the reasoning of future sections and also extends (5), (15) and (18) from finite sets (finite alphabet) to compact spaces.

Theorem 28. *Suppose that X and Y are compact metric spaces and consider the Borel sigma-algebras in X and Y. Then,*

$$H^\nu(\pi) = -\sup\{\int f(x,y) \, d\pi(x,y) | \int e^{f(x,y)} \, d\nu(x) = 1, \, \forall y, \text{ with} f \text{ Lipschitz}\}.$$

Proof. By definition

$$H^\nu(\pi) = -\sup\{\int f(x,y) \, d\pi(x,y) | \int e^{f(x,y)} d\nu(x) = 1 \, \forall y, \, f \in \mathcal{F}(\pi)\}.$$

We denote

$$h^\nu(\pi) := -\sup\{\int f(x,y)\,d\pi(x,y) \mid \int e^{f(x,y)} d\nu(x) = 1,\ \forall y,\ \text{with}\, f\ \text{Lipschitz}\}.$$

It will be necessary to prove that $H^\nu = h^\nu$.

If $\psi : X \times Y \to \mathbb{R}$ is Lipschitz and $\nu-$normalized, Q is any probability on Y and π_ψ is the probability on $X \times Y$, given by

$$\int f(x,y)\,d\pi_\psi(x,y) := \iint f(x,y) e^{\psi(x,y)}\, d\nu(x) dQ(y), \text{ for } f \text{ measurable,}$$

then, e^ψ is a ν-Jacobian of π_ψ. It follows in this case, that

$$H^\nu(\pi_\psi) = -\int \psi(x,y)\,d\pi_\psi(x,y) = h^\nu(\pi_\psi).$$

Suppose by contradiction there exists a probability η on $X \times Y$ such that $-H^\nu(\eta) > -h^\nu(\eta)$. Consequently, $-h^\nu(\eta) \neq +\infty$.

First, we claim that there exists a Lipschitz function $\varphi : X \times Y \to \mathbb{R}$ such that, for any probability π on $X \times Y$,

$$\int \varphi\, d\eta + h^\nu(\eta) > \int \varphi\, d\pi + H^\nu(\pi).$$

The proof of this claim follows the same reasoning of the proof of Theorem 3 in [1] (see also [14] chap. 1). We consider the weak* topology on the space of finite signed-measures and we extend H^ν and h^ν as the value $-\infty$, if π is not a probability.

As $-H^\nu$ is convex, non negative and lower semi-continuous, its epigraph $\text{epi}(-H^\nu) = \{(\pi,t) \mid -H^\nu(\pi) \le t\}$ is convex and closed. As $(\eta, -h^\nu(\eta)) \notin \text{epi}(-H^\nu)$, it follows from Hahn-Banach Theorem that there is $c \in \mathbb{R}$ and a linear functional

$$(\pi, t) \to \int g\, d\pi + at,$$

where g is a fixed continuous function and $a \in \mathbb{R}$ is fixed, such that, for any $(\pi,t) \in \text{epi}(-H^\nu)$ we have

$$\int g\, d\eta - a h^\nu(\eta) < c < \int g\, d\pi + at.$$

Observe that necessarily $a > 0$. We denote $\varphi = -\frac{g}{a}$. If a probability π satisfies $-H^\nu(\pi) < +\infty$, then $(\pi, -H^\nu(\pi)) \in \text{epi}(-H^\nu)$ and finally we get

$$\int \varphi\, d\eta + h^\nu(\eta) > -\frac{c}{a} > \int \varphi\, d\pi + H^\nu(\pi).$$

Even in the case $-H^{\nu}(\pi) = +\infty$ these inequalities remain valid.

Finally, as the set of Lipschitz functions is dense in the set of continuous functions (in the uniform convergence) we can assume that for a Lipschitz function φ we have

$$\int \varphi \, d\eta + h^{\nu}(\eta) > \int \varphi \, d\pi + H^{\nu}(\pi),$$

for any probability π. This finishes the proof of the claim.

Let Q be the y-marginal of η and $\tilde{\varphi}(y) = \log(\int e^{\varphi(x,y)} d\nu(x))$. The function $\psi(x,y) = \varphi(x,y) - \tilde{\varphi}(y)$ is Lipschitz, $\nu-$normalized and for any probability π on $X \times Y$, with $y-$marginal Q, we have

$$\int \psi \, d\eta + h^{\nu}(\eta) > \int \psi \, d\pi + H^{\nu}(\pi).$$

Let $\pi = \pi_{\psi}$ be defined by

$$\int f(x,y) \, d\pi_{\psi}(x,y) := \iint f(x,y) e^{\psi(x,y)} \, d\nu(x) dQ(y).$$

Then, as π_{ψ} has $y-$ marginal Q and $H^{\nu}(\pi_{\psi}) = -\int \psi \, d\pi_{\psi}$, we get that

$$\int \psi \, d\eta + h^{\nu}(\eta) > \int \psi \, d\pi_{\psi} + H^{\nu}(\pi_{\psi}) = 0.$$

This is a contradiction because, by definition of h^{ν}, $\int \psi \, d\eta + h^{\nu}(\eta) \leq 0$. $\square$

The next results will be necessary later; in the direction of getting a different point of view for the concept of information gain.

Proposition 29. *Let X and Y be compact metric spaces and suppose that there exists a continuous/Lipschitz function $\phi : X \times Y \to \mathbb{R}$, such that, $J = e^{\phi}$ is a ν-Jacobian of π. Let P and Q be the marginals of π on X and Y. Then, P is equivalent to ν (each one is absolutely continuous with respect to each other) and*

$$\frac{dP}{d\nu}(x) = \int e^{\phi(x,y)} dQ(y),$$

which is also continuous/Lipschitz. There exist constants $c_2 > c_1 > 0$, such that, $c_1 < \frac{dP}{d\nu} < c_2$, $\forall x \in X$. Finally, defining $\psi := \phi - \log(\frac{dP}{d\nu})$, we have that e^{ψ} is a $P-$Jacobian of π.

Proof. If a measurable and bounded function g depends only of the first coordinate, then, using Fubini's Theorem, we get

$$\int g(x)dP(x) = \int g(x)d\pi(x,y) = \iint e^{\phi(x,y)}g(x)\,d\nu(x)dQ(y)$$
$$= \int [\int e^{\phi(x,y)}dQ(y)]g(x)\,d\nu(x).$$

It follows that $\frac{dP}{d\nu}(x) = \int e^{\phi(x,y)}dQ(y)$ is a density. Clearly the function $\frac{dP}{d\nu}$ is continuous/Lipschitz and there are constants $c_2 > c_1 > 0$, such that, $c_1 < \frac{dP}{d\nu}(x) < c_2$, $\forall x \in X$. This shows that P and ν are equivalent and that $\log(\frac{dP}{d\nu})$ is continuous/Lipschitz.

Let $\psi(x,y) := \phi(x,y) - \log(\frac{dP}{d\nu})(x)$. Then, for any measurable and bounded function $g : X \times Y \to \mathbb{R}$ we have

$$\int e^{\psi(x,y)}g(x,y)\,dP(x) = \int e^{\phi(x,y)}g(x,y)d\nu(x).$$

By integrating both sides with respect to Q and using the fact that e^{ϕ} is a $\nu-$Jacobian of π, we get

$$\int [\int e^{\psi(x,y)}g(x,y))\,dP(x)]dQ(y) = \int g(x,y)d\pi(x,y).$$

This shows that ψ is a $P-$Jacobian of π. □

Proposition 30. *Suppose that X and Y are compact metric spaces. Let ν be a probability on X with $\mathrm{supp}(\nu) = X$ and $\phi : X \times Y \to \mathbb{R}$ be a continuous function. Let π_0 be a probability on $X \times Y$ with ν-Jacobian $J = e^{\phi}$. If π_0 is positive on open sets of $X \times Y$, then ϕ is the unique continuous function, such that, e^{ϕ} is a $\nu-$Jacobian of π_0.*

Proof. Let Q be the y-marginal of π_0 and ϕ_2 be a continuous function, such that, e^{ϕ_2} is a $\nu-$Jacobian of π_0. Then,

$$d\pi_0(x,y) = e^{\phi(x,y)}d\nu(x)dQ(y) \text{ and } d\pi_0(x,y) = e^{\phi_2(x,y)}d\nu(x)dQ(y).$$

This shows that the positive functions e^{ϕ_2} and e^{ϕ} satisfy $e^{\phi_2(x,y)} = e^{\phi(x,y)}$, π_0-a.e. $(x,y) \in X \times Y$. As π_0 is positive on open sets and ϕ, ϕ_2 are also continuous, we get $\phi_2(x,y) = \phi(x,y)$, for all $(x,y) \in X \times Y$. □

Definition 31 (Information Gain for compact spaces)**.** *Let X and Y be compact metric spaces and π_0 be a probability on $X \times Y$ which is positive on open sets and has $x-$marginal P_0. Suppose there exists a continuous function*

$\psi_0 : X \times Y \to \mathbb{R}$, *such that,* e^{ψ_0} *is a* P_0-*Jacobian of* π_0*. For any probability* π *on* $X \times Y$ *we define the information gain of* π *with respect to* π_0 *as*

$$IG(\pi, \pi_0) = -\int \psi_0 \, d\pi - H^{P_0}(\pi). \tag{21}$$

Suppose that π_0 has $y-$marginal Q_0 and $d\pi_0(x, y) = e^{\psi_0(x,y)} dP_0(x) dQ_0(y)$. Observe that in the expression $-\int \psi_0 \, d\pi - H^{P_0}(\pi)$ of (21) appears ψ_0 and P_0 but not Q_0. By considering the probability kernel $\hat{\nu}^y(dx) = e^{\psi_0(x,y)} dP_0(x)$ the next result shows that this definition is coherent with Definition 24.

Proposition 32. *Under the assumptions of Definition 31, we define* $\hat{\nu}^y(dx) = e^{\psi_0(x,y)} dP_0(x)$*. Then,*

$$IG(\pi, \pi_0) = IG(\pi, \hat{\nu}).$$

Proof. It is a consequence of Corollary 27. □

The next proposition shows that the above interpretation of information gain does not depend, in a sense to be explained, on the choice of the *a priori* probability ν.

Proposition 33. *Let* X *and* Y *be compact metric spaces and* π_0 *be a probability on* $X \times Y$ *which is positive on open sets. Let* ν *be an a priori probability on* X*. Suppose there exist a continuous function* ϕ*, such that,* e^{ϕ} *is a* ν*-Jacobian of* π_0*. Let* π *be any probability on* $X \times Y$*. Then,*

$$IG(\pi, \pi_0) = -\int \phi \, d\pi - H^{\nu}(\pi).$$

Proof. Let P and Q be the marginals of π_0. From Lemma 29, P is equivalent to ν, with $\frac{dP}{d\nu}(x) = \int e^{\phi(x,y)} dQ(y)$. Furthermore, the continuous function $\psi := \phi - \log(\frac{dP}{d\nu})$ is such that e^{ψ} is the $P-$Jacobian of π_0. Clearly, $e^{\psi-\phi} = \frac{d\nu}{dP}$, and so a Lipschitz function c is ν-normalized, if and only if, $c + \psi - \phi$ is P-normalized.

It follows that for any probability π we get

$$H^{\nu}(\pi) = H^P(\pi) + \int \psi - \phi \, d\pi.$$

Therefore,

$$IG(\pi, \pi_0) = -\int \psi \, d\pi - H^P(\pi) = -\int \phi \, d\pi - H^{\nu}(\pi).$$

□

From the above, it is legitimate to say that (21) generalizes (18). The next example shows that in a certain sense to be explained, (21) also generalizes (5) and (15).

Example 34. *Suppose* $X = \{1, 2, ..., d\}$, $Y = \{1, 2, ..., d\}^{\{2,3,4,5,...\}}$ *and identify* Ω *with* $X \times Y$ *by the homeomorphism*

$$\Omega \ni |x_1, x_2, x_3, x_4, ...) \to (x_1,\ |x_2, x_3, x_4, ...)) \in X \times Y. \tag{22}$$

When considering the a priori probability ν *as the counting measure on* X*, we get that a* $\nu-$*Jacobian of an invariant probability* π *is given by*

$$J^\pi(x_1, x_2, x_3, ...) = \lim_{n\to\infty} \frac{\pi(x_1, x_2, ..., x_n)}{\pi(x_2, x_3, ...x_n)},$$

for π *a.e.* $x \in \Omega$ *(it can be extended for any point of* Ω *by taking* $J^\pi = \frac{1}{d}$ *in a set of zero measure). Furthermore, the Kolmogorov-Sinai entropy of* π *coincides with* $H^\nu(\pi)$*. A measurable function* $\phi : \Omega \to \mathbb{R}$ *is normalized if* $\sum_a e^{\phi(|a,x_2,x_3,...))} = 1$*, for all* $|x_2, x_3, ...)$*. If* ϕ *is Lipschitz with corresponding equilibrium probability* π_ϕ*, then the unique continuous Jacobian of* π_ϕ *is* e^ϕ *and the pressure of* ϕ *is zero. In this case, it follows from* (5) *and Proposition 33 that for any invariant probability* π *we have*

$$h(\pi, \pi_\phi) = IG(\pi, \pi_\phi).$$

6 Specific information gain in the TFCA model

In this section, we introduce a dictionary connecting the definitions and results of previous sections with analogous ones for the TFCA model. This will allow us to introduce the specific information gain in this setting; a necessary step for introducing the concept of entropy production in the next section. First, we will remember some of the main definitions and results for the TFCA model described in [25] and which will be extensively used here.

Let (M, d) be a compact metric space and denote by $\Omega = \Omega^+$ the space $M^{\mathbb{N}}$. Elements in Ω will be written in the form $x = |x_1, x_2, x_3, ...)$, $x_i \in M$. The space Ω is compact using the metric $d(x, y) = \sum_{i=1}^{\infty} \frac{d(x_i,y_i)}{2^i}$. We also consider the Borel sigma-algebra in Ω.

The relation of the setting of this section with the previous ones can be clarified by considering $X = M$, $Y = M^{\{2,3,4,5,...\}}$ and identifying Ω with $X \times Y$, using the homeomorphism given in (22). Observe that Y can be also identified with $X \times Y$ using the homeomorphism $|x_2, x_3, x_4, ...) \to (x_2, |x_3, x_4, x_5, ...))$. From this identification the shift map $\sigma : \Omega \to \Omega$ given by $\sigma(|x_1, x_2, x_3, ...)) = |x_2, x_3, ...)$ can be also interpreted as the projection on Y. We say that a probability μ on $\Omega = M^{\mathbb{N}}$ is invariant for the shift map σ (or, shift-invariant), if for any continuous function $f : \Omega \to \Omega$, we have $\int f\, d\mu = \int f \circ \sigma\, d\mu$.

Assume we fixed an *a priori* probability ν on M satisfying $\mathrm{supp}(\nu) = M$. For each Lipschitz function $A : \Omega \to \mathbb{R}$ we consider the linear operator $\mathcal{L}_{A,\nu} : C(\Omega) \to C(\Omega)$ defined by

$$\mathcal{L}_{A,\nu}(f)(x) = \int e^{A(|a,x_1,x_2,x_3,...))} f(|a, x_1, x_2, x_3, ...))\, d\nu(a),\ \ x = |x_1, x_2, x_3, ...).$$

We call $\mathcal{L}_{A,\nu}$ the Ruelle operator (or, transfer operator) associated to the Lipschitz potential A and the *a priori* probability ν (we refer the reader to [25] for general properties of this operator).

For this operator there exists a unique (simple) positive eigenvalue λ_A associated to a positive eigenfunction $h = h_A$. If a continuous function $h > 0$ satisfies $\mathcal{L}_{A,\nu}(h) = \lambda_A \cdot h$, then h is Lipschitz. If h_1 and h_2 are eigenfunctions associated to λ_A, then $h_2 = c \cdot h_1$ for some constant c. There exists a unique probability measure ρ_A on Ω satisfying $\mathcal{L}^*_{A,\nu}(\rho_A) = \lambda_A \cdot \rho_A$, which means that

$$\int \mathcal{L}_{A,\nu}(f)\, d\rho_A = \lambda_A \int f\, d\rho_A,$$

for any continuous function $f : \Omega \to \mathbb{R}$. For a matter of convenience we fix the eigenfunction h_A which satisfies $\int h_A\, d\rho_A = 1$.

We point out that in the case the space of symbols M is not countable we really need to introduce an *a priori* probability in order to get a transfer operator.

A Lipschitz function A is called ν-normalized if $\mathcal{L}_{\bar{A},\nu}(1) = 1$, that is,

$$\int e^{A(|a,x_1,x_2,x_3,...))}\, d\nu(a) = 1,\ \ \forall\, x = |x_1, x_2, x_3, ...).$$

The function

$$\bar{A} = A + \log(h_A) - \log(h_A \circ \sigma) - \log(\lambda_A) \tag{23}$$

is ν-normalized. The associated eigenprobability $\rho_{\bar{A}}$ is shift-invariant and it will be denoted also by μ_A. It also satisfies $d\mu_A = h_A d\rho_A$ and $\mathcal{L}^*_{\bar{A},\nu}(\mu_A) = \mu_A$.

The **relative entropy** of an invariant probability μ on Ω with respect to the *a priori* probability ν on M is defined in [25] as

$$h^{\nu}(\mu) = - \sup_{B\, is\, \nu-normalized} \int B\, d\mu.$$

Considering the above identification of Ω and $X \times Y$ and applying Theorem 28 we see that this definition is consistent with the previous definition of H^ν. Let us formally enunciate this result (using also Theorem 13).

Theorem 35. *Denoting by π the probability on $X \times Y$, which corresponds to the shift-invariant probability μ on Ω, we get that the relative entropy $h^\nu(\mu)$, as defined in [25], coincides with $H^\nu(\pi)$. Furthermore, if Q is the $y-$marginal of π (which is identified with π because μ is shift invariant), then we have*

$$h^\nu(\mu) = H^\nu(\pi) = -D_{KL}(\pi \,|\, \nu \times Q).$$

In [1] it is proved that $h^\nu(\mu)$ coincides with the so called specific entropy of Statistical Mechanics (see [17]).

For any Lipschitz function $A : \Omega \to \mathbb{R}$ we have $h^\nu(\mu_A) = -\int \bar{A}\, d\mu_A$. Furthermore,

$$P_\nu(A) := \sup_{\mu \; shift-invariant} \int A\, d\mu + h^\nu(\mu) = \int A\, d\mu_A + h^\nu(\mu_A) = \log(\lambda_A).$$

The number $P_\nu(A)$ is called the $\nu-$pressure of A. A probability μ attaining the supremum value $P_\nu(A)$ is called an **equilibrium probability** for A. In [1] it is proved that μ_A is the unique equilibrium probability for A.

If M is a finite set with d elements and the *a priori* probability ν is set to be the counting measure (which is not a probability), then the relative entropy $h^\nu(\mu)$ above defined coincides with the Kolmogorov-Sinai entropy $h(\mu)$ (see Prop.7 in [25] and Lemma 7 in [26]). If we set the *a priori* probability ν as the normalized counting measure on M (which is a probability), then $h^\nu(\mu) = h(\mu) - \log d \leq 0$.

We remark that all the above concepts depend on the choice of the a priori probability on M. A universe of different choices is possible. If ν_1 and ν_2 are two different a priori probabilities which are not equivalent, then a probability μ on Ω could be a ν_1-equilibrium probability, and, at the same time, it not to be a ν_2-equilibrium probability (see Proposition 36).

We will call an invariant probability μ on Ω of **equilibrium probability** if there exists at least one *a priori* probability ν, satisfying $\text{supp}(\nu) = M$, and also a Lipschitz ν-normalized function A, such that, $\mathcal{L}^*_{(A,\nu)}(\mu) = \mu$. In this case the probability μ is the ν-equilibrium probability for the normalized potential A. If A is a Lipschitz function which is not normalized we can apply the construction given by (23) and we get an associated normalized potential $\bar{A}$ (which is also Lipschitz). The probability μ is the ν-equilibrium probability for both functions A and $\bar{A}$.

Let $A : \Omega \to \mathbb{R}$ be a $\nu-$normalized Lipschitz function. We will call e^A of a $\nu-$Jacobian of the shift-invariant probability μ if for any continuous function $g : \Omega \to \mathbb{R}$ we have

$$\iint e^{A(|a,x_2,x_3,...))} g(|a, x_2, x_3, ...)) d\nu(a) d\mu(|x_2, x_3, ...)) = \int g(x) d\mu(x).$$

The following statements are equivalent for a $\nu-$normalized Lipschitz function A and a shift-invariant probability μ on Ω:
i. e^A is a $\nu-$Jacobian of μ;
ii. μ is the ν-equilibrium probability of A;
iii. $\mathcal{L}^*_{(A,\nu)}(\mu) = \mu$.

Given an equilibrium probability μ we denote by P_μ the projection of μ on the first coordinate. This means that for any continuous function $g : \Omega \to \mathbb{R}$, which depends only of the first coordinate, we have

$$\int_M g(a) dP_\mu(a) := \int_\Omega g(x_1) d\mu(|x_1, x_2, x_3, ...)).$$

The next result is a Corollary of Proposition 29.

Proposition 36. *Suppose that μ is an equilibrium probability and let P_μ be the projection of μ on the first coordinate. If μ is the ν-equilibrium probability for the Lipschitz normalized potential A, then ν is equivalent to P_μ and*

$$\frac{dP_\mu}{d\nu}(a) = \int e^{A(|a,x_1,x_2,x_3...))} d\mu(|x_1, x_2, x_3, ...)),$$

which is also Lipschitz.

It is known that any equilibrium probability μ is positive on open sets of $\Omega = M^{\mathbb{N}}$ (see Prop. 3.1.8. in [33] - see also [7]). Therefore, the next result is a Corollary of Proposition 30.

Proposition 37. *Let μ be an equilibrium probability and let ν be an a priori probability. Suppose A is a Lipschitz $\nu-$normalized function, such that, μ is the $\nu-$equilibrium for A, then A is the unique Lipschitz $\nu-$Jacobian of μ.*

The next definition is inspired by (5) and (21).

Definition 38. *Let η be a shift-invariant probability and μ be an equilibrium probability. Then, we define the* ***specific information gain*** *of η, with respect to μ, by*

$$h(\eta, \mu) = \left[\int B\, d\mu + h^{P_\mu}(\mu)\right] - \left[\int B\, d\eta - h^{P_\mu}(\eta)\right], \tag{24}$$

where B is any Lipschitz function, such that, μ is the P_μ-equilibrium probability of B.

Observe that by the variational principle the information gain is ≥ 0.

There exists a unique P_μ-normalized function $\bar{B}$, such that, μ is the P_μ-equilibrium for $\bar{B}$. If B is not normalized, then there exists a positive function h_B and a positive number λ_B, such that,

$$\bar{B} = B + \log(h_B) - \log(h_B) \circ \sigma - \log(\lambda_B).$$

It follows that

$$\int B\, d\mu + h^{P_\mu}(\mu) - \int B\, d\eta - h^{P_\mu}(\eta) = \int \bar{B}\, d\mu + h^{P_\mu}(\mu) - \int \bar{B}\, d\eta - h^{P_\mu}(\eta).$$

This shows that $h(\eta, \mu)$ is well defined (it does not change if either B is P_μ-normalized, or not). We remark that if B is (the unique possible) normalized potential, then $\int B\, d\mu + h^{P_\mu}(\mu) = 0$ and therefore we get the following result which is a particular version of (21).

Proposition 39. *If μ is an equilibrium probability and e^B is the Lipschitz $P_\mu-$Jacobian of μ, then*

$$h(\eta, \mu) = - \int B\, d\eta - h^{P_\mu}(\eta).$$

The above definition considers, for an equilibrium probability μ, the a priori probability P_μ. In this way, the previous definition of specific information gain does not allow a choice of ν. The next result, which is a Corollary of Proposition 33, shows that if we exchange P_μ by another *a priori* probability ν, then, it is true a similar formula for $h(\eta, \mu)$. This means, that the information gain does not depend on the particular choice of ν, as long as μ is a ν-equilibrium probability.

Proposition 40. *Consider any a priori probability ν and any Lipschitz function A, such that, μ is the ν-equilibrium probability for A. Let η be any invariant probability. Then $h(\eta, \mu)$ as defined in (24) satisfies*

$$h(\eta, \mu) = \left[\int A\, d\mu + h^\nu(\mu) \right] - \left[\int A\, d\eta - h^\nu(\eta) \right]. \tag{25}$$

Proof. First note that if we replace A by its normalization $\bar{A}$, then the value on the right hand side of the above expression does not change. Then, we can suppose that A is $\nu-$normalized. In this case, it is just necessary to prove that

$$h(\eta, \mu) = - \left[\int A\, d\eta - h^\nu(\eta) \right]. \tag{26}$$

It follows from Theorem 35 and above proposition that $h(\eta, \mu)$ corresponds to $IG(\pi, \pi_0)$ in Definition 31. In this way, using this dictionary, equation (26) follows from Proposition 33. $\square$

Example 41. *Consider any a priori probability ν and the Lipschitz function $A = 0$. We observe that $\bar{\nu} = \nu \times \nu \times \nu \times ...$ is the ν-equilibrium probability for $A = 0$. Given any invariant probability η, then*

$$h(\eta, \bar{\nu}) = -h^{\nu}(\eta). \tag{27}$$

Therefore, the specific information gain generalizes the concept of relative entropy in [25].

Now we propose an interpretation of the information gain by using transfer operators defined from *a priori* probability kernels.

Remark 42. *Let μ be an equilibrium probability and suppose that e^B is the Lipschitz $\nu-$Jacobian of μ. Consider the identification of Ω and $X \times Y$ given by* (22) *and then define an a priori probability kernel on Ω by $\hat{\nu}^y(da) = e^{B(a,y)} d\nu(a),\ \ y = |\cdot, x_2, x_3, x_4, ...)$. For a fixed $\hat{\nu}-$normalized function A, let H_A be the operator acting on bounded and measurable functions $f : \Omega \to \mathbb{R}$ by*

$$H_A(f)(y) = \int e^{A(a,y)} f(a, y) \hat{\nu}^y(da),\ \ y = |x_1, x_2, x_3, ...).$$

Let η be a shift-invariant probability on Ω and suppose there exists a $\hat{\nu}$-normalized function A such that $H_A^*(\eta) = \eta$. This means that for any measurable function f, we have

$$\int e^{A(a,y)} f(a, y)\, \hat{\nu}^y(da) d\eta(y) = \int f(y)\, d\eta(y).$$

Then, $h(\eta, \mu) = IG(\eta, \mu) = IG(\eta, \hat{\nu}) = -H^{\hat{\nu}}(\eta) = \int A\, d\eta$. Furthermore,

$$IG(\eta, \mu) = \sup\{\int c\, d\eta\, |\, c\, is\, \hat{\nu} - normalized\}.$$

We finish this section by recalling some results presented in [1]. Let η, μ be two probabilities on Ω. For each $\Gamma \subset \mathbb{N}$ consider the canonical projection $\pi_\Gamma : \Omega \to M^\Gamma$ and, for each $n \in \mathbb{N}$, denote by Λ_n the set $\{1, ..., n\}$. Moreover, denote by $\mathcal{A}_n$, the σ-algebra on Ω generated by the projections $\{\pi_\Gamma,\ \Gamma \subset \Lambda_n\}$. Denote also

$$\mathcal{H}_{\Lambda_n}(\eta\,|\,\mu) = \begin{cases} \int_\Omega \frac{d\eta|_{\mathcal{A}_n}}{d\mu|_{\mathcal{A}_n}} \log\left(\frac{d\eta|_{\mathcal{A}_n}}{d\mu|_{\mathcal{A}_n}}\right) d\mu, & \text{if } \eta \ll \mu \text{ on } \mathcal{A}_n \\ +\infty, & \textit{otherwise} \end{cases}.$$

The next result is a consequence of Theorems 1 and 3 in [1]. From this result, we get an alternative and equivalent way of extending the concept of specific information gain for the TFCA model by considering (4) instead (5) and (21).

Proposition 43. *If μ is an equilibrium probability and η is shift-invariant on Ω, then*

$$\lim_{n\to\infty}\frac{1}{n}\mathcal{H}_{\Lambda_n}(\eta\,|\,\mu) = h(\eta,\mu).$$

7 The Involution kernel and the entropy production in the TFCA model

In the same way as in last section we assume that M is a compact metric space. We denote by Ω^- the space $M^{\mathbb{N}}$ with elements written in the form $y = (..., y_3, y_2, y_1|,\ y_i \in M$, and using the same metric as the one previously defined in $\Omega = \Omega^+$. Furthermore, $\sigma^- : \Omega^- \to \Omega^-$ is defined by $\sigma^-((..., y_4, y_3, y_2, y_1|) = (..., y_4, y_3, y_2|$. Points in $\hat{\Omega} = \Omega^- \times \Omega^+$ are written in the form

$$(y\,|\,x) = (..., y_3, y_2, y_1|x_1, x_2, x_3, ...)$$

and the bidirectional shift map $\hat{\sigma} : \hat{\Omega} \to \hat{\Omega}$ is defined by (7).

Observe that (Ω^-, σ^-) can be identified with (Ω, σ) from the conjugation $\theta : \Omega^- \to \Omega$, given by $\theta((..., z_3, z_2, z_1|) = |z_1, z_2, z_3, ...)$. Using this conjugation any result previously stated for (Ω, σ) has an analogous claim for (Ω^-, σ^-).

Consider a Lipschitz function $A : \Omega^- \times \Omega \to \mathbb{R}$, which does not depend of $y \in \Omega^-$. Then, it is naturally expressed as $A(x) = A\,(\,|x_1, x_2, x_3, ...)\,)$. One can show that there exists a (several, in fact) Lipschitz function $W : \Omega^- \times \Omega \to \mathbb{R}$, which is called an **involution kernel**, and a Lipschitz function A^-, such that

$$A^- := A \circ \hat{\sigma}^{-1} + W \circ \hat{\sigma}^{-1} - W, \tag{28}$$

where the function A^- **does not depend on** $x \in \Omega$ (see [2],[25]). The action of A^- is naturally expressed in coordinates $y = (..., y_3, y_2, y_1|$ as $y \to A^-(y)$ and the action of W can be expressed as $(y\,|\,x) \to W(y\,|\,x)$. All the above can be written in the form:

$$\begin{array}{rl} A^-((..., y_3, y_2, y_1|) = & A(|y_1, x_1, x_2, ...)\,) + W(..., y_3, y_2|y_1, x_1, x_2, ...) \\ & \qquad -W(..., y_3, y_2, y_1|x_1, x_2, ...) \end{array}, \tag{29}$$

for any $(..., y_3, y_2, y_1|x_1, x_2, x_3, ...) \in \hat{\Omega}$.

Definition 44. *We say that a function $A^- : \Omega^- \to \mathbb{R}$ satisfying* (28) *is a* **dual potential** *of A (via W). Moreover, we say that A is* **symmetric** *if $A = A^- \circ \theta^{-1}$, for some dual potential A^-, where θ was defined by* (8).

As we mentioned in the Introduction Section, for several important examples of potentials A it is possible to get the involution kernel W in an explict form (see [2], [3], [8]).

Following [2] and [25] we state two propositions.

Proposition 45. *Let $A : \Omega^+ \to \mathbb{R}$ be a Lipschitz function and $W : \hat{\Omega} \to \mathbb{R}$ be a Lipschitz involution kernel for A. Consider the function A^- which was defined by (28). Fix an a priori probability ν on M. Then, for any $x \in \Omega^+$, $y \in \Omega^-$ and any function $f : \hat{\Omega} \to \mathbb{R}$,*

$$\mathcal{L}_{A^-,\nu}\left(f(\cdot|x)\, e^{W(\cdot|x)}\right)(y) = \mathcal{L}_{A,\nu}\left(f \circ \hat{\sigma}(y|\cdot)\, e^{W(y|\cdot)}\right)(x). \tag{30}$$

Proposition 46. *Let $A : \Omega^+ \to \mathbb{R}$ be a Lipschitz function and $W : \hat{\Omega} \to \mathbb{R}$ be a Lipschitz involution kernel for A. Consider the function A^- as defined by (28). Fix an a priori probability ν on M. Let ρ_A and ρ_{A^-} be the eigenmeasures for $\mathcal{L}^*_{A,\nu}$ and $\mathcal{L}^*_{A^-,\nu}$, respectively. Suppose c is such that $\iint e^{W(y|x)-c}\, d\rho_{A^-}(y) d\rho_A(x) = 1$, and denote $K(y|x) := e^{W(y|x)-c}$. Then,*

1. The probability

$$d\,\hat{\mu}_A = K(y|x)\, d\rho_{A^-}(y)\ d\rho_A(x)$$

is invariant for $\hat{\sigma}$ and it is an extension of the ν-equilibrium probability μ_A.
2. The function $h_A(x) = \int K(y|x)\, d\rho_{A^-}(y)$ is the main eigenfunction for $\mathcal{L}_{A,\nu}$, and the function $h_{A^-}(y) = \int K(y|x)\, d\rho_A(x)$ is the main eigenfunction for $\mathcal{L}_{A^-,\nu}$.
3. $\lambda_A = \lambda_{A^-}$.

An involution kernel can express duality relations. From Proposition 46 we get that the function $(y|x) \to e^{W(y|x)-c}$ is an integral kernel that connects dual objects: the eigenfunction and the eigenprobability for the Ruelle operator. In [11], in the setting of Ergodic Optimization, the authors consider the involution kernel as a cost function in a Transport Theory problem and generic properties. It follows that the concepts of subactions and maximizing probabilities are related by Kantorovich duality.

Now we apply these above results in the understanding of the concept of entropy production. We start by refining item 1. of the last proposition.

Proposition 47. *The probability $d\,\hat{\mu}_A = K(y|x)\, d\rho_{A^-}(y)\ d\rho_A(x)$ is the unique $\hat{\sigma}$-invariant extension to $\hat{\Omega}$ of the equilibrium probability μ_A on Ω.*

Proof. Let $\hat{\mu}$ be any $\hat{\sigma}$-invariant probability on $\hat{\Omega}$ satisfying $\int g\,d\hat{\mu} = \int g\,d\mu_A$, when $g(y|x)$ does not depend of y. Consider any continuous function f on $\hat{\Omega}$. We claim that $\int f\,d\hat{\mu} = \int f\,d\hat{\mu}_A$. Indeed, as $\hat{\Omega}$ is compact, the function f is uniformly continuous. Fix any point $y_0 \in M$ and define the functions f_n on $\hat{\Omega}$, $n \in \mathbb{N}$, by $f_n(y|x) = f(y^n|x)$, where $y^n = (..., y_0, y_0, y_0, y_n, y_{n-1}, ..., y_2, y_1|$.

It follows that $\{f_n\}$ converges uniformly to f, and moreover, the function $f_n((..., y_3, y_2, y_1|x_1, x_2, x_3, ...))$ does not depend of y_k, for $k > n$.

From,

$$\int f_n\,d\hat{\mu} = \int f_n \circ \hat{\sigma}^{-n}\,d\hat{\mu} = \int f_n \circ \hat{\sigma}^{-n}\,d\mu_A = \int f_n \circ \hat{\sigma}^{-n}\,d\hat{\mu}_A = \int f_n\,d\hat{\mu}_A,$$

we conclude that $\int f\,d\hat{\mu} = \int f\,d\hat{\mu}_A$. □

Notation 48. *Let μ be an equilibrium probability on Ω^+. We denote by $\hat{\mu}$ the unique $\hat{\sigma}$-invariant extension to $\hat{\Omega}$ of μ and by μ^- the restriction of $\hat{\mu}$ to Ω^-.*

Proposition 49. *Let $A : \Omega^+ \to \mathbb{R}$ be a Lipschitz function and W be any Lipschitz involution kernel for A. Now, consider the function A^- on Ω^- as defined by (28). Fix an a priori probability ν on M. Let μ_A be the $\nu-$equilibrium of A and let $(\mu_A)^-$ defined as above. Then, $(\mu_A)^-$ is the $\nu-$ equilibrium of A^- in Ω^-, that is*

$$(\mu_A)^- = \mu_{(A^-)}.$$

Proof. From the above

$$d\,\hat{\mu_A} = K(y|x)\,d\rho_{A^-}(y)d\rho_A(x),$$

and $h_{A^-}(y) = \int K(y|x)\,d\rho_A(x)$ is the main eigenfunction for $\mathcal{L}_{A^-,\nu}$. Then, for any continuous function $f : \Omega^- \to \mathbb{R}$ we get

$$\int f(y)\,d(\mu_A)^- = \int f(y)\,d\hat{\mu_A} = \iint f(y)K(y|x)\,d\rho_A(x)d\rho_{A^-}(y)$$

$$= \int f(y)h_{A^-}(y)d\rho_{A^-}(y) = \int f(y)\,d\mu_{A^-}.$$

□

Definition 50. *The* **entropy production** *of the equilibrium probability μ is defined as*

$$e_p(\mu) = h(\mu, \theta_*\mu^-),$$

where $\theta_*\mu^-$ on Ω^+ is the push-forward of μ^- by the conjugation $\theta : \Omega^- \to \Omega^+$ given by (8).

Observe that as a consequence of the variational principle we get $e_p(\mu) \geq 0$, and it is zero, if and only if, $\mu^- = \mu$. As the specific information gain $h(\mu, \mu^-)$ "does not depend of ν", the above definition also "does not depend of ν". In fact, by definition, we should have to consider the *a priori* probability P_{μ^-}, but, if for some *a priori* probability ν the measure μ is a ν-equilibrium probability, then it follows that the probability μ^- also satisfies this property. Now, applying Proposition 40 we get an alternative formula for computing expression $e_p(\mu)$, but now using the *a priori* probability ν.

We will exhibit below other alternative ways for computing the entropy production.

Let $\hat{\theta} : \hat{\Omega} \to \hat{\Omega}$ be given by

$$\hat{\theta}(..., y_3, y_2, y_1 | x_1, x_2, x_3, ...) = (...x_3, x_2, x_1 | y_1, y_2, y_3, ...). \tag{31}$$

Observe that $\hat{\theta}^{-1} = \hat{\theta}$ and $\hat{\theta} \circ \hat{\sigma}^{-1} = \hat{\sigma} \circ \hat{\theta}$.

Proposition 51. *Let $A : \Omega \to \mathbb{R}$ be a Lipschitz function, $W : \hat{\Omega} \to \mathbb{R}$ be any Lipschitz involution kernel for A and let $A^- : \Omega^- \to \mathbb{R}$ be defined by (28). Let μ be any equilibrium probability on Ω and consider $\hat{\mu}$ and μ^- defined as above. Then,*

1. $\int A \, d\mu = \int A^- \, d\mu^-$
2. $\int A^- \circ \theta^{-1} \, d\mu = \int A \circ \theta \, d\mu^-$.

Proof. In order to prove item 1. we observe that

$$\int A \, d\mu = \int A \, d\hat{\mu} = \int A \circ \sigma^{-1} + W \circ \hat{\sigma}^{-1} - W \, d\hat{\mu} = \int A^- \, d\hat{\mu} = \int A^- d\mu^-.$$

Now we will prove item 2.

$$\int A^- \circ \theta^{-1} \, d\mu = \int A^- \circ \hat{\theta} \, d\hat{\mu} = \int A \circ \hat{\sigma}^{-1} \circ \hat{\theta} + W \circ \hat{\sigma}^{-1} \circ \hat{\theta} - W \circ \hat{\theta} \, d\hat{\mu}$$

$$= \int A \circ \hat{\theta} \circ \hat{\sigma} + W \circ \hat{\theta} \circ \hat{\sigma} - W \circ \hat{\theta} \, d\hat{\mu} = \int A \circ \hat{\theta} \, d\hat{\mu} = \int A \circ \theta \, d\mu^-.$$

□

Proposition 52. *Let μ be an equilibrium probability and ν be an a priori probability. Then, $h^\nu(\mu) = h^\nu(\mu^-)$.*

Proof. For each Lipschitz function $A : \Omega^+ \to \mathbb{R}$ we can consider a Lipschitz involution kernel W, and then, we get an associated Lipschitz function $A^- : \Omega^- \to \mathbb{R}$.

For the fixed *a priori* probability ν we have $\lambda_A = \lambda_{A^-}$. Then,

$$h^\nu(\mu) = - \sup_{A\, is\, \nu-normalized} \int A\, d\mu = - \sup_{A\, is\, Lipschitz\, on\, \Omega^+} \int A\, d\mu - \log(\lambda_A)$$

$$= - \sup_{A^-\, given\, from\, some\, Lip.\, A^+} \int A^-\, d\mu^- - \log(\lambda_{A^-})$$

$$\geq - \sup_{B^-\, is\, Lipschitz\, on\, \Omega^-} \int B^-\, d\mu^- - \log(\lambda_{B^-}) = h^\nu(\mu^-).$$

In order to get the opposite inequality, we follow a similar argument. We exchange the reasoning by $\hat{\theta}$: for each Lipschitz function $B^- : \Omega^- \to \mathbb{R}$, we take an involution kernel, and, an associated Lipschitz function $B^+ : \Omega^+ \to \mathbb{R}$. Now, we just have to proceed in the same way as before.

□

As a consequence we get the following claim for the entropy production:

Proposition 53. *Suppose that μ is an equilibrium probability and consider the associated probability μ^-. Suppose that for an a priori probability ν and for a Lipschitz function A^- we have that μ^- is the ν-equilibrium probability for A^-. Now, assume that μ^- and A^- are defined on Ω^+ via the conjugation θ. Then, the entropy production of μ satisfies*

$$e_p(\mu) = \int A^-\, d\mu^- - \int A^-\, d\mu.$$

We can take A^-, such that, $J^- = e^{A^-}$ is the $\nu-$Jacobian of μ^-.

Theorem 54. *Suppose that μ is the ν-equilibrium probability for the Lipschitz function $A : \Omega^+ \to \mathbb{R}$. Let W be any Lipschitz involution kernel for A and $A^- : \Omega^- \to \mathbb{R}$ be the function defined by (28). Suppose that A^- is defined on Ω^+ using the conjugation θ. Then,*

$$e_p(\mu) = \int A - A^-\, d\mu.$$

Proof. The claim follows from the previous result and Proposition 51. □

As a consequence of above theorem we get the result below.

Theorem 55. *Suppose that the potential $A : \Omega \to \mathbb{R}$ is symmetric, then, the equilibrium probability for A has entropy production zero.*

There are several examples of potentials A that are symmetric (see for instance [2], [3], [8]). Note that in order to check if the equilibrium probability μ for the Holder potential A has entropy production zero one have to follow a process of finding the eigenfunction and the eigenprobability; which is a procedure that in general we do not have explicit expressions. All this can be avoided when it is possible to show that for some involution kernel the potential is symmetric.

Proposition 56. *Suppose that μ is an equilibrium probability. Then,*

$$e_p(\mu) = h(\mu, \mu^-) = h(\mu^-, \mu) = e_p(\mu^-).$$

Proof. It follows from Proposition 46 and 47 that $(\mu^-)^- = \mu$. Consider an *a priori* probability ν, such that, μ is the $\nu-$equilibrium probability for a Lipschitz function A. Let A^- defined by (28) using any involution kernel. From Proposition 53 we get

$$e_p(\mu) = h(\mu, \mu^-) = \int A^- \, d\mu^- - \int A^- \circ \theta^{-1} \, d\mu$$

and

$$e_p(\mu^-) = h(\mu^-, \mu) = \int A \, d\mu - \int A \circ \theta \, d\mu^-.$$

Now, from Proposition 51 we get

$$\int A^- \, d\mu^- - \int A^- \circ \theta^{-1} \, d\mu = \int A \, d\mu - \int A \circ \theta \, d\mu^-.$$

This ends the proof. □

The next example considers the more simple case where $M = \{1, 2, ..., d\}$ is a finite set.

Example 57. *Take $M = \{1, 2, ..., d\}$ and consider as the a priori measure ν the counting measure on M.*

Any invariant probability μ for (Ω, σ) can be extended to a $\hat{\sigma}$-invariant probability $\hat{\mu}$ on $\hat{\Omega}$ by defining

$$\hat{\mu}([a_m, ..., a_1 | b_1, ..., b_n]) := \mu(\, |a_m, ..., a_1, b_1, ..., b_n]\,),$$

and using the extension Theorem. The restriction of $\hat{\mu}$ to Ω^- satisfies

$$\mu^-(\, [a_m, ..., a_2, a_1|\,) = \mu(\, |a_m, ..., a_2, a_1]\,).$$

Now, using the conjugation $\theta : \Omega^- \to \Omega$ in order to transfer μ^- to Ω^+, we get

$$\theta_* \mu^-(|a_1, a_2 .... a_m]) = \mu(|a_m, ..., a_2, a_1]). \tag{32}$$

As the Kolmogorov-Sinai entropy of μ is given by

$$h(\mu) = \lim_{n\to\infty} -\frac{1}{n} \sum_{i_1,\dots,i_n} \mu(|i_1, ..., i_n]) \log(\mu(|i_1, ..., i_n])),$$

we conclude that $h(\mu) = h(\mu^-)$.

Suppose now, that μ is the equilibrium probability for the Lipschitz normalized potential A. Then, $e^A = J$ is the Jacobian of μ, that is,

$$e^A(|x_1, x_2, x_3, ...)) = J(|x_1, x_2, x_3, ...)) = \lim_{n\to\infty} \frac{\mu(|x_1, x_2, x_3, ..., x_n])}{\mu(|x_2, x_3, ..., x_n])}.$$

Let J^- be the Jacobian of μ^- and define $A^- := \log(J^-)$. Then, using (32),

$$e^{A^-}(..., y_3, y_2, y_1|) = J^-(..., y_3, y_2, y_1|) = \lim_{n\to\infty} \frac{\mu(|y_n, ..., y_2, y_1])}{\mu(|y_n, ..., y_2])}.$$

The next example computes the entropy production for a Markov measure μ. Our estimate is coherent with expression (1) in [22].

Example 58. *Consider the line stochastic matrix $M = (p_{ij})$ and the initial probability vector $P = (\pi_i)$, such that, $PM = P$.*

We denote by μ the associated Markov measure, that is, for any cylinder $|x_1, x_2, ..., x_n]$ we set

$$\mu(|x_1, x_2, ..., x_n]) = \pi_{x_1} \cdot p_{x_1 x_2} \cdots p_{x_{n-1} x_n}.$$

Then,

$$J(|i, j, x_3, ...) = \frac{\pi_i p_{ij}}{\pi_j}.$$

We also get

$$J^-(..., y_3, j, i|) = \lim_{n\to\infty} \frac{\mu(|y_n, ..., y_3, j, i])}{\mu(|y_n, ..., y_3, j])} =$$

$$\lim_{n\to\infty} \frac{\pi_{y_n} \cdot p_{y_n y_{n-1}} \cdots p_{y_3 j} \cdot p_{ji}}{\pi_{y_n} \cdot p_{y_n y_{n-1}} \cdots p_{y_3 j}} = p_{ji}.$$

As J^- depends only on two coordinates, μ^- is also a Markov measure.

Considering the conjugation θ, we get,

$$\mu(|i,j]) = \pi_i p_{ij} \ \text{and} \ \mu^-(|i,j]) = \mu(|j,i]) = \pi_j p_{ji}.$$

Taking $A = \log(J)$ and $A^- = \log(J^-)$, we also get

$$e^{A(|i,j,x_3,\ldots)} = \frac{\pi_i p_{ij}}{\pi_j} \ \text{and} \ e^{A^-(|i,j,z_3,z_4,\ldots))} = p_{ji}.$$

Then, using the Theorem 54, we derive

$$e_p(\mu) = \int A - A^- \, d\mu = \sum_{i,j} \log\left(\frac{\pi_i p_{ij}}{\pi_j p_{ji}}\right) \pi_i p_{ij}. \tag{33}$$

We can compute $e_p(\mu)$, alternatively, using Proposition 53:

$$e_p(\mu) = \int A^- \, d\mu^- - \int A^- \, d\mu = \sum_{i,j} \log(p_{ji}) \mu^-(|i,j)) - \sum_{i,j} \log(p_{ji}) \mu(|i,j))$$

$$= \sum_{i,j} \log(p_{ji}) \pi_j p_{ji} - \sum_{i,j} \log(p_{ji}) \pi_i p_{ij} = \sum_{i,j} \log(p_{ij}) \pi_i p_{ij} - \sum_{i,j} \log(p_{ji}) \pi_i p_{ij}$$

$$= \sum_{i,j} \log(p_{ij}) \pi_i p_{ij} - \sum_{i,j} \log(p_{ji}) \pi_i p_{ij} + \left[\sum_i \pi_i \log(\pi_i) - \sum_j \pi_j \log(\pi_j)\right]$$

$$= \sum_{i,j} \log(p_{ij}) \pi_i p_{ij} - \sum_{i,j} \log(p_{ji}) \pi_i p_{ij} + \left[\sum_{i,j} \pi_i p_{ij} \log(\pi_i) - \sum_{i,j} \pi_i p_{ij} \log(\pi_j)\right]$$

$$= \sum_{i,j} \log(\frac{\pi_i p_{ij}}{\pi_j p_{ji}}) \pi_i p_{ij}.$$

In this case an involution kernel for A is the function $W : \{1,2\}^{\mathbb{N}} \to \mathbb{R}$ given by

$$W(\ldots, y_2, y_1 | x_1, x_2, ..) = \log p_{y_1\, x_1} - \log \pi_{x_1}$$

and the corresponding A^- is given by the $A^-(i,j,y_3,y_4..) = p_{j\,i}$.

The case with just two symbols is quite special as we will see now.

Example 59. *Entropy production zero - Suppose $\Omega = \{1,2\}^{\mathbb{N}}$ and assume that μ is a Markov measure (as defined above). Then, $e_p(\mu) = 0$. Indeed, as μ is invariant we get $\mu(|1,2)) = \mu(|2,1))$, and therefore, $\mu^-(|i,j)) = \mu(|j,i)) = \mu(|i,j)$, for any $i,j \in \{1,2\}$. It follows that $J^- = J^+$, and therefore, $\mu^- = \mu$. Consequently,*

$$e_p(\mu) = \int \log(J) - \log(J^-) \, d\mu = 0.$$

That is, in this case, the entropy production is zero.

Markov measures on $\Omega = \{1,2,3\}^{\mathbb{N}}$ may have non zero entropy production.

Example 60. *Consider the line stochastic matrix*

$$M = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/4 & 1/4 & 1/2 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$$

and the probability vector $P = (\pi_1, \pi_2, \pi_3) = (4/13, 4/13, 5/13)$*, which satisfies* $PM = P$*. Consider the Markov measure* μ *associated to* P *and* M*, that is,* μ *is given by*

$$\mu([j_1, j_2, j_3, ..., j_n]) = \pi_{j_1} p_{j_1 j_2} \cdots p_{j_{n-1} j_n}.$$

From expression (33) *we get*

$$e_p(\mu) = \sum_{j=1}^{3} \sum_{i=1}^{3} \log\left(\frac{\pi_i p_{ij}}{\pi_j p_{ji}}\right) \pi_i p_{ij},$$

and this value is approximately $0,01777 \neq 0$*.*

8 Appendix: Variational form of $H(\pi)$

In this section, we propose to study the entropy $H(\pi)$ which appears in (1) in a similar way as in [34].

If $(a_1, ..., a_n)$ and $(b_1, ..., b_n)$ are probability vectors such that $b_i > 0$, $\forall\, i$, then,

$$\sum_{i=1}^{n} a_i \log(a_i) \geq \sum_{i=1}^{n} a_i \log(b_i), \qquad (34)$$

with equality only if $a_i = b_i$, $\forall i$. This classical result can be found for example in [37] Lemma 3.3.

We will say that $f : X \times Y \to \mathbb{R}$ is a **normalized function**, if it satisfies

$$\sum_{x \in X} e^{f(x,y)} = 1,\ \forall y.$$

If the probability π on $X \times Y$ satisfies $\pi_{x,y} > 0$, $\forall (x,y)$, then $\log(J^\pi)$ is a normalized function.

Proposition 61. *Let π be a probability on $X \times Y = \{1, ..., d\} \times \{1, ..., r\}$ and f be a normalized function. Then,*

$$\sum_{x=1}^{d}\sum_{y=1}^{r} \pi_{x,y} \log(J^\pi(x,y)) \geq \sum_{x=1}^{d}\sum_{y=1}^{r} \pi_{x,y} f(x,y).$$

The equality occurs only if $J^\pi_{x,y} = e^{f(x,y)}$, $\forall (x,y)$, such that, $\pi_{x,y} > 0$.

Proof. Let $q_y = \sum_x \pi_{x,y}$. From (34), if $q_y > 0$, we have

$$\sum_x J^\pi(x,y) \log(J^\pi(x,y)) \geq \sum_x J^\pi(x,y), \log(e^{f(x,y)}),$$

with equality only if $J^\pi_{x,y} = e^{f(x,y)}$, $\forall x$. By definition $J^\pi(x,y) = \frac{\pi_{x,y}}{q_y}$, if $q_y > 0$, then we get

$$\sum_x \pi(x,y) \log(J^\pi(x,y)) \geq \sum_x \pi(x,y) f(x,y)).$$

If we assume that $J^\pi(x_0, y_0) \neq e^{f(x_0,y_0)}$, for some (x_0, y_0), such that $\pi(x_0, y_0) > 0$, then, we get

$$\sum_{x,y} \pi(x,y) \log(J^\pi(x,y)) > \sum_{x,y} \pi(x,y) f(x,y).$$

□

Proposition 62. *Let π be a probability on $X \times Y = \{1, ..., d\} \times \{1, ..., r\}$. Then,*

$$H(\pi) = -\sup\{\sum_{x,y} f(x,y)\pi_{x,y} \mid \sum_{x \in X} e^{f(x,y)} = 1, \forall y\}.$$

Proof. If $\pi_{x,y} > 0$, $\forall (x,y)$, then J^π is well defined, normalized and positive in $X \times Y$. From the last proposition, we get that the function $\log(J^\pi)$ attains the supremum. In this case, the proof is finished. If $\pi(x_0, y_0) = 0$, for some point (x_0, y_0), then the function $\log(J^\pi)$ is only well defined for π a.e. (x, y). In this case we get, from the last proposition,

$$H(\pi) \leq -\sup\{\sum_{x,y} f(x,y)\pi_{x,y} \mid \sum_{x \in X} e^{f(x,y)} = 1, \forall y\}.$$

In order to prove the opposite inequality we consider for each $\epsilon > 0$ the function f^ϵ defined in the following way: for fixed y_0, if $\pi(x, y_0) > 0$, for any

x, then $f^\epsilon(x,y_0)=\log(J^\pi(x,y_0))$, $\forall x$. For fixed y_0, if $\pi(x_0,y_0)=0$, for some x_0, we define

$$f^\epsilon(x,y_0)=\begin{cases}\log((1-\epsilon)J^\pi(x,y_0)) & \text{if}\,\pi(x,y_0)>0\\ a(\epsilon,y_0) & \text{if}\,\pi(x,y_0)=0\end{cases},$$

where $a(\epsilon,y_0)$ is chosen in such way that $\sum_x e^{f(x,y_0)}=1$.

With this construction we get that $f(x,y)$ is well defined for any $(x,y)\in X\times Y$ and $\sum_x e^{f(x,y)}=1,\ \forall y$. Furthermore,

$$\begin{aligned}\sum_{x,y} f^\epsilon(x,y)\pi_{x,y} &\geq \sum_{x,y}\log((1-\epsilon)J^\pi(x,y_0))\pi_{x,y}\\ &= \log(1-\epsilon)+\sum_{x,y}\log(J^\pi(x,y_0))\pi_{x,y}.\end{aligned}$$

Then,

$$\begin{aligned}H(\pi) &= -\sum_{x=1}^{d}\sum_{y=1}^{r}\log(J^\pi(x,y))\pi_{x,y}\\ &\geq -\sup\{\sum_{x,y} f(x,y)\pi_{x,y}\mid \sum_{x\in X} e^{f(x,y)}=1,\,\forall y\}-\log(1-\epsilon).\end{aligned}$$

Taking $\epsilon\to 0$, we finish the proof. □

A. O. Lopes partially is supported by CNPq grant.

References

[1] D. Aguiar, L. Cioletti and R. Ruviaro. A variational principle for the specific entropy for symbolic systems with uncountable alphabets. *Math. Nachr.* 291, no. 17 - 18, 2506 - 2525 (2018).

[2] A. Baraviera, A. Lopes and Ph. Thieullen. A large deviation principle for equilibrium states of Hölder potencials: the zero temperature case. *Stochastics and Dynamics*, **6**, 77-96 (2006).

[3] A. Baraviera and R. Leplaideur and A. O. Lopes, Ergodic Optimization, Zero Temperature Limits and the Max-Plus Algebra, mini-course in XXIX Coloquio Brasileiro de Matematica (2013)

[4] T. Benoist, V. Jaksic, Y. Pautrat and C-A. Pillet. On entropy production of repeated quantum measurements I. General theory. *Comm. Math. Phys.*, 357, no. 1, 77 - 123 (2018).

[5] M. Capinski and E. Kopp. Measure Integral and Probability. Springer-Verlag (2004).

[6] J-R. Chazottes, E. Floriani and R. Lima. Relative entropy and identification of Gibbs measures in dynamical systems. *J. Statist. Phys.*, 90, no. 3-4, 679 - 725 (1998).

[7] L. Cioletti, L. Melo, R. Ruviaro and E. Silva, On the dimension of the space of harmonic functions on transitive shift spaces. Advances in Math, 38, Article 1077585 (2021)

[8] L. Cioletti and A. O. Lopes, Correlation Inequalities and Monotonicity Properties of the Ruelle Operator, Stoch. and Dyn, 19, no. 6, 1950048, 31 pp (2019)

[9] L. Cioletti, M. Denker, A. O. Lopes and M. Stadlbauer, Spectral Properties of the Ruelle Operator for Product Type Potentials on Shift Spaces, Journal of the London Mathematical Society - Volume 95, Issue 2, 684-704 (2017)

[10] L. Cioletti and A. O. Lopes, Phase Transitions in One-dimensional Translation Invariant Systems: a Ruelle Operator Approach, Journ. of Statistical Physics, 159 - Issue 6, 1424-1455 (2015)

[11] G. Contreras, A. O. Lopes and E. Oliveira, Ergodic Transport Theory, periodic maximizing probabilities and the twist condition, "Modeling, Optimization, Dynamics and Bioeconomy I", Springer Proceedings in Mathematics and Statistics, Volume 73, Edit. David Zilberman and Alberto Pinto, 183-219 (2014)

[12] T. Cover and J. Thomas. Elements of information theory. 2 ed. Wiley-Interscience (2006)

[13] G. Crooks. Entropy production fluctuation Theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E*, 60, 2721 (1999).

[14] I. Ekeland and R. Témam. Convex Analysis and Variational Problems. North-Holland (1976).

[15] A. Fisher and A. O. Lopes, Exact bounds for the polynomial decay of correlation, 1/f noise and the central limit Theorem for a non-Holder Potential, Nonlinearity, Vol 14, Number 5, pp 1071–1104 (2001).

[16] G. Gallavotti and E. G. D. Cohen. Dynamical Ensembles in Nonequilibrium Statistical Mechanics, *Phys. Rev. Lett.* 74, 2694 (1995).

[17] H-O. Georgii. Gibbs measures and phase transitions. second edition. Ed. Gruyter, (2011).

[18] R. Gray. Entropy and information theory. New York : Springer-Verlag (1990).

[19] L. Y. Hataishi, Spectral Triples em Formalismo Termodinâmico e Kernel de Involução para potenciais Walters, Master Dissertation, Pos. Grad. Mat - UFRGS (2020)

[20] L. Y. Hataishi and A. O. Lopes, The involution Kernel for Potentials on the Walters' family, to appear

[21] F. Hofbauer, Examples for the nonuniquenes of the equilibrium state, Transactions AMS, 228, 133141, (1977)

[22] Da-quan Jiang, Min Qian and Min-ping Qian. Entropy Production and Information Gain in Axiom-A Systems. *Commun. Math. Phys.*, 214, 389 - 409 (2000).

[23] S. Kullback and R. A. Leibler. On Information and Sufficiency. *Ann. Math. Statist.* 22, no. 1, 79-86 (1951).

[24] A. O. Lopes and J. K. Mengue. Thermodynamic Formalism for Haar systems in Noncommutative Integration: probability kernels and entropy of transverse measures. *Erg. theo. Dyn. Sys.*, 41, 18351863 (2021)

[25] A. O. Lopes, J. K. Mengue, J. Mohr and R. R. Souza. Entropy and Variational Principle for one-dimensional Lattice Systems with a general a priori probability: positive and zero temperature. *Ergodic Theory and Dynamical Systems*, 35 (6), 1925-1961 (2015).

[26] A. Lopes, J. Mengue, J. Mohr and R. Souza. Entropy, Pressure and Duality for Gibbs plans in Ergodic transport. *Bull. Braz. Math. Soc.* Vol. 46, Issue 3, 353-389 (2015).

[27] A. O. Lopes, The Zeta Function, Non-Differentiability of Pressure and The Critical Exponent of Transition, Advances in Mathematics, Vol. 101, pp. 133-167, (1993).

[28] A. O. Lopes, E. Oliveira and Ph. Thieullen. The Dual Potential, the involution kernel and Transport in Ergodic Optimization, Dynamics, Games and Science -International Conference and Advanced School Planet Earth DGS II, Portugal (2013), Edit. J-P Bourguignon, R. Jelstch, A. Pinto and M. Viana, Springer Verlag, pp 357-398 (2015).

[29] A. O. Lopes, J. K. Mengue, J. Mohr and C. G. Moreira. Large Deviations for Quantum Spin probabilities at temperature zero. *Stochastics and Dynamics*, Vol. 18, No. 06, 1850044 (2018).

[30] A. O. Lopes and R. Ruggiero. Nonequilibrium in Thermodynamic Formalism: the Second Law, gases and Information Geometry, to appear in Qualitative Theory of Dynamical Systems.

[31] R. J. McEliece, The theory of information and coding, Addison-Wesley (1977)

[32] C. Maes. The fluctuation Theorem as a Gibbs property. *J. Statist. Phys.*, 95 , no. 1-2, 367 - 392 (1999).

[33] L. C. Melo. On the Maximal Eigenspace of the Ruelle Operator. PhD Thesis. UNB (2020) (available online from: https://repositorio.unb.br/handle/10482/39599).

[34] J. Mengue. Tópicos de álgebra linear e probabilidade. SBM (2016).

[35] J. Mengue and E. Oliveira. Duality results for iterated function systems with a general family of branches. *Stochastics and Dynamics*, Vol. 17, No. 03, 1750021 (2017).

[36] J. Mohr, Product type potential on the XY model: selection of maximizing probability and a large deviation principle, to appear in Qual. Theo. of Dyn. Syst.

[37] W. Parry and M. Pollicott. Zeta functions and the periodic orbit structure of hyperbolic dynamics. Astérisque, Vol 187-188, pp 1-268 (1990).

[38] J. R. Quinlan. Induction of decision trees. *Machine learning*, vol 1, issue 1., 81 - 106, (1986).

[39] D. Ruelle. A generalized detailed balance relation. *J. Stat. Phys.* 164, no. 3, 463-471 (2016).

[40] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal* Vol. 27 Issue 3, 379 - 423 (1948).

[41] M. Viana and K. Oliveira. Foundations of Ergodic Theory. Cambridge Press (2016).

[42] P. Walters. A natural space of functions for the Ruelle operator Theorem. *Ergodic Theory and Dynamical Systems*, 27, 1323–1348, (2007).

[43] P. Walters. An introduction to Ergodic Theory. Springer Verlag (1982).