BAYES POSTERIOR CONVERGENCE FOR LOSS FUNCTIONS VIA ALMOST ADDITIVE THERMODYNAMIC FORMALISM

ARTUR O. LOPES, SILVIA R. C. LOPES, AND PAULO VARANDAS †

Abstract. Statistical inference can be seen as information processing involving input information and output information that updates belief about some unknown parameters. We consider the Bayesian framework for making inferences about dynamical systems from ergodic observations, where the Bayesian procedure is based on the Gibbs posterior inference, a decision process generalization of standard Bayesian inference (see [7, 37]) where the likelihood is replaced by the exponential of a loss function. In the case of direct observation and almost-additive loss functions, we prove an exponential convergence of the a posteriori measures to a limit measure. Our estimates on the Bayes posterior convergence for direct observation are related and extend those in [47] to a context where loss functions are almost-additive. Our approach makes use of non-additive thermodynamic formalism and large deviation properties [40, 39, 57] instead of joinings.

1. Introduction and statement of the main results

1.1. **Bayesian inference.** Statistical inference aims to update beliefs about uncertain parameters as more information becomes available. The Bayesian inference, one of the most successful methods used in decision theory, builds over Bayes' theorem:

$$\text{Prob}(H \mid E) = \frac{\text{Prob}(E \mid H) \cdot \text{Prob}(H)}{\text{Prob}(E)} = \frac{\text{Prob}(E \mid H)}{\text{Prob}(E)} \cdot \text{Prob}(H) \tag{1}$$

which expresses the conditional probability of the hypothesis H conditional to the event E with the probability that the event or evidence E occurs given the hypothesis H. In the previous expression, the *posterior probability* $\text{Prob}(H \mid E)$ is inferred as an outcome of the *prior probability* $\text{Prob}(H)$ on the hypothesis, the model evidence $\text{Prob}(E)$ and the likelihood $\text{Prob}(E \mid H)$. Bayes' theorem has been widely used as an inductive learning model to transform prior and sample information into posterior information and, consequently, in decision theory. One should not make confusion between $\text{Prob}(E \mid H)$ and $\text{Prob}(H \mid E)$. Let us provide a simple example. Suppose one is tested for covid-19, and the test turns out to be positive. If the test is 99% accurate, the latter means that Prob(Positive test | Covid-19) = 0.99. However, the most relevant information is Prob(Covid-19 | Positive test), namely the probability of having covid-19 once one is tested positive, which is related with the former conditional probability by (1). If proportion Prob(Covid-19) of infected persons in the total population is 0.001 it it possible to compute the normalizing term Prob(Positive test) and to conclude that Prob(Covid-19 | Positive test) = 0.5, which provides a different and rather relevant information (see e.g. [13]

Date: August, 2021

† Email: paulo.varandas@ufba.br (corresponding author).

2010 *Mathematics Subject Classification.* 62F15, 37D35, 60F10, 62C12, 62E10.

Key words and phrases. Bayesian inference, thermodynamic formalism, Gibbs posterior convergence, large deviations.

for all computations in a similar example). The conclusion is that both the prior and the data contain important information, and so neither should be neglected.

The process of drawing conclusions from available information is called inference. However, in many physical phenomena the available information is often insufficient to reach certainty through a logical reasoning. In these cases, one may use different approaches for doing inductive inference, and the most common methods are those involving probability theory and entropic inference (cf. [13]). The frequentist interpretation advocates that the probability of a random event is given by the relative number of occurrences of the event in a sufficiently large number of identical and independent trials. An alternative approach is given by the Bayesian interpretation which became more popular in the recent decades and sustains that a probability reflects the degree of belief of an agent in the truth of a proposition. Citing [13], "the crucial aspect of Bayesian probability measures is that different agents may have different degrees of belief in the truth of the very same proposition, a fact that is described by referring to Bayesian probability measures as being subjective".

In the framework of parametric Bayesian statistics, one is interested in updating beliefs, or the degree of confidence, on the space of parameters Θ, which play the role of the variable H in the expression (1) above. In rough terms, the formula (1) expresses that the belief on a certain set of parameters is updated from the original belief, after an event E, by how likely such event is for all parameterized models. This supports the idea that while frequentists say the data are random and the parameters are fixed, Bayesians say the data are fixed and the parameters are random. The basic idea in classical Bayesian inference is the updating of a prior belief distribution to a posterior belief distribution when the parameter of interest is connected to observations via the likelihood function. In [7], Bissiri et al propose a general framework for the Bayesian inference arguing that a valid update of a prior belief distribution to the posterior one can be made for parameters which are connected to observations through a loss function which accumulates information as time passes rather than the likelihood function. In their framework, the classical inference process corresponds to the special case where the loss function is expressed as the negative log likelihood function. In this more general framework, the choice of loss function determines the way that the data are analyzed contribute to the mechanism of updating the belief distribution on the space of parameters, and such choice is often subjective and depends on the kind of feature one desires to highlight from the data. Moreover, the purpose is that the successive updated belief distributions, called posterior distributions, either converge or concentrate around the unknown targeted parameters. We refer the reader to [1, 21, 22, 49, 51, 54] for more information on classical Bayesian inference formalism.

The Bayesian inference in the context of observations arising from dynamical systems faces some natural challenges. The first one is that the process of taking time series (via Birkhoff theorem) lacks independence: if $T : (Y, \nu) \to (Y, \nu)$ is a measure preserving map and $\phi : Y \to \mathbb{R}$ is an observable then the sequence of random variables $(\phi \circ T^n)_{n\geqslant 1}$ is identically distributed but the random variables are not even pairwise independent. The second one concerns the choice of the loss function to make update of beliefs on the space of parameters. From the Physics and the Dynamical Systems viewpoints it is natural that loss functions should value some of the geometric or chaotic properties of the dynamical system, identified either in terms of Lyapunov exponents, joint spectral radius of matrix cocycles, entropy or estimates on the Charathéodory, box-counting or Hausdorff dimension of repellers and attractors, and with applications in wavelets and multifractal analysis, just to mention a few. These concepts, central

in mathematical physics (see e.g. [4, 5, 6, 10, 8, 20, 29, 27, 32, 33] and references therein) appear naturally as limits of either Birkhoff averages of potentials, sub-additive or almost-additive potentials (or several other versions of non-additivity, to be defined in Subsection 3.2). As a first example, if T is a C^1-smooth volume preserving and ergodic diffeomorphism on a surface then its largest Lyapunov exponent is by the random product of $SL(2,\mathbb{R})$-matrices as

$$\lambda_+(T,\mathrm{Leb}) = \lim_{n\to\infty} \frac{1}{n} \log \|A(T^{n-1}(y)) \dots A(T(y))A(y)\|$$

for Lebesgue almost every $y \in Y$, where $A = DT : Y \to TY$ is the derivative cocycle. In general, the sequence of observables $\Phi = (\varphi_n)_{n\geqslant 1}$ defined by $\varphi_n(y) = \log \|A(T^{n-1}(y)) \dots A(T(y))A(y)\|$ is sub-additive and, in the special case that the linear cocycle has an invariant cone-field, this sequence is actually almost-additive (cf. [28]). A second example concerns the Shannon-McMillan-Breiman formula for entropy on one-sided subshifts of finite type $\sigma : \Omega \to \Omega$, where the set $\Omega \subset \{1, 2, ..., q\}^{\mathbb{N}}$ is σ-invariant determined by a transition matrix $M_\Omega \in \mathcal{M}_{q\times q}(\{0,1\})$ and

$$h_\mu(\sigma) = \lim_{n\to\infty} -\frac{1}{n} \log \mu(C_n(x)), \qquad \text{for } \mu\text{-a.e. } x \tag{2}$$

where $C_n(x) \subset \Omega$ denotes the n-cylinder set containing the sequence $x = (x_1, x_2, x_3, \dots)$. The sequence of observables $\Phi = (\varphi_n)_{n\geqslant 1}$ defined by $\varphi_n(y) = -\log \mu(C_n(x))$, which is non-additive in general, is additive and almost-additive in the relevant classes of Bernoulli and Gibbs measures, respectively (see Lemma 3.3). Finally, it is worth to mention that sub-additive and almost-additive sequences appear naturally also in applications to several other areas of knowledge and appear for instance in the study of factorial languages by Thue, Morse and Hedlund in the beginning of the twentieth century (see [52] and references therein).

In this article, inspired by the relevant physical quantities arising from non-additive sequences of potentials, we will establish a bridge between non-additive thermodynamic formalism of dynamical systems and Gibbs posterior inference in statistics (to be defined in Subsection 1.2 below), two areas of research in connection with statistical physics. We refer the interested reader to the introduction of [47] for a careful and wonderful exposition on the link between Bayesian inference and thermodynamic formalism, and a list of cornerstone contributions. We will mostly be interested in the parametric formulation of Bayesian inference, as described below. Let $\sigma : \Omega \to \Omega$ be a subshift of finite type. This will serve as the underlying dynamical system, with respect to which ergodic samples are obtained along finite orbits $\{y, \sigma(y), \dots, \sigma^{n-1}(y)\}$, $y \in \Omega$, taken according to a certain fixed reference ergodic probability. We take a family of Gibbs probability measures $\{\mu_\theta\}_{\theta\in\Theta}$ as the models in the inference procedure for their relevance and ubiquity in the thermodynamic formalism of dynamical systems, and are of crucial importance in several other fields as in the study of the randomness of time-series, decision theory, quantum information and information gain, just to mention a few (cf. [2, 13, 30, 36, 46]). In our context, Gibbs measures appear as fixed points of the dual of certain transfer operators. Let us be more precise. For any Lipschitz continuous potential $A : \Omega \to \mathbb{R}$, the Ruelle-Perron-Frobenius transfer operator associated to A is defined by

$$\mathcal{L}_A(\varphi)(x) = \sum_{\sigma(y)=x} e^{A(y)}\varphi(y).$$

The potential A is called normalized if $\mathcal{L}_A(1) = 1$, and in this case, it is natural to write $A = \log J$, and we call J the Lipschitz continuous Jacobian. A Gibbs measure μ is any σ-invariant probability measure obtained as a fixed point of the dual operator $\mathcal{L}^*_{\log J}$ acting on

the space of probability measures on Ω, for some Lipschitz continuous and normalized Jacobian J. In this way, it is natural to parametrize Gibbs probabilities by the space of normalized Lipschitz continuous Jacobians J, hence this space can be observed as an infinite dimensional Riemannian analytic manifold [35, 45, 46]. Invariant Gibbs measures are equilibrium states, namely they satisfy a variational relations (cf. Subsection 1.3 for more details). Given a prior probability measure Π_0 on the space Θ of parameters and an ergodicity sample taken according to a Gibbs measure μ_{θ_0}, the posterior probability (i.e. updated belief distribution) is determined using the loss functions $\ell_n : \Theta \times \Omega \times \Omega \to \mathbb{R}$, where $\ell_n(\theta, x, y)$ encodes the information on the parameter θ accumulated along the ergodic sample $\{y, \sigma(y), \ldots, \sigma^{n-1}(y)\}$ and influenced by the measurements along the orbit $\{x, \sigma(x), \ldots, \sigma^{n-1}(x)\}$. The Shannon-McMillan-Breiman formula (2) suggests the use of loss functions to collect the information of the measure on cylinder sets in Ω (cf. expressions (4), (5) and (9) below). The relative entropy, also called Kullback-Leibler divergence and defined by (27), makes the comparison between the measurements of cylinders according to two different Gibbs measures. This notion is of paramount importance in Physics and will be used to interconnect log likelihood inference with the direct observation analysis of Gibbs probability measures. Our main results guarantee that posterior consistency for certain classes of loss functions determined by almost-additive sequences of potentials: the posterior distributions asymptotically concentrate around the unknown targeted parameter θ_0, often with exponential speed (we refer the reader to Theorems A, B and C for the precise statements). The main ingredient to obtain quantitative estimates on the convergence for the parameter θ_0 is the use of large deviations for non-additive sequences of potentials [57].

Our results are strongly inspired, and should be compared, with those by McGoff, Mukherjee and Nobel [47], where the authors established posterior consistency of (hidden) Gibbs processes on mixing subshifts of finite type using properties of Gibbs measures. For that purpose, they consider a more general framework, where the dynamical system $T : Y \to Y$ on a Polish space does not necessarily coincide with the subshift of finite type $\sigma : \Omega \to \Omega$. In particular, the ergodic time series is determined by a T-invariant and ergodic probability measure ν, that could be unrelated to the Gibbs measures $\{\mu_\theta\}_{\theta\in\Theta}$ for the shift. If the loss functions are additive (i.e. $\ell_n = \sum_{j=0}^{n-1} \ell(\theta, \sigma^j(x), T^j(y))$ for some function $\ell : \Theta \times \Omega \times Y \to \mathbb{R}$ satisfying a mild regularity condition then the main results in [47] ensure that it is possible to formulate the problem as a limiting variational problem and to identity the parameters, obtained as minimizing parameters for a lower semicontinuous function $V : \Theta \to \mathbb{R}$, for which the posterior consistency holds: if $\Theta_{\min} = \operatorname{argmin}_{\theta\in\Theta} V(\theta)$ then the posterior distributions $\Pi_n(\cdot \mid y)$, defined by (6), satisfy

$$\lim_{n\to\infty} \Pi_n(\Theta \setminus U \,|y) = 0$$

for each open neighborhood U of $\Theta_{\min}$ and for ν-almost every $y \in Y$ (cf. [47, Theorem 2]). The proof of this result requires the use of joinings (or couplings) of the model system and the observed system, and results on fibered entropy. Our framework corresponds to the special case of direct observation, that the dynamical system T coincides with the subshift of finite type σ and the target parameter is a single $\theta_0 \in \Theta$, with a subtler difference that our assumptions ensure that $\mu_\theta \neq \mu_{\tilde\theta}$ for every distinct $\theta, \tilde\theta \in \Theta$. Our results complement the ones in [47] in the sense that the information can be collected by more general loss functions ℓ_n. Furthermore, the more direct use of large deviation techniques allows to prove an exponential speed of convergence in the posterior consistency (cf. Theorem A), which were not known even in the context of direct observation (cf. [47, Theorem 2 and Remark 8]). Summarizing, the three main novelties are the

extension to non-additive loss functions, the exponential rate of convergence and the proof which is not based on joinings and fiber entropy. It is also worth noticing that, more recently, Su and Mukherjee [55] also used a large deviations approach for posterior consistency, using Varadhan's large deviation principle for stochastic processes. A different point of view of the Bayesian *a priori* and *a posteriori* formalism will appear in [26] where results on thermodynamic formalism for plans are used (see [42, 43]). In [36] the author considered log-likelihood estimators in classical thermodynamic formalism and the inference concerns Hölder potentials and not probabilities.

To finalize, one should mention that there is an increasing interest to explore the strong connection between Statistical Inference and Physics in general. There are several such connections in this regard, including a Bayesian approach to the dynamics of the classical ideal gas [58, Section 31.3], prior sensitivity in the Bayesian model selection context to some galaxy data sets [11]. In the monograph [13], the author clarifies the conceptual foundations of Physics by deriving the fundamental laws of statistical mechanics and of quantum mechanics as examples of inductive inference, while he also advocates that, in view of the fact that models may need to change as time evolves, it may be the case that all areas of Physics may be modeled using inductive inference.

1.2. **Gibbs posterior inference.** According to the Gibbs posterior paradigm [7, 37], the beliefs should be updated according to the Gibbs posterior distribution. Let us recall the formulation of this posterior measure following [47].

Observed system. Assume that Y is a complete and separable metric space and that $T : Y \to Y$ is a Borel measurable map endowed with a T-invariant, ergodic probability measure ν. This dynamical system represents the observed system and will be used to update information for the model. This is the analogue of the data in the context of Statistics. The updated belief, given by the *a posteriori* measure, is obtained by feeding data obtained from the observed system on a model by means of a loss function.

Model families. Consider a transitive subshift of finite type $\sigma\colon \Omega \to \Omega$ where σ denotes the right-shift map, acting on a compact invariant set $\Omega \subset \{1, 2, ..., q\}^{\mathbb{N}}$ determined by a transition matrix $M_\Omega \in \mathcal{M}_{q\times q}(\{0, 1\})$. The map σ presents different statistical behaviors (e.g. measured in terms of different convergences for Cesàro averages of continuous observables) according to any of its equilibrium states associated to Lipschitz continuous observables, each of which satisfies a Gibbs property (see e.g. Remark 1 in [48, Section 2] or [41]).

Consider a compact metric space Θ and a family of σ-invariant probability measures

$$\mathscr{G} = \big\{\mu_\theta\colon \theta \in \Theta\big\}$$

so that: (i) for every $\theta \in \Theta$ the probability measure μ_θ is a Gibbs measure associated to a Lipschitz continuous potential $f_\theta : \Omega \to \mathbb{R}$, that is, there exists $K_\theta > 1$ and $P_\theta \in \mathbb{R}$ so that

$$\frac{1}{K_\theta} \leqslant \frac{\mu_\theta(C_n(x))}{e^{-nP_\theta} + S_n f_\theta(x)} \leqslant K_\theta, \qquad \forall n \geqslant 1, \tag{3}$$

where $S_n f_\theta = \sum_{j=0}^{n-1} f_\theta \circ \sigma^j$ and $C_n(x) \subset \Omega$ denotes the n-cylinder set in the shift space Ω containing the sequence $x = (x_1, x_2, x_3, \dots)$; and (ii) the family $\Theta \ni \theta \mapsto f_\theta$ is continuous (in the Lipschitz norm). We assume Gibbs measures to be normalized, hence probability measures. It is well known that the previous conditions ensure the continuity of the pressure function $\Theta \ni \theta \mapsto P_\theta$ and of the map $\Theta \ni \theta \mapsto \mu_\theta$ (in the weak* topology) [48]. In particular, one can

take a uniform constant $K > 0$ in (3). The problem to be considered here involves a formulation and analysis of an iterative procedure (based on an ergodic sample and updated information) on the family $\mathscr{G}$ of models.

Loss functions and Gibbs posterior distributions. Consider the product space $\Theta \times \Omega$ endowed with the metric d defined as $d(\,(\theta, x), (\theta', x')\,) = \max\{d_\Theta(\theta, \theta'), d_\Omega(x, x')\,\}$. A fully supported probability measure Π_0 on Θ describes the *a priori* uncertainty on the Gibbs measure.

Given such an *a priori* probability measure Π_0 on the space of parameters Θ and a sample of size n (determined by the observed system T) we will get the *a posteriori* probability measure Π_n on the space of parameters Θ, taking into account the updated information from the data. More precisely, given Π_0 and a family $(\mu_\theta)_{\theta\in\Theta}$, consider the probability measure P_0 on the product space $\Theta \times \Omega$ given by

$$P_0(E) = \int \int \mathbf{1}_E(\theta, x)\, d\mu_\theta(x)\, d\Pi_0(\theta)$$

for all Borel sets $E \subset \Theta \times \Omega$. In other words, P_0 has the *a priori* measure Π_0 as marginal on Θ and admits a disintegration on the partition by vertical fibers where the fibered measures are exactly the Gibbs measures $(\mu_\theta)_{\theta\in\Theta}$. There is no action of the dynamics T on this product space. Indeed, the *a posteriori* measures are defined using loss functions. For each $n \in \mathbb{N}$ consider a continuous *loss function* ℓ_n of the form

$$\ell_n : \Theta \times \Omega \times Y \to \mathbb{R},$$

consider the probability measure P_n on $\Theta \times \Omega$ given by

$$P_n(E \mid y) = \int \int \mathbf{1}_E(\theta, x) e^{-\,\ell_n(\theta,x,y)}\, d\mu_\theta(x)\, d\Pi_0(\theta) \tag{4}$$

for all Borel sets $E \subset \Theta \times \Omega$, and set

$$Z_n(y) = \int_\Theta \int_\Omega e^{-\,\ell_n(\theta,x,y)}\, d\mu_\theta(x)\, d\Pi_0(\theta), \tag{5}$$

where $x = (x_1, x_2, ..., x_n, \dots) \in \Omega$ and $y \in Y$. In the special case that $Y = \Omega$, that $-\ell_n : \Theta \times \Omega \times \Omega \to \mathbb{R}$ coincides with a n-Birkhoff sum of a fixed observable ψ with respect to T and Π_0 is a Dirac measure, the expression (5) resembles the partition function in statistical mechanics whose exponential asymptotic growth coincides with the topological pressure of T with respect to ψ.

Given $y \in Y$ and $n \geqslant 1$, the *a posteriori* Borel probability measure $\Pi_n(\cdot \,|\, y)$ on the parameter space Θ (at time n and determined by the sample of y) is defined by

$$\Pi_n(B \,|\, y) = \frac{1}{Z_n(y)} \int_B \int_\Omega e^{-\,\ell_n(\theta,x,y)} d\mu_\theta(x) d\Pi_0(\theta)\,, \tag{6}$$

for every measurable $B \subset \Theta$ and appears as marginals of the probability measures $P_n(\cdot \mid y)$ given above.

The general question is to describe the set of probability measures $\Pi_n(.\,|\,y)$ on the parameter space Θ, namely if their marginals converge and to formulate the locus of convergence in terms of some variational principle or as points of maximization for a certain function (see e.g. [47, Theorem 2] for a context where the loss functions are chosen such that the support of such measures on the minimization locus of a certain rate function).

The main problem we are interested in is to understand whenever an ergodic sampling process, taken according to a fixed probability measure, can help to identify it from a recursive process involving Bayesian inference. Assume that $Y = \Omega$, that $T = \sigma$ is the shift and that one is interested in a specific probability measure $\mu_{\theta_0} \in \mathscr{G}$, where $\theta_0 \in \Theta$. If $\nu = \mu_{\theta_0}$ then the ergodic time series $\{y, T(y), T^2(y), \dots T^{n-1}(y)\}$ is distributed according to this probability measure. From the Birkhoff time series is it possible to successively update the initial a priori probability measure Π_0 in order to get a sequence of probability measures $\Pi_n(\cdot \mid y)$ on Θ (the a posteriori probability measure at time n) as described. We ask the following:

- Does the limit $\lim_{n\to\infty} \Pi_n$ exist?
- If the previous question has an affirmative answer:
  - is it the Dirac measure δ_{θ_0} on $\theta_0 \in \Theta$?
  - is it possible to estimate the speed of convergence to the limiting measure?

In this paper we answer the previous questions for loss functions that are not necessarily arising from Birkhoff averaging but that keep some almost additive property. For that reason our approach will make use of results from non-additive thermodynamic formalism, hence it differs from the one considered in [47]. We refer the reader to [16] for a related work which does not involve Bayesian statistics.

This paper is organized as follows. In the rest of this first section we formulate the precise setting we are interested in and state the main results. In Section 2 we present several examples and applications of our results. Section 3 is devoted to some preliminaries on relative entropy, large deviations and non-additive thermodynamic formalism. Finally, the proofs of the main results are given in Section 4.

1.3. **Setting and Main results.** Let $\sigma : \Omega \to \Omega$ be a subshift of finite type endowed with the metric $d_\Omega(x, y) = 2^{-n(x,y)}$, where $n(x, y) = \inf\{n \geqslant 1 \colon x_n \neq y_n\}$, and denote by $\mathcal{M}_\sigma(\Omega)$ the space of σ-invariant probability measures. The space $\mathcal{M}_\sigma(\Omega)$ is metrizable and we consider the usual topology on it (compatible with weak* convergence). Let D_Ω be a metric on $\mathcal{M}_\sigma(\Omega)$ compatible with the weak* topology. The set $\mathcal{G} \subset \mathcal{M}_\sigma(\Omega)$ of Gibbs measures for Lipschitz continuous potentials is dense in $\mathcal{M}_\sigma(\Omega)$ (see for instance [39]). Given a Lipschitz continuous potential $A : \Omega \to \mathbb{R}$ we denote by μ_A the associated Gibbs measure. We say that the Lipschitz continuous potential $A : \Omega \to \mathbb{R}$ is *normalized* if $\mathcal{L}_A(1) = 1$, where

$$\mathcal{L}_A \colon \mathrm{Lip}(\Omega, \mathbb{R}) \to \mathrm{Lip}(\Omega, \mathbb{R}) \quad \text{given by} \quad \mathcal{L}_A g(x) = \sum_{\sigma(y)=x} e^{A(y)}\, g(y)$$

is the usual Ruelle-Perron-Frobenius transfer operator (cf. [48, Chapter 2]). We will always assume that potentials are normalized and write $J = e^A > 0$ (or alternatively $A = \log J$) as the Jacobian of the associated probability measure $\mu_A = \mu_{\log J}$. That is, $\mathcal{L}^*_{\log J}(\mu_{\log J}) = \mu_{\log J}$ and, equivalently, $\mu_{\log J}(\sigma(E)) = \int_E J\, d\mu_{\log J}$ for every measurable set $E \subset \Omega$ so that $\sigma\,|_E$ is injective. We consider the Lipschitz norm $|\,.\,| = \|\cdot\|_\infty + |\cdot|_{Lip}$ on the space of Lipschitz continuous potentials A, where $|A|_{Lip} = \sup_{x\neq y} \frac{|A(x)-A(y)|}{d_\Omega(x,y)}$. Moreover, it is a classical result in thermodynamic formalism (see e.g. [48]) that the following variational principle holds

$$\sup_{\mu\in\mathcal{M}_\sigma(\Omega)} \left\{ h(\mu) + \int \log J\, d\mu \right\} = 0 \tag{7}$$

for any Lipschitz and normalized potential $\log J$. A particularly relevant context is given by the space of stationary Markov probability measures on shift spaces (cf. Example 2.1). One should emphasize that, replacing the metric on Ω, it is possible to deal instead with the space of Lipschitz continuous potentials (cf. [48, Chapter 1]).

In the direct observation context, the procedure of taking ergodic time series on the Bayesian inference is determined by $T = \sigma$ and a fixed T-invariant Gibbs measure ν on Ω associated to a normalized potential $\log J$. The time series will describe the interaction (expressed in terms of the loss functions) over certain families of potentials (and Gibbs measures) which are parameterized on a compact set, where the sampling will occur. More precisely, consider the set of parameters $\Theta \subset \mathbb{R}^k$ of the form

$$\Theta = [a_1, b_1] \times [a_2, b_2] \times ... \times [a_k, b_k],$$

endowed with the metric d_Θ given by $d_\Theta(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|$, $\forall \theta_1, \theta_2 \in \Theta$, and denote by $f : \Theta \to \mathcal{G} \subset \mathcal{M}_\sigma(\Omega)$ a continuous function of potentials parameterized over Θ such that:

(1) f is an homeomorphism over its image;
(2) for each θ the potential $f(\theta)$ is normalized (we use the notation $f(\theta) = \log J_\theta$).

The assumptions guarantee that for each $\theta \in \Theta$ there exists a unique invariant Gibbs measure μ_θ with respect to the associated normalized potential $f(\theta)$, and that these vary continuously in the weak* topology. Moreover, as the parameter space Θ is compact and $f : \Theta \to \mathcal{G}$ is a continuous function (expressed in the form $f(\theta) = \log J_\theta$, where f is a continuous function on $\theta \in \Theta$ and $J_\theta > 0$), we deduce that the quotient

$$\frac{J_{\theta_1}(x)}{J_{\theta_2}(x)} > 0 \tag{8}$$

is uniformly bounded for every $x \in \Omega$ and all $\theta_1, \theta_2 \in \Theta$.

Remark 1.1. At this moment we are not requiring the probability measure ν of the observed system $Y = \Omega$ to belong to the family of probability measures $(\mu_\theta)_{\theta \in \Theta}$. We refer the reader to Example 2.5 for an application in the special case that $\nu = \mu_{\theta_0}$, for some $\theta_0 \in \Theta$.

The statistics is described by an *a priori* Bayes probability measure Π_0 on the space of parameters Θ satisfying *Hypothesis A*:

$$\Pi_0(dz_1, dz_2, ..., dz_k) = \Pi_0(d\theta) \quad \text{is a fixed continuous strictly positive density fully supported on the compact set } \Theta. \tag{A}$$

In many examples the *a priori* measure appears as the Lebesgue or an equidistributed measure on the parameter space. We refer the reader to Section 2 for examples.

The previous full support assumption not only expresses the uncertainty on the choice of the parameters, as it ensures that all parameters in Θ will be taken into account in the inference independently of the initial belief (distribution of Π_0). In this case of direct observations of Gibbs measures, let $\theta_0 \in \Theta$ be fixed. The probability measure μ_{θ_0} will play the role of the measure ν (on the observed system Y) considered abstractly on the previous subsection. We will consider the *loss functions* $\ell_n : \Theta \times \Omega \times \Omega \to \mathbb{R}$, $n \geqslant 1$, given by

$$\ell_n(\theta, x, y) = \begin{cases} \log\left(\mu_{\theta_0}\left(C_n(y)\right)\right) & , \text{if } x \in C_n(y) \\ +\infty & , \text{if } x \notin C_n(y). \end{cases} \tag{9}$$

If one denotes by $\mathbf{1}_{C_n(y)}$ the indicator function of the n-cylinder set centered at y and defined by $C_n(y) = \{(x_j)_{j\geqslant 1} \colon x_j = y_j, \forall 1 \leqslant j \leqslant n\}$, such choice of loss functions ensures that

$$\begin{aligned} Z_n(y) &= \int_\Theta \int e^{-\,\ell_n(\theta,x,y)}\, d\mu_\theta(x)\, d\Pi_0(\theta) = \int_\Theta \int_{C_n(y)} e^{-\,\ell_n(\theta,x,y)}\, d\mu_\theta(x)\, d\Pi_0(\theta) \\ &= \int_\Theta \int \frac{\mathbf{1}_{C_n(y)}(x)}{\mu_{\theta_0}\,(\,C_n(y)\,)}\, d\mu_\theta(x)\, d\Pi_0(\theta) = \int_\Theta \frac{\mu_\theta\,(\,C_n(y)\,)}{\mu_{\theta_0}\,(\,C_n(y)\,)}\, d\Pi_0(\theta) \end{aligned}$$

for each $y \in Y$. Therefore, using equalities (25) and (27) (see Subsection 3.1 below), Jensen inequality and the monotone convergence theorem, one obtains that

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n} \log Z_n(y) \; &= \limsup_{n\to\infty} \frac{1}{n} \log \int_\Theta \frac{\mu_\theta\,(\,C_n(y)\,)}{\mu_{\theta_0}\,(\,C_n(y)\,)} d\Pi_0(\theta) \\ &\geqslant \limsup_{n\to\infty} \frac{1}{n} \int_\Theta \log \frac{\mu_\theta(C_n(y))}{\mu_{\theta_0}(C_n(y))}\, d\Pi_0(\theta) \\ &= -\int_\Theta h(\mu_{\theta_0} \mid \mu_\theta)\, d\Pi_0(\theta) \\ &= \int_\Theta \Big[\, h(\mu_{\theta_0}) + \int_\Omega \log J_\theta \, d\mu_{\theta_0} \,\Big]\, d\Pi_0(\theta) \end{aligned} \tag{10}$$

for μ_{θ_0}-almost every y.

On this context of direct observation we are interested in estimating the family of *a posteriori measures*

$$\Pi_n(E \mid y) \;=\; \frac{\int_E \mu_\theta(C_n(y))\, d\Pi_0(\theta)}{\int_\Theta \mu_\theta(C_n(y))\, d\Pi_0(\theta)}, \tag{11}$$

on Borel sets $E \subset \Theta$ which *do not contain* θ_0 and $y \in \Omega$ is a point chosen according to μ_{θ_0}. An equivalent form of (11) which may be useful is

$$\Pi_n(E \mid y) \;=\; \frac{\int_E \frac{\mu_\theta\,(\,C_n(y)\,)}{\mu_{\theta_0}\,(\,C_n(y)\,)}\, d\Pi_0(\theta)}{\int_\Theta \frac{\mu_\theta\,(\,C_n(y)\,)}{\mu_{\theta_0}\,(\,C_n(y)\,)}\, d\Pi_0(\theta)}. \tag{12}$$

Actually, given such kind of $E \subset \Theta$, one can ask wether the limit

$$\lim_{n\to\infty} \Pi_n(E \mid y) = \lim_{n\to\infty} \frac{1}{n} \log \frac{\int_E \mu_\theta(C_n(y)) d\Pi_0(\theta)}{\int_\Theta \mu_\theta(C_n(y)) d\Pi_0(\theta)} \tag{13}$$

exists for μ_{θ_0}-almost every y. The following result gives an affirmative answer to this question.

Theorem A. *In the previous context,*

$$\lim_{n\to\infty} \Pi_n(\cdot \mid y) = \delta_{\theta_0}, \quad \textit{for } \mu_{\theta_0}\textit{-a.e. } y \in \Omega.$$

Moreover the convergence is exponentially fast: for every $\delta > 0$ there exists a constant $c_\delta > 0$ so that the ball B_δ of radius δ around θ_0 satisfies $|\Pi_n(B_\delta \mid y) - 1| \leqslant e^{-c_\delta\, n}$ for every large $n \geqslant 1$.

The previous result guarantees that the parameter θ_0, or equivalently the ergodic measure μ_{θ_0}, is identified as the limit of the Bayesian inference process determined by the loss function (9). This result arises as a consequence of the quantitative estimates in Theorem 4.1, given in the proofs section below. The direct observation of Gibbs measures was also considered in [47, Section 2.1] although with a different approach. For a parameterized family of loss functions of

the form $\beta \cdot \ell_n(\theta, x, y)$ it is also analyzed in section 3.7 of [47] the zero temperature limit (ground states). This is a topic which can be associated to ergodic optimization. Our results are related in some sense to the so called Maximum Likelihood Identification described [15, 14, 17, 18, 16]

The previous context fits in the wider scope of non-additive thermodynamic formalism, using almost-additive sequences of continuous functions (see Subsection 3.2 for the definition). Indeed, the loss functions $(\ell_n)_{n\geqslant 1}$ described in (9) form an almost-additive family (cf. Definition 3.2 and Lemma 3.3). Furthermore, we will consider loss functions $\ell_n : \Theta \times \Omega \times Y \to \mathbb{R}$ which form an almost-additive sequence of continuous functions, and for which one can write

$$\ell_n(\theta, x, y) = -\varphi_n(\theta, x, y), \tag{14}$$

where $\varphi_n : \Theta \times \Omega \times Y \to \mathbb{R}_+$ are continuous observables satisfying:

(A1) for ν-almost every $y \in Y$ the following limit exists

$$\Gamma^y(\theta) := \lim_{n\to\infty} \frac{1}{n} \log \int_\Omega e^{\varphi_n(\theta,x,y)}\, d\mu_\theta(x),$$

(A2) $\Theta \ni \theta \mapsto \Gamma^y(\theta)$ is upper semicontinuous.

Given $y \in Y$ and the loss functions ℓ_n satisfying (A1)-(A2), the *a posteriori* measures are

$$\Pi_n(E \mid y) \;=\; \frac{\int_E \int_\Omega e^{\varphi_n(\theta,x,y)}\, d\mu_\theta(x)\, d\Pi_0(\theta)}{\int_\Theta \int_\Omega e^{\varphi_n(\theta,x,y)}\, d\mu_\theta(x)\, d\Pi_0(\theta)}. \tag{15}$$

Remark 1.2. *The expression appearing in assumption (A1), which resembles the logarithm of the moment generating function for i.i.d. random variables, is in special cases referred to as the free energy function. Consider the special case where $T = \sigma$ is the shift, ν is an equilibrium state with respect to a Lipschitz continuous potential ψ and $\varphi_n(\theta, x, y) = \varphi_{n,1}(\theta, x) + \varphi_{n,2}(\theta, y)$, where $\varphi_{n,1}(\theta, x) = \sum_{j=0}^{n-1} \phi_\theta \circ \sigma^j(x)$, $\phi_\theta : \Omega \to \mathbb{R}$ is Lipschitz continuous and $(\varphi_{n,2}(\theta, \cdot))_{n\geqslant 1}$ is sub-additive. Then using the fact that the pressure function defined over the space of Lipschitz continuous observables is Gateaux differentiable and the sub-additive ergodic theorem one obtains that*

$$\begin{aligned}
\Gamma^y(\theta) &= \lim_{n\to\infty} \frac{1}{n} \log \int_\Omega e^{\sum_{j=0}^{n-1} \phi_\theta(\sigma^j(x))}\, d\mu_\theta(x) + \inf_{n\geqslant 1} \frac{1}{n} \int \varphi_{n,2}(\theta, \cdot)\, d\nu \\
&= P_{top}(\sigma, \log J_\theta + \phi_\theta) - P_{top}(\sigma, \log J_\theta) + \inf_{n\geqslant 1} \frac{1}{n} \int \varphi_{n,2}(\theta, \cdot)\, d\nu \\
&= P_{top}(\sigma, \log J_\theta + \phi_\theta) + \inf_{n\geqslant 1} \frac{1}{n} \int \varphi_{n,2}(\theta, \cdot)\, d\nu,
\end{aligned}$$

for ν-almost every $y \in \Omega$, hence it is independent of y. We refer the reader to Subsection 3.2 for the concept of topological pressure and further information.

The following result guarantees that the previous Bayesian inference procedure accumulates on the set of probability measures on the parameter space Θ that maximize the free energy function Γ^y. By assumption (A2) the set $\operatorname{argmax} \Gamma^y := \{\theta_0 \in \Theta \colon \Gamma^y(\theta) \leqslant \Gamma^y(\theta_0), \forall \theta \in \Theta\}$ is non-empty. Then we prove the following:

Theorem B. *Assume ℓ_n is a loss function of the form (14) satisfying assumptions (A1)-(A2). There exists a full ν-measure subset $Y' \subset Y$ so that, for any $\delta > 0$ and $y \in Y'$,*

$$\lim_{n\to\infty} \Pi_n(\Theta \setminus B_\delta^y \mid y) = 0 \quad \textit{where} \quad B_\delta^y = \{\theta \in \Theta \colon d_\Theta\big(\theta, \operatorname{argmax} \Gamma^y\big) > \delta\} \tag{16}$$

is the open δ-neighborhood of the maximality locus of Γ^y. In particular, if $y \in Y'$ is such that $\Theta \ni \theta \mapsto \Gamma^y(\theta)$ has a unique point of maximum $\theta_0^y \in \Theta$ then $\lim_{n\to\infty} \Pi_n(\cdot \mid y) = \delta_{\theta_0^y}$.

Finally, inspired by the log-likelihood estimators in the context of Bayesian statistics it is also natural to consider the loss functions $\ell_n : \Theta \times X \times Y \to \mathbb{R}$ defined by

$$\ell_n(\theta, x, y) = -\log \varphi_n(\theta, x, y) \tag{17}$$

associated to an almost additive sequence $\Phi = (\varphi_n)_{n\geqslant 1}$ of continuous observables $\varphi_n : \Theta \times X \times Y \to \mathbb{R}_+$ satisfying

(H1) for each $\theta \in \Theta$ and $x \in X$ there exists a constant $K_{\theta,x} > 0$ so that, for every $y \in Y$,

$$\varphi_n(\theta, x, y) + \varphi_m(\theta, x, T^n(y)) - K_{\theta,x} \leqslant \varphi_{m+n}(\theta, x, y) \leqslant \varphi_n(\theta, x, y) + \varphi_m(\theta, x, T^n(y)) + K_{\theta,x}$$

(H2) $\int K_{\theta,x} d\mu_\theta(x) < \infty$ for every $\theta \in \Theta$.

In this context, the loss functions induce the *a posteriori* measures

$$\Pi_n(E \mid y) \;=\; \frac{\int_E \psi_n(\theta, y)\, d\Pi_0(\theta)}{\int_\Theta \psi_n(\theta, y)\, d\Pi_0(\theta)}, \quad \text{where} \quad \psi_n(\theta, y) = \int_\Omega \varphi_n(\theta, x, y)\, d\mu_\theta(x). \tag{18}$$

Therefore, even though the loss functions are not almost-additive, due to the logarithmic term, we have the following result for the latter non-additive loss functions:

Theorem C. *Assume that the loss function of the form (17) satisfies assumptions (H1)-(H2) above. There exists a non-negative function $\psi_* : \Theta \to \mathbb{R}_+$ (depending on $\Psi^\theta = (\psi_n(\theta, \cdot))_{n\geqslant 1}$) so that for ν-almost every $y \in Y$ the* a posteriori *measures $(\Pi_n(\cdot \mid y))_{n\geqslant 1}$ are convergent and*

$$\Pi_n(\cdot \mid y) \;=\; \frac{\int \psi_n(\theta, y)\, d\Pi_0(\theta)}{\int_\Theta \psi_n(\theta, y)\, d\Pi_0(\theta)} \longrightarrow \Pi_*(\cdot) := \frac{(\psi_* \Pi_0)(\cdot)}{(\psi_* \Pi_0)(\Theta)}$$

as $n \to \infty$. Moreover, if $T = \sigma$ is a subshift of finite type, $\nu \in \mathcal{M}_\sigma(\Omega)$ is a Gibbs measure with respect to a Lipschitz continuous potential and $\inf_{\theta\in\Theta} \psi_*(\theta) > 0$ then for each $g \in C(\Theta, \mathbb{R})$ there exists $c > 0$ so that

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n} \;\log\, \nu(\{\, y \in \Omega : \Big| \int g\, d\Pi_n(\cdot \mid y) - \int g\, d\Pi_* \Big| \geqslant \delta\}) \\ \leqslant \sup_{\theta\in\Theta} \sup_{\{\eta:\, |\mathcal{F}(\eta,\Psi^\theta) - \psi_*(\theta)| \geqslant c\delta\}} \Big\{ -P(\sigma, \varphi) + h_\eta(\sigma) + \int \varphi\, d\eta \Big\}, \end{aligned} \tag{19}$$

where $\mathcal{F}(\eta, \Psi^\theta) := \lim_{n\to\infty} \frac{1}{n} \int \psi_n(\theta, \cdot)\, d\eta$. If, additionally, the map $\Theta \ni \theta \mapsto \mathcal{F}(\eta, \Psi^\theta)$ is continuous for each $\eta \in \mathcal{M}_\sigma(\Omega)$ then the right hand-side in (19) is strictly negative.

The previous theorem ensures that, in the context of loss functions of the form (17) satisfying properties (H1) and (H2) above, the a posteriori measures do converge exponentially fast to a probability measure on the parameter space which is typically fully supported. We refer the reader to Example 2.2 for more details in the special case the loss function depends exclusively on one parameter.

Remark 1.3. *For completeness, let us mention that the results by Kifer [38] suggest that level-2 large deviations estimates (ie, the rate of convergence of $\Pi_n(\cdot \mid y)$ to Π_* on the space of probability measures on Θ) are likely to hold under the assumption that the limit $\lim_{n\to\infty} \frac{1}{n} \log \int e^{\varphi_n}\, d\nu$ exists for all almost-additive sequences $\Phi = (\varphi_n)_{n\geqslant 1}$ of continuous observables and defines a*

non-additive free energy function which is related to the non-additive topological pressure. This extrapolates the scope of our interest here.

2. Examples

In what follows we give some examples which illustrate the intuition and utility of the Bayesian inference and also the meaning of the *a priori* measures.

Example 2.1. The space of all stationary Markov probability measures μ in $\Omega = \{1,2\}^{\mathbb{N}}$ is described by the space of column stochastic transition matrices P with all positive entries. These matrices P can be parameterized by the open square $\Theta = (0,1) \times (0,1)$ through the parameterization

$$M_{(a,b)} = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} = \begin{pmatrix} a & 1-b \\ 1-a & b \end{pmatrix}, \qquad (a,b) \in (0,1) \times (0,1).$$

In this case the associated normalized Jacobian $J_{(a,b)}(w)$ has constant value on cylinders of size two. More precisely, for w on the cylinder $[i,j] \subset \Omega$ we get $J = \frac{\pi_i P_{i,j}}{\pi_j}$, where (π_1, π_2) is the initial invariant probability vector. For each value (a,b) denote by $\mu_{(a,b)}$ the stationary Markov probability measure associated to the stochastic matrix $M_{(a,b)}$. In this case we get that $h(\mu_{(a,b)}) + \int \log J_{(a,b)}\, d\mu_{(a,b)} = 0$ and $\mathcal{L}^*_{\log J_{(a,b)}}(\mu_{(a,b)}) = \mu_{(a,b)}$ (see [41, 53]). We refer the reader to [23, 24, 25, 56] for applications of the use of the maximum likelihood estimator in this context of Markov probability measures. One possibility would be to take the probability measure Π_0 on the Θ space as the Lebesgue probability measure on $(0,1) \times (0,1)$. Different choices of loss functions would lead to different solutions for the claim of Theorem B.

The first of the following examples are very simple and illustrate some trivial contexts. Whenever the parameter space Θ (or Y) is a singleton the Bayesian inference is trivial, hence it carries no information. The first example we shall consider is when the loss function depends exclusively on a single variable. Nevertheless, as loss functions are non-additive, these results could not be handled with the previous literature in the subject.

Example 2.2. Assume that $\Theta \subset \mathbb{R}^d$ is a compact set, $Y = \Omega$ and $T = \sigma : \Omega \to \Omega$ is a subshift of finite type. In the case that the loss functions $\ell_n \colon \Theta \times \Omega \times Y \to \mathbb{R}$ are generated by an almost-additive sequence of continuous observables $\Phi = (\varphi_n)_{n \geqslant 1}$ by $\ell_n(\theta, x, y) = -\log \varphi_n(y)$ which are independent of θ and x, the loss function gives no information on the parameter space. For that reason it is natural that the *a posteriori* measures are

$$\Pi_n(E \mid y) \;=\; \frac{\int_E \varphi_n(y)\, d\Pi_0(\theta)}{\int_\Theta \varphi_n(y)\, d\Pi_0(\theta)} = \Pi_0(E) \tag{20}$$

for every ergodic time series $y, T(y), \ldots, T^{n-1}(y) \in Y$.

Now, assuming alternatively that the loss function is given by $\ell_n(\theta, x, y) = -\log \varphi_n(\theta)$, which is independent on both x and y, a simple computation shows that

$$\Pi_n(E \mid y) \;=\; \frac{\int_E \varphi_n(\theta)\, d\Pi_0(\theta)}{\int_\Theta \varphi_n(\theta)\, d\Pi_0(\theta)}. \tag{21}$$

In this case the loss function neglects the observable dynamical system T, hence the a posteriori measures are independent of the ergodic time series. Yet, as the family Φ is almost-additive it is easy to check that there exists $C > 0$ so that $\{\varphi_n + C\}_{n \geqslant 1}$ is sub-additive. In particular, a

simple application of Fekete's lemma (cf. Lemma 3.2) ensures that the limit $\lim_{n\to\infty} \frac{\phi_n(\theta)}{n}$ does exists and coincides with $\phi_*(\theta) := \inf_{n\geqslant 1} \frac{\phi_n(\theta)}{n}$, for every $\theta \in \Theta$. In consequence,

$$\lim_{n\to\infty} \Pi_n(E \mid y) = \lim_{n\to\infty} \frac{\int_E \frac{\varphi_n(\theta)}{n}\, d\Pi_0(\theta)}{\int_\Theta \frac{\varphi_n(\theta)}{n}\, d\Pi_0(\theta)} = \Pi(E) := \frac{\int_E \varphi_*(\theta)\, d\Pi_0(\theta)}{\int_\Theta \varphi_*(\theta)\, d\Pi_0(\theta)}, \tag{22}$$

independently of the point y taken according to the ergodic theorem. In particular the limit measure Π is fully supported on Θ if and only if $\varphi_*(\theta) > 0$ for every $\theta \in \Theta$.

Finally, for each $n \geqslant 1$ and almost-additive sequence of continuous observables $\Phi = (\varphi_n)_{n\geqslant 1}$ on X, consider the loss function

$$\ell_n(\theta, x, y) = -\log \varphi_n(x),$$

In this case a simple computation shows that one obtains *a posteriori* measures

$$\Pi_n(E \mid y) \;=\; \frac{\int_E \psi_n(\theta)\, d\Pi_0(\theta)}{\int_\Theta \psi_n(\theta)\, d\Pi_0(\theta)}, \tag{23}$$

where the sequence $\psi_n(\theta) = \int_\Omega \varphi_n(x)\, d\mu_\theta(x)$ is almost additive. Indeed, the σ-invariance of μ_θ and the almost-additivity condition $\varphi_n(x)+\varphi_m(\sigma^n(x))-C \leqslant \varphi_{m+n}(x) \leqslant \varphi_n(x)+\varphi_m(\sigma^n(x))+C$ ensures that $\psi_n(\theta)+\psi_m(\theta)-C \leqslant \psi_{m+n}(\theta) \leqslant \psi_n(\theta)+\psi_m(\theta)+C$ for every $m, n \geqslant 1$ and $\theta \in \Omega$. Hence, even though the feed of information is given through the x-variable, the *a posteriori* measures are of the form (20), and their convergence is described by Lemma 3.2. In particular, this example shows that the situation is much simpler to describe when the loss functions depend exclusively on a single variable.

In the following two simple examples, we will make explicit computations on the limit of posterior distributions which shows that the assumption (A) on the space of parameters and a priori distribution cannot be removed. In particular, these will show that the posterior distributions $\Pi_n(\cdot \mid y)$ may converge but not for a Dirac measure on the parameter θ_0 corresponding to the measure with respect to which the ergodic time series was taken.

Example 2.3. Set $\Theta = \{-1, 1\}$, $T = \sigma : \{0,1\}^{\mathbb{N}} \to \{0,1\}^{\mathbb{N}}$ be the full shift and let $\mathbb{B}(a, b)$ denote the Bernoulli measure with $\nu[0] = a$ and $\nu[1] = b$, for $a + b = 1$, $0 < a < 1$. If $\phi : \{0,1\}^{\mathbb{N}} \to \mathbb{R}$ is a locally constant normalized potential so that $\phi\,|_{[0]} = c < 0$ then it is not hard to deduce (see e.g. [9]) that $\phi\,|_{[1]} = \log(1 - e^c)$ and the unique equilibrium state for σ with respect to ϕ is the probability measure $\mathbb{B}(e^c, 1 - e^c)$. Assume that $\mu_{-1} = \mathbb{B}(\frac{1}{3}, \frac{2}{3})$ and $\mu_1 = \mathbb{B}(\frac{2}{3}, \frac{1}{3})$ which are the unique equilibrium states for the potentials

$$\phi_{-1}(x) := \begin{cases} -\log 3, x \in [0] \\ -\log \frac{3}{2}, x \in [1] \end{cases} \qquad \text{and} \quad \phi_1(x) := \begin{cases} -\log \frac{3}{2}, x \in [0] \\ -\log 3, x \in [1], \end{cases}$$

respectively. Take $\Pi_0 = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ and $\nu = \mathbb{B}(\frac{1}{2}, \frac{1}{2})$ and notice that ν does not belong to the family $(\mu_\theta)_\theta$. On the context of direct observation we are interested in describing the *a posteriori* measures

$$\Pi_n(E \mid y) \;=\; \frac{\int_E \mu_\theta(C_n(y))\, d\Pi_0(\theta)}{\int_\Theta \mu_\theta(C_n(y))\, d\Pi_0(\theta)},$$

where y was taken according to the ergodic theorem for ν. The ergodic theorem ensures that

$$\lim_{n\to\infty} \frac{1}{n} \#\{0 \leqslant j \leqslant n-1 \colon \sigma^j(y) \in [0]\} = \lim_{n\to\infty} \frac{1}{n} \#\{0 \leqslant j \leqslant n-1 \colon \sigma^j(y) \in [1]\} = \frac{1}{2}$$

for ν-almost every y. The Bernoulli property of $\mu_{\pm 1}$ then implies that, for ν-a.e. y,

$$\frac{\mu_1(C_n(y))}{\mu_{-1}(C_n(y))} \to 1 \quad \text{as } n \to \infty$$

and, consequently, the sequence of probability measures $\Pi_n(\cdot \mid y)$ on $\{-1, 1\}$ is convergent as

$$\lim_{n\to\infty} \Pi_n(\{\pm 1\} \mid y) = \lim_{n\to\infty} \frac{\mu_{\pm 1}(C_n(y))}{\mu_{-1}(C_n(y)) + \mu_1(C_n(y))} = \frac{1}{2}.$$

In other words, $\lim_{n\to\infty} \Pi_n(\cdot \mid y) = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1 = \Pi_0$. This convergence reflects the fact that $\int \phi_{-1}\, d\nu = \int \phi_1\, d\nu$. Finally, it is not hard to check that for any a priori measure $\Pi_0 = \alpha\delta_{-1} + (1-\alpha)\delta_1$ for some $0 < \alpha < 1$ it still holds that $\lim_{n\to\infty} \Pi_n(\cdot \mid y) = \Pi_0$.

Example 2.4. In the context of Example 2.3, assume that the ergodic time series was taken according to a non-symmetric Bernoulli measure $\hat{\nu} = \mathbb{B}(\alpha, 1-\alpha)$ for some $0 < \alpha < \frac{1}{2}$. The ergodic theorem guarantees that, for $\hat{\nu}$-a.e. y,

$$\frac{\mu_1(C_n(y))}{\frac{2^{\alpha n}}{3^n}} \to 1 \quad \text{and} \quad \frac{\mu_{-1}(C_n(y))}{\frac{2^{(1-\alpha)n}}{3^n}} \to 1 \quad \text{as } n \to \infty$$

and, consequently, $\mu_1(C_n(y))/\mu_{-1}(C_n(y)) \to 0$ as $n \to \infty$. Altogether we get

$$\lim_{n\to\infty} \Pi_n(\{1\} \mid y) = \lim_{n\to\infty} \frac{\mu_1(C_n(y))}{\mu_{-1}(C_n(y)) + \mu_1(C_n(y))} = \lim_{n\to\infty} \frac{\frac{\mu_1(C_n(y))}{\mu_{-1}(C_n(y))}}{1 + \frac{\mu_1(C_n(y))}{\mu_{-1}(C_n(y))}} = 0$$

and $\lim_{n\to\infty} \Pi_n(\{-1\} \mid y) = 1$. In other words, $\lim_{n\to\infty} \Pi_n(\cdot \mid y) = \delta_{-1}$ for $\hat{\nu}$-almost every y, which reflects the fact that $\int \phi_1\, d\hat{\nu} < \int \phi_{-1}\, d\hat{\nu}$.

Example 2.5. Take $\Theta = [0, 1]$ and let Π_0 be the Lebesgue measure. Take $\log J_0$ and $\log J_1$ two normalized Lipschitz continuous Jacobians, where $J_0, J_1 : \{1, 2, ..., q\}^{\mathbb{N}} \to \mathbb{R}_+$, and consider the family of Lipschitz continuous potentials

$$f_\theta = \log J_\theta := \log(\theta J_1 + (1-\theta)J_0), \qquad \theta \in [0, 1].$$

For each $\theta \in [0, 1]$ let μ_θ be the unique Gibbs measure associated to the Lipschitz continuous potential f_θ (see also section 6 in [30] for a related work). Assume further that the observed probability measure associated to the ergodic time series is $\nu = \mu_{\theta_0}$ for some $\theta_0 \in [0, 1]$. The probability measure Π_0 describes our ignorance of the exact value θ_0 among all possible choices $\theta \in [0, 1]$. For each $n \in \mathbb{N}$ consider a continuous *loss function* $\ell_n : \Theta \times \Omega \times Y \to \mathbb{R}$ expressed as

$$\ell_n((a, b), x, y) = -\sum_{j=0}^{n-1} \log J_\theta(\sigma^j(x)) + \sum_{j=0}^{n-1} \log J_{\theta_0}(\sigma^j(y)) + \theta \log \theta - \theta \log \theta_0\,.$$

Similar expressions are often referred as cross-entropy loss functions. By compactness of the parameter space Θ we conclude that the third and fourth expressions above are uniformly bounded, hence $(\ell_n)_{n\geqslant 1}$ forms an almost-additive family on the y-variable, hence it fits in the context of Theorem B. In particular we conclude that the *a posteriori* measures $\Pi_n(\cdot \mid y)$ converge to the probability measure μ_{θ_0} as n tends to infinity, for μ_{θ_0}-almost every y. Alternatively, consider the continuous *loss function* $\ell_n : \Theta \times \Omega \times Y \to \mathbb{R}$ given by

$$\ell_n((a, b), x, y) = -\sum_{j=0}^{n-1} \log J_\theta(\sigma^j(x)) + \sum_{j=0}^{n-1} \log J_{\theta_0}(\sigma^j(y)) - \| \theta - \theta_0 \|^2.$$

The minimization of $-\ell_n$ corresponds, in rough terms, to what is known in statistics as the minimization of the mean squared error on the set of parameters. As the previous loss function is also almost-additive on the y-variable, Theorem B ensures that the corresponding *a posteriori* measures $\Pi_n(\cdot \mid y)$ converge exponentially fast to the ergodic probability measure μ_{θ_0}, as n tends to infinity, μ_{θ_0}-almost everywhere (we refer the reader to [47] where the methods which were developed there can provide an alternative argument leading to the same conclusion).

Example 2.6. Let $\sigma : \{1,2\}^{\mathbb{N}} \to \{1,2\}^{\mathbb{N}}$ be the full shift and for each $\theta = (\theta_1, \theta_2)$ in the parameter space $\Theta := [-\varepsilon, \varepsilon]^2$ let μ_θ be a continuous family of Bernoulli measures. These are equilibrium states for a continuous family of potentials. Consider also the locally constant linear cocycle $A_\theta : \Omega \to SL(2,\mathbb{R})$ given by

$$A_\theta \mid_{[i]} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix}, \quad \text{for every } i = 1, 2.$$

Given $n \geqslant 1$ and $(x_1, \ldots, x_n) \in \{1,2\}^n$, take the matrix product

$$A_\theta^{(n)}(x_1, \ldots, x_n) := A_{\theta_{x_n}} \ldots A_{\theta_{x_2}} A_{\theta_{x_1}}.$$

The limit $\lambda_{\theta,i} := \lim_{n\to\infty} \frac{1}{n} \log \|A_\theta^{(n)}(x)\, v_i\|$ is the largest Lyapunov exponent along the orbit of x, it is well defined for ν-almost every x and independs on the vector $v_i \in E^i_{\theta,x} \setminus \{0\}$, $(i = \pm)$ (cf. Subsection 3.2.3 for more details). Somewhat dual to the context of joint spectral radius [8], the problem here is the selection of a certain Gibbs measure from the information on the norm of the products of matrices, along orbits of typical points. More precisely, take the loss function $\ell_n(\theta, x, y) = -\log \|A_\theta^{(n)}(x)\|$ and notice that, for ν-almost every $y \in Y$ and every $\theta \in \Theta$,

$$\begin{aligned}
\int_\Omega e^{\varphi_{n+m}(\theta,x,y)}\, d\mu_\theta(x) &= \int_\Omega \|A_\theta^{(n+m)}(x)\|\, d\mu_\theta(x) = \sum_{C_{n+m}(z)} \|A_\theta^{(n+m)}(z)\|\, \mu_\theta(C_{n+m}(z)) \\
&\leqslant \sum_{C_n(z)} \|A_\theta^{(n)}(z)\| \|A_\theta^{(m)}(\sigma^n(z))\|\, \mu_\theta(C_n(z))\, \mu_\theta(C_m(\sigma^n(z))) \\
&\leqslant \int_\Omega e^{\varphi_n(\theta,x,y)}\, d\mu_\theta(x) \cdot \int_\Omega e^{\varphi_m(\theta,x,y)}\, d\mu_\theta(x)
\end{aligned}$$

for every $m, n \geqslant 1$, where we used that μ_θ is a σ-invariant Bernoulli measure. In particular, Fekete's lemma implies that the following limit exists

$$\lim_{n\to\infty} \frac{1}{n} \log \int_\Omega e^{\varphi_n(\theta,x,y)}\, d\mu_\theta(x) = \inf_{n\geqslant 1} \frac{1}{n} \log \int_\Omega \|A_\theta^{(n)}(x)\|\, d\mu_\theta(x).$$

exists and independs on y. As the right hand-side above is the infimum of continuous functions on the parameter θ, the limit function $\Theta \ni \theta \mapsto \Gamma^y(\theta)$ is upper semicontinuous. We remark that $\theta_0 = (0,0)$ is the unique parameter for which the Lyapunov exponent is the largest possible (see Lemma 3.5). Hence, as assumptions (A1) and (A2) are satisfied, Theorem B implies that

$$\Pi_n(E \mid y) \;=\; \frac{\int_E \int \|A_\theta^{(n)}(x)\| d\mu_\theta(x)\, d\Pi_0(\theta)}{\int_\Theta \int \|A_\theta^{(n)}(x)\| d\mu_\theta(x)\, d\Pi_0(\theta)} \longrightarrow \begin{cases} 1, \text{if } (0,0) \in E \\ 0, \text{otherwise} \end{cases}$$

for every measurable subset $E \subset \Theta$, In other words, the a posteriori measures converge to the Dirac measure $\delta_{(0,0)}$. In particular, one has posterior consistency in the problem of determining the measure with largest Lyapunov exponent.

Alternatively, taking the loss function $\ell_n(\theta,x,y) = -\varphi_n(\theta,x,y) = -\log\|A_\theta^{(n)}(y)\|$, note that the *a posteriori* measures are given by

$$\Pi_n(E \mid y) \;=\; \frac{\int_E \|A_\theta^{(n)}(y)\|\, d\Pi_0(\theta)}{\int_\Theta \|A_\theta^{(n)}(y)\|\, d\Pi_0(\theta)}$$

and that, by the Oseledets theorem and the sub-additive ergodic theorem, the limit

$$\lim_{n\to\infty}\frac{1}{n}\log\int_\Omega e^{\varphi_n(\theta,x,y)}\,d\mu_\theta(x) = \lim_{n\to\infty}\frac{1}{n}\log\|A_\theta^{(n)}(y)\| = \inf_{n\geqslant 1}\frac{1}{n}\int\log\|A_\theta^{(n)}(\cdot)\|\,d\nu$$

for ν-almost every $y\in Y$. The map $\Theta\ni\theta\mapsto\Gamma^y(\theta) := \inf_{n\geqslant1}\frac{1}{n}\int\log\|A_\theta^{(n)}(\cdot)\|\,d\nu$ is upper semicontinuous because it is the infimum of continuous maps. In particular, Theorem B implies once more that for ν-almost every $y\in Y$

$$\lim_{n\to\infty}\Pi_n(\cdot\mid y) \;=\; \lim_{n\to\infty}\frac{\int_\cdot\|A_\theta^{(n)}(y)\|\,d\Pi_0(\theta)}{\int_\Theta\|A_\theta^{(n)}(y)\|\,d\Pi_0(\theta)} = \delta_{(0,0)}$$

Example 2.7. In the context of Example 2.6, noticing that all matrices are in $SL(2,\mathbb{R})$ it makes sense to consider alternatively the loss function $\ell_n(\theta,x,y) = -\log\varphi_n(\theta,x,y) = -\log\log\|A_\theta^{(n)}(y)\|$, and to observe that $\varphi_n(\theta,x,y)$ is almost-additive, meaning it satisfies (H1)-(H2) with a constant K uniform on θ. The loss functions induce the *a posteriori* measures

$$\Pi_n(E\mid y) \;=\; \frac{\int_E\int\log\|A_\theta^{(n)}(x)\|d\mu_\theta(x)\,d\Pi_0(\theta)}{\int_\Theta\int\log\|A_\theta^{(n)}(x)\|d\mu_\theta(x)\,d\Pi_0(\theta)}. \tag{24}$$

A simple computation involving Fekete's lemma guarantees that, for each $\theta\in\Theta$, the annealed Lyapunov exponent

$$\lambda(\theta) := \lim_{n\to\infty}\frac{1}{n}\int\log\|A_\theta^{(n)}(x)\|d\mu_\theta(x) = \inf_{n\geqslant1}\frac{1}{n}\int\log\|A_\theta^{(n)}(x)\|d\mu_\theta(x)\geqslant 0$$

does exist. Theorem C implies that the *a posteriori* measures (24) converge and

$$\lim_{n\to\infty}\Pi_n(E\mid y) \;=\; \frac{\int_E\lambda(\theta)\,d\Pi_0(\theta)}{\int_\Omega\lambda(\theta)\,d\Pi_0(\theta)}$$

for every measurable subset $E\subset\Theta$. In particular the limit measure is absolutely continuous with respect to the *a priori* measure Π_0 and with density given by the normalized Lyapunov exponent function. Moreover, the continuous dependence of the Lyapunov exponents with respect to the parameter θ implies the exponential large deviations estimates in Theorem C.

3. Preliminaries

3.1. **Relative entropy.** Let us recall some relevant concepts of entropy in the context of shifts. Given $x=(x_1,x_2,...,x_k,...)\in\Omega$ and $n\geqslant1$, recall

$$C_n(x) = \{y\in\Omega\,|\,y_j = x_j, j=1,2,..,n\}$$

the *n-cylinder* in Ω that contains the point x. The concept of relative entropy will play a key role in the analysis. Let $\phi\colon\Omega\to\mathbb{R}$ be a Lipschitz continuous potential and let μ_ϕ be its unique

Gibbs measure, thus ergodic. Following [16, Section 3], given an ergodic probability measure $\mu \in \mathcal{M}_\sigma(\Omega)$ the limit

$$h(\mu \mid \mu_\phi) := \lim_{n\to\infty} \frac{1}{n} \log\Big(\frac{\mu(C_n(x))}{\mu_\phi(C_n(x))}\Big) \tag{25}$$

exists and is non-negative for μ-almost every $x = (x_1, x_2, ..., x_n, ..) \in \Omega$, and it is called the *relative entropy* of μ with respect to μ_ϕ. Notice that any two distinct ergodic probability measures are mutually singular, hence no Radon-Nykodym derivative is well defined. In (25), a sequence of nested cylinder sets is used as an alternative to compute relative entropy when Radon-Nykodym derivatives are not well defined (see [16] for more details). Moreover,

$$h(\mu \mid \mu_\phi) = P_{\text{top}}(\sigma, \phi) - \int \phi \, d\mu - h(\mu) \tag{26}$$

and, by the variational principle and uniqueness of equilibrium states for Lipschitz continuous potentials, $h(\mu \mid \mu_\phi) = 0$ if and only if $\mu = \mu_\phi$ (cf. Subsection 3.2 in [16]). Furthermore, if $\mu = \mu_\psi$ is a Gibbs measure then $h(\mu \mid \mu_\phi) = 0$ if and only if ϕ and ψ are cohomologous, i.e, if there exists a Lipschitz continuous function $u : \Omega \to \mathbb{R}$ so that $\phi = \psi + u \circ \sigma - u$. The relative entropy is also known as the Kullback-Leibler divergence. For proofs of general results on the topic in the context of shifts we refer the reader to [16] and [44] which deal with finite and compact alphabets, respectively. We refer the reader to [34] for an application of Kullback-Leibler divergence in statistics.

Remark 3.1. In the special case that $(\mu_\theta)_{\theta\in\Theta}$ is a parameterized family of Gibbs measures associated to normalized potentials then for μ_θ-almost every $x = (x_1, x_2, ..., x_n, ..) \in \Omega$ we have

$$\frac{\mu_\theta\,(C_n(x))}{\mu_{\theta_0}\,(C_n(x))} \sim e^{-n\,h(\mu_{\theta_0}|\mu_\theta)} \to 0,$$

as $n \to \infty$, whenever f_θ and f_{θ_0} are not cohomologous. Furthermore, as the pressure function is zero in this context the relative entropy $h(\mu_{\theta_0} \mid \mu_\theta)$ can be written as

$$h(\mu_{\theta_0} \mid \mu_\theta) = -h(\mu_{\theta_0}) - \int \log J_\theta \, d\mu_{\theta_0}. \tag{27}$$

Expression (26) allows to obtain uniform estimates on the relative entropy of nearby invariant measures. More precisely:

Lemma 3.1. *Let $\phi : \Omega \to \mathbb{R}$ be a Lipschitz continuous potential and let μ_ϕ be its unique Gibbs measure. Then, for any small $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$\inf_{\mu\in\mathcal{M}_\sigma(\Omega)} \Big\{ h(\mu \mid \mu_\phi) : \; D_\Omega(\mu, \mu_\phi) > \delta \Big\} > \varepsilon.$$

Proof. Fix $\varepsilon > 0$. By the continuity of the map $\mu \mapsto \int \phi \, d\mu$, upper-semicontinuity of the entropy map $\mu \mapsto h(\mu)$ and uniqueness of the equilibrium state, there exists $\delta > 0$ so that any invariant probability measure μ so that $D_\Omega(\mu, \mu_\phi) > \delta$ satisfies $h(\mu) + \int \phi \, d\mu < P_{\text{top}}(\sigma, \phi) - \varepsilon$. This, together with (26) proves the lemma. □

3.2. **Non-additive thermodynamic formalism.** As mentioned before, we are mostly interested in non-additive loss functions which keep some almost additivity condition. Let us recall some of the basic notions associated to the non-additive thermodynamic formalism.

3.2.1. *Basic notions.* There are several notions of non-additive sequences which appear naturally in the description of thermodynamic objects. Let us recall some of these notions.

Definition 3.2. A sequence $\Psi := \{\psi_n\}_{n\geqslant 1}$ of continuous functions $\psi_n : \Omega \to \mathbb{R}$ is called:

(1) *almost additive* if there exists $C > 0$ such that

$$\psi_n + \psi_m \circ \sigma^n - C \leqslant \psi_{m+n} \leqslant \psi_n + \psi_m \circ \sigma^n + C, \quad \forall m, n \geqslant 1;$$

(2) *asymptotically additive* if for any $\xi > 0$ there is a continuous function ψ_ξ so that

$$\limsup_{n\to\infty} \frac{1}{n} \|\psi_n - S_n\psi_\xi\| < \xi;$$

(3) *sub-additive* if

$$\psi_{m+n} \leqslant \psi_m + \psi_n \circ \sigma^m, \quad \forall m, n \geqslant 1.$$

The convergence in the case of constant functions, ie sub-additive sequences is given by the following well known lemma.

Lemma 3.2 (Fekete's lemma)**.** *Let $(a_n)_{n\geqslant 1}$ be a sequence of real numbers so that $a_{n+m} \leqslant a_n + a_m$ for every $n, m \geqslant 1$. Then the sequence $(a_n)_{n\geqslant 1}$ is convergent to $\inf_{n\geqslant 1} \frac{a_n}{n}$.*

In order to recall the variational principle and equilibrium states for sequences of dynamical systems we need to obtain an almost sure convergence. Given a probability measure $\rho \in \mathcal{M}(\Omega)$, Kingman's sub-additive ergodic theorem ensures that any almost additive or sub-additive sequence $\Psi := \{\psi_n\}_{n\geqslant 1}$ of continuous functions is such that

$$\mathcal{F}(\rho, \Psi) = \lim_{n\to\infty} \frac{1}{n} \int \psi_n \, d\rho \,. \tag{28}$$

Definition 3.3. *We denote by $P_{top}(\sigma, \Phi)$ the pressure of the almost additive family Φ, associated to the family φ_n, where*

$$P_{top}(\sigma, \Phi) = \sup_{\rho\in\mathcal{M}_\sigma(\Omega)} \Big\{ h(\rho) \, + \, \mathcal{F}(\rho, \Phi) \Big\}.$$

A probability measure $\mu = \mu_\Phi \in \mathcal{M}_\sigma(\Omega)$ is called a Gibbs measure for the almost additive family Φ, if it attains the supremum.

The previous topological pressure for non-additive sequences can also be defined, in the spirit of information theory, as the maximal topological complexity of the dynamics with respect to such sequences of observables (cf. [3]). The unique Gibbs measure associated to the family $\Phi = (\varphi_n)_{n\geqslant 1}$, $\varphi_n = \sum_{j=0}^{n-1} \log J_{\theta_0} \circ \sigma^j$, $n \in \mathbb{N}$, is μ_{θ_0}. Moreover, in this case $P_{\rm top}(\sigma, \Phi) = 0$. For the family $\Phi := \{\varphi_n\}$ the claim is under the domain of the classical Thermodynamic Formalism as described before by expression (7). In this case

$$P_{\rm top}(\sigma, \Phi) = \sup_{\mu\in\mathcal{M}_\sigma(\Omega)} \{ \, h(\mu) \, + \, \int \log J_{\theta_0} d\mu \, \} = 0. \tag{29}$$

Remark 3.4. In [19], the author proved that any sequence Ψ of almost additive or asymptotically additive potentials is equivalent to standard additive potentials: there exists a continuous potential φ with the same topological pressure, equilibrium states, variational principle, weak Gibbs measures, level sets (and irregular set) for the Lyapunov exponent and large deviations properties. Yet, it is still unknown wether any sequence of Lipschitz continuous potentials has a Lipschitz continuous additive representative.

3.2.2. *Almost-additive potentials related to entropy.* The next proposition says that Gibbs measures determine in a natural way some sequences of almost additive potentials.

Lemma 3.3. *Given $\theta \in \Theta$, the family $\psi^\theta_{n,1}(y) := \log(\mu_\theta(\, C_n(y)\,)$, $n \in \mathbb{N}$, is almost additive.*

Proof. Recall that all potentials f_θ are normalized, thus each μ_θ satisfies the Gibbs property (3) with $P_\theta = 0$. Thus, for $\theta \in \Theta$ there exists $K_\theta > 0$ such that for all $n \geqslant 1$ and $x \in \Omega$

$$\begin{aligned}\mu_\theta(C_{m+n}(x)) &\leqslant K_\theta^3\ \mu_\theta(C_n(x))\,\mu_\theta(\sigma^n(C_{m+n}(x))\,)\\ &= K_\theta^3\ \mu_\theta(C_n(x))\,\mu_\theta(C_m(\sigma^n(x))\,).\end{aligned}$$

Similarly, $\mu_\theta(C_{m+n}(x)) \geqslant K_\theta^{-3}\,\mu_\theta(C_n(x))\,\mu_\theta(C_m(\sigma^n(x))\,)$ for all $n \geqslant 1$. Therefore, the family $\psi^\theta_{n,1}(y) = \log(\mu_\theta(\, C_n(y)\,)$ satisfies

$$\psi^\theta_{n,1} + \psi^\theta_{m,1} \circ \sigma^n - 3\,\log K_\theta \leqslant \psi^\theta_{(n+m),1} \leqslant \psi^\theta_{n,1} + \psi^\theta_{m,1} \circ \sigma^n + 3\,\log K_\theta$$

for all $m, n \geqslant 1$, hence it is almost-additive. □

Note that the natural family

$$y \to -\log \int_E \frac{\mu_\theta(\, C_n(y)\,)}{\mu_{\theta_0}(\, C_n(y)\,)}\, d\Pi_0(\theta),$$

$n \in \mathbb{N}$, which seems at first useful, may *not* be almost additive as one first evaluate fluctuations on the different ways the measures see cylinders and only afterwards take its logarithm. We consider alternatively the sequence of potentials given below.

Lemma 3.4. *For any fixed $y \in Y$ and any Borel set $E \subset \Theta$, the family*

$$\psi_n(y) = \psi_n^E(y) = -\int \mathbf{1}_E(\theta)\ \log(\mu_\theta(\, C_n(y)\,)\, d\Pi_0(\theta),$$

$n \in \mathbb{N}$, is almost additive. In particular, for each $\theta_0 \in \Theta$ and $E \subset \Theta$, the family $\Psi^E := \{\Psi_n^E\}_n$,

$$y \to \Psi_n^E(y) := -\int_E \log\Big(\frac{\mu_\theta(\, C_n(y)\,)}{\mu_{\theta_0}(\, C_n(y)\,)}\Big)\, d\Pi_0(\theta), \tag{30}$$

is almost additive.

Proof. The first assertion is a direct consequence of the previous lemma and linearity of the integral. For the second one just notice that

$$-\int_E \log\Big(\frac{\mu_\theta(\, C_n(y)\,)}{\mu_{\theta_0}(\, C_n(y)\,)}\Big)\, d\Pi_0(\theta) = \psi_n(y) + \psi^{\theta_0}_{n,1}$$

is the sum of two almost-additive sequences, hence almost additive. □

3.2.3. *Almost-additive potentials related to Lyapunov exponents.* Let $\sigma : \Omega \to \Omega$ be a subshift of finite type and for each $\theta = (\theta_1, \theta_2, \ldots, \theta_p) \in [-\varepsilon, \varepsilon]^2$ consider the locally constant linear cocycle $A_\theta : \Omega \to SL(2, \mathbb{R})$ given by

$$A_\theta\,|_{[i]} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix}, \quad \text{for every } i = 1, 2.$$

To each $n \geqslant 1$ and $(x_1, \ldots, x_n) \in \{1, 2\}^n$ one associates the product matrix

$$A_\theta^{(n)}(x_1, \ldots, x_n) := A_{\theta_{x_n}} \ldots A_{\theta_{x_2}} A_{\theta_{x_1}}.$$

If $\varepsilon > 0$ is chosen small the previous family of matrices preserve a constant cone field in $\mathbb{R}^2$, hence have a dominated splitting. Furthermore, if $\mu \in \mathcal{M}_\sigma(\Omega)$ is ergodic the Oseledets theorem ensures that for μ-almost every $x \in \Omega$ there exists a cocycle invariant splitting $\mathbb{R}^2 = E^+_{\theta,x} \oplus E^-_{\theta,x}$ so that the limit

$$\lambda_{\theta,i} := \lim_{n\to\infty} \frac{1}{n} \log \|A^{(n)}_\theta(x)\, v_i\|$$

exists and is independent of the vector $v_i \in E^i_{\theta,x} \setminus \{0\}$, $(i = \pm)$. Actually, Oseledets theorem also ensures that the largest Lyapunov exponent can be obtained by means of sub-additive sequences, as

$$\lambda^+(A_\theta, \mu) = \lambda_{\theta,+} = \inf_{n\geqslant 1} \frac{1}{n} \log \|A^{(n)}_\theta(x)\|,$$

for μ-almost every x. Since all matrices preserve a cone field then for each $\theta \in [-\varepsilon, \varepsilon]^2$ the sequence $(\log \|A^{(n)}_\theta(x)\|)_{n\geqslant 1}$ is known to be almost-additive on the x-variable (cf. [28]). Most surprisingly, in this simple context the largest annealed Lyapunov exponent

$$\lim_{n\to\infty} \frac{1}{n} \int \log \|A^{(n)}_\theta(x)\|\, d\nu(x)$$

varies analytically with the parameter θ (cf. [50]). We will need the following localization result.

Lemma 3.5. *$\lambda^+(A_{(0,0)}, \nu) > \lambda^+(A_\theta, \nu)$ for every $\theta \in [-\varepsilon, \varepsilon]^2 \setminus \{(0,0)\}$*

Proof. First observe that, as all matrices are obtained by a rotation of the original hyperbolic matrix, we have that $\log \|A_\theta\| = \log(\frac{3+\sqrt{5}}{2})$ for all $\theta \in [-\varepsilon, \varepsilon]^2$. Second, it is clear from the definition that $\lambda^+(A_{(0,0)}, \nu)$ is the logarithm of the largest eigenvalue of the unperturbed hyperbolic matrix, hence it is $\log(\frac{3+\sqrt{5}}{2})$. Finally, Furstenberg [31] proved that

$$\lambda^+(A_\theta, \nu) = \int \int_{\mathbb{S}^1} \log \frac{\|A_\theta(x) \cdot v\|}{\|v\|}\, d\mathbb{P}(v)\, d\nu(x)$$

where $\mathbb{S}^1$ stands for the projective space of $\mathbb{R}^2$ and $\mathbb{P}$ is a ν-stationary measure, meaning that $\nu \times \mathbb{P}$ is invariant by the projectivization of the map $F(x, v) = (\sigma(x), A_{x_0}(v))$ for $(x, v) \in \Omega \times \mathbb{R}^2$. Altogether this guarantees that

$$\lambda^+(A_\theta, \nu) = \log(\frac{3+\sqrt{5}}{2}) \quad \text{if and only if} \quad \mathbb{P} = \delta_{v_+}$$

where v_+ is the leading eigenvector of $A_{(0,0)}$, which cannot occur because $\nu \times \delta_{v_+}$ is not invariant by the projectivized cocycle. This proves the lemma. □

3.3. **Large deviations: speed of convergence.** Large deviations estimates are commonly used in decision theory (see e.g. [12, 30, 56]). In the context of dynamical systems, the exponential rate of convergence in large deviations are defined in terms of rate functions, often described by thermodynamic quantities as pressure and entropy. In the case of level-1 large deviation estimates these can be defined as follows. Given a family $\Psi^E := \{\psi^E_n\}$, where $\psi^E_n : \Omega \to \mathbb{R}$, E is a Borel set of parameters, $n \in \mathbb{N}$ and $-\infty \leqslant c < d \leqslant \infty$, we define

$$\overline{R}_\nu(\Psi^E, [c, d]) = \limsup_{n\to\infty} \frac{1}{n}\ \log\, \nu(\{\, y \in \Omega : \frac{1}{n}\ \psi^E_n(y) \in [c, d]\})$$

and

$$\underline{R}_\nu(\Psi^E,(c,d)) = \liminf_{n\to\infty} \frac{1}{n}\ \log \nu(\{\, y\in\Omega : \frac{1}{n}\ \psi_n^E(y)\in(c,d)\}).$$

Since the subshift dynamics satisfies the transitive specification property (also referred as gluing orbit property, the [57, Theorem B] ensures the following large deviations principle for the subshift and either asymptotically additive or certain sequences of sub-additive potentials.

Theorem 3.6. *Let $\Phi=\{\varphi_n\}$ be an almost additive family of potentials with $P(\sigma,\Phi)>-\infty$ and let ν be a Gibbs measure for σ with respect to Φ. Assume that either:*

(a) *$\Psi=\{\psi_n\}$ is an asymptotically additive family of potentials, or;*
(b) *$\Psi=\{\psi_n\}$ is a sub-additive family of potentials such that:*
   i. *$\Psi=\{\psi_n\}$ satisfies the weak Bowen condition: there exists $\delta>0$ so that*

$$\limsup_{n\to\infty}\frac{\sup\{|\varphi_n(y)-\varphi_n(z)|:\ y,z\in B_n(x,\delta)\}}{n}=0;$$

   ii. *$\inf_{n\geqslant 1}\frac{\psi_n(x)}{n}>-\infty$ for all $x\in\Omega$; and*
   iii. *the sequence $\{\psi_n/n\}$ is equicontinuous.*

Given $c\in\mathbb{R}$, it holds that:

(1) *$\overline{R}_\nu(\Psi,[c,\infty))\leqslant \sup\big\{-P(\sigma,\Phi)+h_\eta(\sigma)+\mathcal{F}(\eta,\Phi)\big\}\}$, where the supremum is over all $\eta\in\mathcal{M}_\sigma(\Omega)$ such that $\mathcal{F}(\eta,\Psi)\geqslant c$;*
(2) *$\underline{R}_\nu(\Psi,(c,\infty))\geqslant \sup\big\{-P(\sigma,\Phi)+h_\eta(\sigma)+\mathcal{F}(\eta,\Phi)\big\}$ where the supremum is taken over all $\eta\in\mathcal{M}_\sigma(\Omega)$ satisfying $\mathcal{F}(\eta,\Psi)>c$.*

While in the previous theorem both invariant measures and sequences of observables may be generated by non-additive sequences of potentials (we refer the reader e.g. to [3] for the construction of equilibrium states associated to almost-additive sequences of potentials) we will be mostly concerned with Gibbs measures generated by a single Lipschitz continuous potential. In the special case of the almost-additive sequences considered in Subsection 3.2.2 the previous theorem can read as follows:

Corollary 3.5. *Let $\Phi=\{\varphi_n\}$ be defined by $\varphi_n=\sum_{j=0}^{n-1}\log J_{\theta_0}$, $n\in\mathbb{N}$ and let μ_{θ_0} denote the corresponding Gibbs measure. For a given Borel set $E\subset\Theta$, take $\Psi^E:=\{\psi_n^E\}$, where ψ_n^E, $n\in\mathbb{N}$ was defined in Lemma 3.4. Then, given $\infty\geqslant d>c\geqslant-\infty$ we have:*

a. $\overline{R}_{\mu_{\theta_0}}(\Psi^E,[c,d])\leqslant \sup\Big\{h(\eta)+\int\log J_{\theta_0}\,d\eta:\eta\in\mathcal{S}(Y)\ \text{so that}\,\mathcal{F}(\eta,\Psi^E)\in[c,d]\Big\}$
b. $\overline{R}_{\mu_{\theta_0}}(\Psi^E,(c,d))\geqslant \sup\Big\{h(\eta)+\int\log J_{\theta_0}\,d\eta:\eta\in\mathcal{S}(Y)\ \text{so that}\,\mathcal{F}(\eta,\Psi^E)\in(c,d)\Big\}$

As the entropy function of the subshift is upper-semicontinuous, any sequence of invariant measures whose free energies associated to a continuous potential tend to the topological pressure accumulate on the space of equilibrium states. Thus, in the special case that there exists a unique equilibrium state, any such sequence is convergent to the equilibrium state. Altogether the previous argument gives the following:

Lemma 3.7. *Consider the sequence of functions $\Phi=\{\varphi_n\}_{n\geqslant1}$ where $\varphi_n(y)=\sum_{j=0}^{n-1}\log J_{\theta_0}(\sigma^j(y))$ and $\log J_{\theta_0}$ is Lipschitz continuous, and let μ_Φ denote the corresponding Gibbs measure. If U is an open neighborhood of the Gibbs measure μ_Φ then there exists $\alpha_1>0$ such that*

$$\sup_{\mu\in\mathcal{M}_\sigma(\Omega)\setminus U}\{\,h(\mu)\,+\,\int\log J_{\theta_0}d\mu\}=\sup_{\mu\in\mathcal{M}_\sigma(\Omega)\setminus U}\{\,h(\mu)\,+\,\mathcal{F}(\mu,\Phi)\,\}\,<\,-\alpha_1.$$

We are particularly interested in the δ-neighborhood of the parameter $\theta_0 \in \Theta$ defined by

$$B_\delta = \{\, \theta \in \Theta \,|\, d_\theta(\theta, \theta_0) < \delta \,\}, \qquad \text{for some } \delta > 0. \tag{31}$$

The next result establishes large deviations estimates for relative entropy associated to Gibbs measures close to μ_{θ_0}. More precisely:

Proposition 3.8. *Let Ψ^E be defined by (30). For any $\delta > 0$ there exists $d_\delta > 0$ satisfying*

$$\mathcal{F}(\mu_{\theta_0}, \Psi^{B_\delta}) \; < d_\delta < \mathcal{F}(\mu_{\theta_0}, \Psi^{\Theta}) = \int_\Theta h(\mu_{\theta_0} \mid \mu_\theta)\, d\Pi_0(\theta).$$

d_δ can be taken small if δ is small.
Moreover, for every small $\delta > 0$ there exists $\alpha_1 > 0$ so that

$$\limsup_{n\to\infty} \frac{1}{n} \; \log \Big[\mu_{\theta_0} \Big(\{\, y \in \Omega : \; -\frac{1}{n} \int_{B_\delta} \log \Big(\frac{\mu_\theta(\, C_n(y)\,)}{\mu_{\theta_0}(\, C_n(y)\,)} \Big) \; d\Pi_0(\theta) \in [d_\delta, \infty) \} \Big) \Big] \leqslant -\alpha_1.$$

Proof. Remember that, given $\eta \in \mathcal{M}_\sigma(\Omega)$ and $E \subset \Theta$,

$$-\lim_{n\to\infty} \frac{1}{n} \int \int_E \log \Big(\frac{\mu_\theta(\, C_n(y)\,)}{\mu_{\theta_0}(\, C_n(y)\,)} \Big)\, d\Pi_0(\theta)\, d\eta(y) = \lim_{n\to\infty} \frac{1}{n} \int \psi_n^E(y)\, d\eta(y) \; = \; \mathcal{F}(\eta, \Psi^E).$$

Taking $\eta = \mu_{\theta_0}$ and $E = \Theta$ we get from (25), (27) and Lemma 3.1 that

$$\begin{aligned} \mathcal{F}(\mu_{\theta_0}, \Psi^{\Theta}) &= \int_\Theta \int -\lim_{n\to\infty} \frac{1}{n} \log \Big(\frac{\mu_\theta(\, C_n(y)\,)}{\mu_{\theta_0}(\, C_n(y)\,)} \Big) \; d\mu_{\theta_0}(y) \; d\Pi_0(\theta) \\ &= \int_\Theta h(\mu_{\theta_0} \mid \mu_\theta)\, d\Pi_0(\theta) \\ &= -h(\mu_{\theta_0}) - \int_\Theta \int \log J_\theta \, d\mu_\theta \, d\Pi_0. \end{aligned} \tag{32}$$

Similarly, one obtains $\mathcal{F}(\mu_{\theta_0}, \Psi^E) = -h(\mu_{\theta_0})\, \Pi_0(E) - \int_E \int \log J_\theta \, d\mu_\theta \, d\Pi_0$ for any $E \subset \Theta$. Using that $h(\mu_{\theta_0} \mid \mu_\theta) > 0$ for all $\theta \neq \theta_0$ and that Π_0 is fully supported on Θ, Lemma 3.1 ensures that

$$\int_{\Theta \setminus B_\delta} h(\mu_{\theta_0} \mid \mu_\theta)\, d\Pi_0(\theta) > 0$$

for every small δ. In consequence,

$$\begin{aligned} \mathcal{F}(\mu_{\theta_0}, \Psi^{B_\delta}) = \int_{B_\delta} h(\mu_{\theta_0}, \mu_\theta)\, d\Pi_0(\theta) \; &< \int_{B_\delta} \int -\lim_{n\to\infty} \frac{1}{n} \log(\frac{\mu_\theta(\, C_n(y)\,)}{\mu_{\theta_0}(\, C_n(y)\,)})\, d\mu_{\theta_0}(y) \; d\Pi_0(\theta) \\ &+ \int_{\Theta \setminus B_\delta} \int -\lim_{n\to\infty} \frac{1}{n} \log(\frac{\mu_\theta(\, C_n(y)\,)}{\mu_{\theta_0}(\, C_n(y)\,)})\, d\mu_{\theta_0}(y) \; d\Pi_0(\theta) \\ &= \int_\Theta h(\mu_{\theta_0} \mid \mu_\theta) d\Pi_0(\theta) = \mathcal{F}(\mu_{\theta_0}, \Psi^{\Theta}). \end{aligned}$$

for every small δ, hence there exists $d_\delta > 0$ so that

$$\mathcal{F}(\mu_{\theta_0}, \Psi^{B_\delta}) < d_\delta < \mathcal{F}(\mu_{\theta_0}, \Psi^{\Theta}). \tag{33}$$

Now, on the one hand, by continuity of $\eta \mapsto \mathcal{F}(\eta, \Psi^{B_\delta})$, the set $U = \{\, \eta \in \mathcal{M}_\sigma(\Omega) \;:\; \mathcal{F}(\eta, \Psi^{B_\delta}) < d_\delta\}$ is an open neighborhood of μ_{θ_0}. On the other hand, according to Lemma 3.7 there exists $\alpha_1 > 0$ such that

$$\sup_{\eta \in \mathcal{M}_\sigma(\Omega)\setminus U} \Big\{ h(\eta) + \int \log J_{\theta_0}\, d\eta \Big\} \leqslant -\alpha_1 < 0.$$

Therefore, from Theorem 3.5

$$\limsup_{n\to\infty} \frac{1}{n}\; \log \mu_{\theta_0}\Big(\{\, y \in \Omega \,|\, -\frac{1}{n}\int_{E^\delta} \log(\frac{\mu_\theta(\,C_n(y)\,)}{\mu_{\theta_0}(\,C_n(y)\,)})\, d\Pi_0(\theta) \;\in [d_\delta, \infty)\}\Big)$$
$$\leqslant \sup_{\{\eta \in \mathcal{M}_\sigma(\Omega)\,:\; \mathcal{F}(\eta, \Psi^{B_\delta}) > d_\delta\}} \{h(\eta) + \int \log J_{\theta_0}\, d\eta\,\} \leqslant -\,\alpha_1 < 0.$$

□

Remark 3.9. From Hypothesis A the value $d_\delta > 0$ can be taken small, if $\delta > 0$ is small, because $\mathcal{F}(\mu_{\theta_0}, \Psi^{B_\delta}) = \int_{B_\delta} h(\mu_{\theta_0} \mid \mu_\theta)\, d\Pi_0(\theta)$.

Corollary 3.10. *Given $\delta > 0$ small let $B_\delta \subset \Theta$ be the δ-open neighborhood of θ_0 defined in (31) and let $d_\delta > 0$ be given by Proposition 3.8. The following holds:*

$$\limsup_{n\to\infty} -\frac{1}{n} \log \int_{B_\delta} \frac{\mu_\theta(\,C_n(y)\,)}{\mu_{\theta_0}(\,C_n(y)\,)}\, d\Pi_0(\theta) \leqslant d_\delta \tag{34}$$

for μ_{θ_0}-almost every point y. Moreover, for μ_{θ_0}-almost every point y

$$\liminf_{n\to\infty} \frac{1}{n} \log \int_{B_\delta} \mu_\theta(\,C_n(y)\,)\, d\Pi_0(\theta) \geqslant -\Pi_0(B_\delta)\cdot h(\mu_{\theta_0}) - d_\delta. \tag{35}$$

Proof. For each $n \geqslant 1$ consider the set

$$A_n = \Big\{\, y \in \Omega \,|\, -\frac{1}{n}\int_{B_\delta} \log \frac{\mu_\theta(\,C_n(y)\,)}{\mu_{\theta_0}(\,C_n(y)\,)}\, d\Pi_0(\theta) \in [d_\delta, \infty)\Big\}.$$

By Proposition 3.8, we get that $\sum_n \mu_{\theta_0}(A_n) < \infty$. It follows from Borel-Cantelli Lemma that for μ_{θ_0}-almost every point $y \in \Omega$ there exists an N, such that $y \notin A_n$ for all $n > N$. Equivalently, $-\frac{1}{n}\int_{B_\delta} \log(\frac{\mu_\theta(\,C_n(y)\,)}{\mu_{\theta_0}(\,C_n(y)\,)})\, d\Pi_0(\theta) < d_\delta$ for all $n > N$, which proves (34). Therefore, from Jensen inequality, we get for μ_{θ_0}-almost every $y \in \Omega$ and every large $n \geqslant 1$

$$\frac{1}{n}\log \int_{B_\delta} \mu_\theta\;(\,C_n(y)\,)\, d\Pi_0(\theta) \;-\frac{1}{n}\int_{B_\delta} \log(\mu_{\theta_0}(\,C_n(y)\,))\, d\Pi_0(\theta)$$
$$\geqslant \frac{1}{n}\int_{B_\delta} \log(\,\mu_\theta(\,C_n(y)\,)\, d\Pi_0(\theta) - \frac{1}{n}\int_{B_\delta} \log(\mu_{\theta_0}(\,C_n(y)\,))\, d\Pi_0(\theta)$$
$$= \frac{1}{n}\int_{B_\delta} \log \frac{\mu_\theta(\,C_n(y)\,)}{\mu_{\theta_0}(\,C_n(y)\,)}\, d\Pi_0(\theta) \geqslant -d_\delta. \tag{36}$$

Moreover, as $\lim_{n\to\infty} -\frac{1}{n}\log(\mu_{\theta_0}(\,C_n(y)\,)) = h(\mu_{\theta_0})$ for μ_{θ_0}-almost every y, it follows from the previous inequalities that

$$\liminf_{n\to\infty} \frac{1}{n} \log \int_{B_\delta} \mu_\theta(C_n(y))\, d\Pi_0(\theta) \;+\Pi_0(B_\delta)\, h(\mu_{\theta_0}) \geqslant -d_\delta$$

for μ_{θ_0} almost every y, which proves (35), as desired. □

Remark 3.11. *The previous corollary ensures that for any $\zeta > 0$ and μ_{θ_0}-a.e. $y \in \Omega$*

$$\int_{B_\delta} \mu_\theta(\, C_n(y)\,)\, d\Pi_0(\theta) \geqslant e^{-[d_\delta + \Pi_0(B_\delta)\; h(\mu_{\theta_0}) + \zeta]\, n} \quad \textit{for every large } n \geqslant 1.$$

Moreover, Remark 3.9 guarantees that $d_\delta > 0$ can be chosen small provided that δ is small. In particular, the absolute continuity assumption on the a priori *measure Π_0 (hypothesis A) implies that $\Pi_0(B_\delta)\, h(\mu_{\theta_0}) + d_\delta$ can be taken arbitrarily small, provided that δ is small.*

Lemma 3.12. *For small $\delta > 0$ and μ_{θ_0}-almost every $y \in \Omega$*

$$\limsup_{n\to\infty} \frac{1}{n} \log \int_{\Theta\setminus B_\delta} \mu_\theta(C_n(y))\, d\Pi_0(\theta) \leqslant \sup_{\theta\in\Theta\setminus B_\delta} \int \log J_\theta d\mu_{\theta_0} < 0. \tag{37}$$

Moreover, $\sup_{\theta\in\Theta\setminus B_\delta} \int \log J_\theta \, d\mu_{\theta_0} \to -h(\mu_{\theta_0})$ *as* $\delta \to 0$.

Proof. Recalling the Gibbs property (3) for μ_θ the continuous dependence of the constants K_θ and compactness of Θ we conclude that there exist uniform constants $c_1, c_2 > 0$ so that

$$c_1 \leqslant \frac{\mu_\theta(C_n(x))}{e^{-nP_\theta} + S_n f_\theta(x)} \leqslant c_2 \qquad \forall \theta \in \Theta, \forall y \in \Omega, \forall n \geqslant 1. \tag{38}$$

Furthermore, as the potentials are assumed to be normalized then $P_\theta = 0$ for every $\theta \in \Theta$. Therefore, there exists $C_1 > 0$ and $C_2 > 0$, such that, for all $y \in \Omega$, $\theta \in \Theta$ and $n \geqslant 1$

$$C_1 < \frac{\mu_\theta(C_n(y))}{\mu_{\theta_0}(C_n(y))} \frac{e^{\sum_{j=0}^{n-1} \log J_{\theta_0}(\sigma^j(y))}}{e^{\sum_{j=0}^{n-1} \log J_\theta(\sigma^j(y))}} < C_2.$$

Then,

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log C_1 &< \limsup_{n\to\infty} \frac{1}{n} \log \frac{\mu_\theta(C_n(y))}{\mu_{\theta_0}(C_n(y))} \\
&\quad + \limsup_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \log J_{\theta_0}(\sigma^j(y)) - \sum_{j=0}^{n-1} \log J_\theta(\sigma^j(y))] \\
&< \limsup_{n\to\infty} \frac{1}{n} \log C_2.
\end{aligned}$$

The above expression is independent of y.

In consequence, using the ergodic theorem and that $h(\mu_{\theta_0}) = \int -\log J_{\theta_0}\, d\mu_{\theta_0}$, one gets

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{\mu_\theta(C_n(y))}{\mu_{\theta_0}(C_n(y))} \leqslant h(\mu_{\theta_0}) + \int \log J_\theta d\mu_{\theta_0}, \quad \text{for } \mu_{\theta_0}\text{-a.e. } y.$$

Fix $\zeta > 0$ arbitrary and small. The previous expression ensures that, for μ_{θ_0}-a.e. $y \in \Omega$, there exists a $N = N(\zeta, y)$ such that

$$\frac{\mu_\theta(C_n(y))}{\mu_{\theta_0}(C_n(y))} \leqslant e^{n\,(h(\mu_{\theta_0}) + \int \log J_\theta d\mu_{\theta_0} + \zeta)} \quad \text{for every large } n \geqslant N = N(\zeta, y)$$

Given a small $\delta > 0$, by uniqueness of the equilibrium state for $\log J_\theta$, we have that

$$\rho_\delta := h(\mu_{\theta_0}) + \sup_{\theta\in\Theta\setminus B_\delta} \int \log J_\theta d\mu_{\theta_0} = \sup_{\theta\in\Theta\setminus B_\delta} [h(\mu_{\theta_0}) + \int \log J_\theta d\mu_{\theta_0}] < 0,$$

and that ρ_δ tends to zero as $\delta \to 0$. Then, for μ_{θ_0}-almost every point y, and $n \geq N(\zeta, y)$,

$$\int_{\Theta\setminus B_\delta} \frac{\mu_\theta(C_n(y))}{\mu_{\theta_0}(C_n(y))} d\Pi_0(\theta) \leqslant \int_{\Theta\setminus B_\delta} e^{n\,(h(\mu_{\theta_0})+\log J_\theta d\mu_{\theta_0}+\zeta)} d\Pi_0(\theta) \leqslant \Pi_0(\Theta \setminus B_\delta) e^{n(\rho_\delta+\zeta)},$$

which implies for small arbitrary ζ

$$\limsup_{n\to\infty} \frac{1}{n} \log \int_{\Theta\setminus B_\delta} \frac{\mu_\theta(C_n(y))}{\mu_{\theta_0}(C_n(y))} d\Pi_0(\theta) \leqslant \rho_\delta + \zeta.$$

$\zeta > 0$ small enough we conclude that, for μ_{θ_0}-almost every point y,

$$\limsup_{n\to\infty} \frac{1}{n} \log \int_{\Theta\setminus B_\delta} \mu_\theta(C_n(y)) d\Pi_0(\theta) \leqslant -h(\mu_{\theta_0}) + \rho_\delta + \zeta = \sup_{\theta\in\Theta\setminus B_\delta} \int \log J_\theta d\mu_{\theta_0} + \zeta < 0.$$

For fixed y, as ζ is arbitrary we get the (37).

□

Proposition 3.13. *For μ_{θ_0}-almost every $y \in \Omega$*

$$0 \leqslant -\limsup_{n\to\infty} \frac{1}{n} \log Z_n(y) \;\leqslant -\int_\Theta \int_\Omega \log J_\theta(y) d\mu_{\theta_0}(y) d\Pi_0(\theta) - h(\mu_{\theta_0}).$$

Proof. For y, μ_{θ_0}-almost every everywhere, if

$$0 < \limsup_{n\to\infty} \frac{1}{n} \log Z_n(y) = \limsup_{n\to\infty} \frac{1}{n} \log \int_\Theta \frac{\mu_\theta\,(\,C_n(y)\,)}{\mu_{\theta_0}\,(\,C_n(y)\,)} d\Pi_0(\theta),$$

taking $\delta \to 0$ in (37), we would reach a contradiction.

□

The statement of the second inequality in the above Proposition is nothing more than the expression (10).

4. Proof of the main results

4.1. **Proof of Theorem A.** We proceed to show that the *a posteriori* measures in Theorem A do converge, for μ_{θ_0}-typical points y. In order to prove that $\Pi_n(\cdot, y) \to \delta_{\theta_0}$ (in the weak* topology) it is sufficient to prove that, for every $\delta > 0$, one has that $\Pi_n(\Theta \setminus B_\delta, y) \to 0$ as $n \to \infty$. This is the content of the following theorem.

Theorem 4.1. *Let $\Pi_n(\cdot \mid y)$ be the a posteriori measures defined by (11) and let B_δ be the δ-neighborhood of θ_0 defined by (31). Then, for every small $\delta > 0$ and μ_{θ_0}-a.e. y,*

$$\Pi_n(B_\delta \mid y) = \frac{\int_{B_\delta} \mu_\theta\,(\,C_n(y)\,)\, d\Pi_0(\theta)}{\int_\Theta \mu_\theta(C_n(y)\,) d\Pi_0(\theta)} \to 1 \tag{39}$$

exponentially fast as $n \to \infty$.

Proof. Fix $\delta > 0$ small. We claim that $\Pi_n(\Theta \setminus B_\delta \mid y)$ tends to zero exponentially fast as $n \to \infty$. We have to estimate

$$\limsup_{n\to\infty} \frac{1}{n} \log \int_{B_\delta} \mu_\theta(C_n(y)) d\Pi_0(\theta)$$

and

$$-\limsup_{n\to\infty} \frac{1}{n} \log \int_{\Theta\setminus B_\theta} \mu_\theta(C_n(y)) d\Pi_0(\theta).$$

From (35), for μ_{θ_0} almost every point y

$$\limsup_{n\to\infty} \frac{1}{n} \log \int_{B_\delta} \mu_\theta(C_n(y))\, d\Pi_0(\theta) \geqslant -h(\mu_{\theta_0})\, \Pi_0(B_\delta) - d_\delta, \tag{40}$$

where d_δ can be taken small if $\delta > 0$ is small. Fix $0 < \zeta < \frac{h(\mu_{\theta_0})}{2}$. Therefore, from Remark 3.11 we get that for μ_{θ_0} almost every point y

$$\int_{B_\delta} \mu_\theta(\, C_n(y)\,)\, d\Pi_0(\theta) \geqslant e^{-[d_\delta + \Pi_0(B_\delta)\; h(\mu_{\theta_0}) + \zeta]\, n} \quad \text{for every large } n \geqslant 1. \tag{41}$$

Observe that the map $\delta \mapsto \sup_{\theta\in\Theta\setminus B_\delta} \int \log J_\theta d\mu_{\theta_0}$ is monotone increasing and recall that $\sup_{\theta\in\Theta\setminus B_\delta} \int \log J_\theta d\mu_{\theta_0} \to -h(\mu_{\theta_0})$ as $\delta \to 0$. On the other hand, $-h(\mu_{\theta_0})\, \Pi_0(B_\delta) - d_\delta$ tends to zero as $\delta \to 0$ (cf. Remark 3.11). Thus,

$$\sup_{\theta\in\Theta\setminus B_\delta} \int \log J_\theta d\mu_{\theta_0} < -h(\mu_{\theta_0})\, \Pi_0(B_\delta) - d_\delta - \zeta \tag{42}$$

for every small $\delta > 0$. As

$$\frac{\int_{B_\delta} \mu_\theta\,(\, C_n(y))\, d\Pi_0(\theta)}{\int_\Theta \mu_\theta(C_n(y)\,) d\Pi_0(\theta)} + \frac{\int_{\Theta\setminus B_\delta} \mu_\theta\,(\, C_n(y)\, d\Pi_0(\theta)}{\int_\Theta \mu_\theta(C_n(y)\, d\Pi_0(\theta)} = 1$$

we just have to show that

$$\frac{\int_{B_\delta} \mu_\theta\,(\, C_n(y))\, d\Pi_0(\theta)}{\int_{\Theta\setminus B_\delta} \mu_\theta(C_n(y)\,) d\Pi_0(\theta)} \to \infty,$$

when $n \to \infty$.

Indeed,

$$\frac{\int_{\Theta\setminus B_\delta} \mu_\theta\,(\, C_n(y)\, d\Pi_0(\theta)}{\int_\Theta \mu_\theta(C_n(y))\, d\Pi_0(\theta)} = \frac{1}{1 + \frac{\int_{B_\delta} \mu_\theta(C_n(y))\, d\Pi_0(\theta)}{\int_{\Theta\setminus B_\delta} \mu_\theta\,(\, C_n(y))\, d\Pi_0(\theta)}}.$$

Now, equations (37) and (41) and the choice of δ in (42) ensure that, for μ_{θ_0}-almost every $y \in \Omega$,

$$\frac{\int_{B_\delta} \mu_\theta\,(\, C_n(y))\, d\Pi_0(\theta)}{\int_{\Theta\setminus B_\delta} \mu_\theta(C_n(y)\,) d\Pi_0(\theta)} \geqslant \frac{e^{-[h(\mu_{\theta_0})\, \Pi_0(B_\delta) + d_\delta + \zeta]\; n}}{e^{n\, \sup_{\theta\in\Theta\setminus B_\delta} \int \log J_\theta d\mu_{\theta_0}}}$$

which tends to infinity as $n \to \infty$. Finally the previous expression also ensures that

$$|\Pi_n(B_\delta \mid y) - 1| = \frac{\int_{\Theta\setminus B_\delta} \mu_\theta\,(\, C_n(y))\, d\Pi_0(\theta)}{\int_\Theta \mu_\theta(C_n(y)\,) d\Pi_0(\theta)} \leqslant e^{n\,[\sup_{\theta\in\Theta\setminus B_\delta} \int \log J_\theta d\mu_{\theta_0} + h(\mu_{\theta_0})\, \Pi_0(B_\delta) + d_\delta + \zeta]}$$

decreases exponentially fast with exponential rate that can be taken uniform for all small $\delta > 0$. This finishes the proof of the theorem. □

4.2. **Proof of Theorem B.** By assumption, there exists a full ν-measure subset $Y' \subset Y$ so that the limit

$$\Gamma^y(\theta) := \lim_{n\to\infty} \frac{1}{n} \log \int_\Omega e^{\varphi_n(\theta,x,y)}\, d\mu_\theta(x)$$

exists for every $y \in Y'$. Given an arbitrary $y \in Y'$ we proceed to estimate the asymptotic behavior of the *a posteriori* measures $\Pi_n(\cdot \mid y)$ given by (23).

Given $\delta > 0$, by upper semicontinuity of $\Gamma^y(\cdot)$ the function Γ^y has a maximum value and there exists $d_\delta > 0$ (which may be chosen to converge to zero as $\delta \to 0$) so that

$$B_\delta^y = \big\{\theta \in \Theta\colon d_\Theta\big(\theta, \operatorname{argmax} \Gamma^y\big) > \delta\big\} \subset (\Gamma^y)^{-1}((-\infty, \alpha^y - d_\delta))$$

is non-empty and open subset, where $\alpha^y := \max_{\theta\in\Theta} \Gamma^y(\theta)$.

There are two cases to consider. On the one hand, if $\Gamma^y(\cdot) \equiv \alpha^y$ is constant then $B_\delta^y = \Theta$ and we conclude that $\Pi_n(B_\delta^y \mid y) = 1$ for all $n \geqslant 1$ and the convergence in (16) is trivially satisfied. On the other hand, as Π_0 is fully supported and absolutely continuous then $\int_\Theta \Gamma^y(\theta)\, d\Pi_0(\theta) < \alpha^y$. Actually, this allows to estimate the double integral

$$\int_{\Theta\setminus B_\delta^y} \int_\Omega e^{\varphi_n(\theta,x,y)}\, d\mu_\theta(x)\, d\Pi_0(\theta)$$

without making use of the features of the set B_δ^y. More precisely, using Jensen inequality and taking the limsup under the sign of the integral,

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log \int_{\Theta\setminus B_\delta^y} \int_\Omega e^{\varphi_n(\theta,x,y)}\, d\mu_\theta(x)\, d\Pi_0(\theta) &\leqslant \int_{\Theta\setminus B_\delta^y} \limsup_{n\to\infty} \frac{1}{n} \log \int_\Omega e^{\varphi_n(\theta,x,y)}\, d\mu_\theta(x)\, d\Pi_0(\theta)\\ &= \int_{\Theta\setminus B_\delta^y} \Gamma^y(\theta)\, d\Pi_0(\theta).\end{aligned}$$

As φ_n are assumed non-negative we conclude that $\Gamma^y(\cdot)$ is a non-negative function and

$$\limsup_{n\to\infty} \frac{1}{n} \log \int_\Theta \int_\Omega e^{\varphi_n(\theta,x,y)}\, d\mu_\theta(x)\, d\Pi_0(\theta) \leqslant \int_\Theta \Gamma^y(\theta)\, d\Pi_0(\theta) < \alpha^y. \tag{43}$$

In consequence, if $0 < \zeta < \frac{1}{2}\big[\alpha^y - \int_\Theta \Gamma^y(\theta)\, d\Pi_0(\theta)\big]$ then

$$\int_\Theta \int_\Omega e^{\varphi_n(\theta,x,y)}\, d\mu_\theta(x)\, d\Pi_0(\theta) \leqslant e^{(\alpha^y-\zeta)n}$$

for every large $n \geqslant 1$. Now, in order to estimate the measures $\Pi_n(\cdot \mid y)$ on the nested family $(B_\delta^y)_{\delta>0}$ we observe that $\int_\Omega e^{\varphi_n(\theta,x,y)}\, d\mu_\theta(x) \geqslant e^{(\alpha^y-d_\delta)n}$, $\forall \theta \in B_\delta^y$, thus

$$\int_{B_\delta^y} \int_\Omega e^{\varphi_n(\theta,x,y)}\, d\mu_\theta(x)\, d\Pi_0(\theta) \geqslant e^{(\alpha^y-d_\delta)n} \Pi_0(B_\delta^y)$$

for every large $n \geqslant 1$. In particular, if $\delta > 0$ is small so that $0 < d_\delta < \zeta$, putting together the last expression, inequality (43) and the fact that $0 < \Pi_0(B_\delta^y) < 1$, one concludes that

$$\begin{aligned}\Pi_n(\Theta \setminus B_\delta^y \mid y) &= \frac{\int_{\Theta\setminus B_\delta^y}\int_\Omega e^{\varphi_n(\theta,x,y)}\,d\mu_\theta(x)\,d\Pi_0(\theta)}{\int_\Theta\int_\Omega e^{\varphi_n(\theta,x,y)}\,d\mu_\theta(x)\,d\Pi_0(\theta)}\\ &\leqslant \frac{\int_{\Theta\setminus B_\delta^y}\int_\Omega e^{\varphi_n(\theta,x,y)}\,d\mu_\theta(x)\,d\Pi_0(\theta)}{\int_{B_\delta^y}\int_\Omega e^{\varphi_n(\theta,x,y)}\,d\mu_\theta(x)\,d\Pi_0(\theta)}\\ &\leqslant \frac{1}{\Pi_0(B_\delta^y)}e^{-(\zeta-d_\delta)n}\end{aligned}$$

tends exponentially fast to zero, as claimed. Hence, any accumulation point of $(\Pi_n(\cdot \mid y))_{n\geqslant 1}$ (in the weak* topology) is supported on the compact set $\operatorname{argmax}\Gamma^y$, which proves the first statement in the theorem. As the second assertion is immediate from the first one, this concludes the proof of the theorem. □

4.3. **Proof of Theorem C.** Consider the family of loss functions $\ell_n : \Theta \times X \times Y \to \mathbb{R}$ defined by (17) associated to an almost additive sequence $\Phi = (\varphi_n)_{n\geqslant 1}$ of continuous and non-negative observables $\varphi_n : \Theta \times X \times Y \to \mathbb{R}_+$ satisfying assumptions (H1)-(H2).

(H1) for each $\theta \in \Theta$ and $x \in X$ there exists a constant $K_{\theta,x} > 0$ so that, for every $y \in Y$,

$$\varphi_n(\theta,x,y) + \varphi_m(\theta,x,T^n(y)) - K_{\theta,x} \leqslant \varphi_{m+n}(\theta,x,y) \leqslant \varphi_n(\theta,x,y) + \varphi_m(\theta,x,T^n(y)) + K_{\theta,x}$$

(H2) $\int K_{\theta,x}d\mu_\theta(x) < \infty$ for every $\theta \in \Theta$.

The *a posteriori* measures are

$$\Pi_n(E \mid y) = \frac{\int_E \psi_n(\theta,y)\,d\Pi_0(\theta)}{\int_\Theta \psi_n(\theta,y)\,d\Pi_0(\theta)}, \tag{44}$$

where the sequence $\psi_n(\theta,y) = \int_\Omega \varphi_n(\theta,x,y)\,d\mu_\theta(x)$ is almost additive in the y-variable. Indeed, this family satisfies

$$\psi_n(\theta,y)+\psi_m(\theta,T^n(y)) - \int K_{\theta,x}d\mu_\theta(x) \leqslant \psi_{m+n}(\theta,y) \leqslant \psi_n(\theta,y)+\psi_m(\theta,T^n(y)) + \int K_{\theta,x}d\mu_\theta(x)$$

for every $m, n \geqslant 1$, every $\theta \in \Omega$ and $y \in Y$. Now, for each fixed $\theta \in \Theta$, we note that the sequence of observables

$$\Big(\psi_n(\theta,\cdot) + \int K_{\theta,x}d\mu_\theta(x)\Big)_{n\geqslant 1}$$

is subadditive. Hence, Kingman's subadditive ergodic theorem ensures that the limit $\lim_{n\to\infty}\frac{\psi_n(\theta,y)}{n}$ does exist and is ν-almost everywhere constant to the non-negative function $\psi_*(\theta) := \inf_{n\geqslant 1}\frac{1}{n}\int \psi_n(\theta,y)\,d\nu(y)$. The function ψ_* is measurable and integrable, because it satisfies $0 \leqslant \psi_* \leqslant \psi_1$. Thus, taking the limit under the sign of the integral and noticing that the denominator is a normalizing term we conclude that

$$\lim_{n\to\infty}\Pi_n(E \mid y) = \frac{\int_E \psi_*(\theta)\,d\Pi_0(\theta)}{\int_\Theta \psi_*(\theta)\,d\Pi_0(\theta)} = \frac{\int 1_E\psi_*(\theta)\,d\Pi_0(\theta)}{\int_\Theta \psi_*(\theta)\,d\Pi_0(\theta)} \tag{45}$$

for every measurable subset $E \subset \Theta$. This proves the first statement of the theorem.

We proceed to prove the level-1 large deviations estimates on the convergence of the *a posteriori* measures $\Pi_n(\cdot \mid y)$ to Π_*, whenever T is a subshift of finite type and ν is a Gibbs measure

associated to a Lipschitz continuous potential φ. We will make use of the following instrumental lemma, whose proof is left as a simple exercise to the reader.

Lemma 4.2. *Given arbitrary functions $A, B : \Omega \to \mathbb{R}_+$ and constants $a, b, \delta > 0$ and $0 < \xi < b$, the following holds:*

$$\Big\{ \Big| \frac{A(y)}{B(y)} - \frac{a}{b} \Big| > \delta \Big\} \subset S_1 \ \cup \ S_2 \ \cup \ S_3$$

where $S_1 = \Big\{ \Big| B(y) - b \Big| > \xi \Big\}$, $\quad S_2 = \Big\{ \frac{1}{b-\xi} \Big| A(y) - a \Big| > \frac{\delta}{2} \Big\}$ *and* $S_3 = \Big\{ \frac{a}{b(b-\xi)} \Big| B(y) - b \Big| > \frac{\delta}{2} \Big\}$.

Let us return to the proof of the large deviation estimates. Given $g \in C(\Theta, \mathbb{R})$ it is not hard to check using (44) and (45) that

$$\int g \, d\Pi_n(\cdot \mid y) = \frac{\int g(\theta) \frac{\psi_n(\theta, y)}{n} \, d\Pi_0(\theta)}{\int_\Theta \frac{\psi_n(\theta, y)}{n} \, d\Pi_0(\theta)} \quad \text{and} \quad \int g \, d\Pi_* = \frac{\int g(\theta) \psi_*(\theta) \, d\Pi_0(\theta)}{\int_\Theta \psi_*(\theta) \, d\Pi_0(\theta)}. \tag{46}$$

Fix $\delta > 0$. In order to provide an upper bound for

$$\limsup_{n \to \infty} \frac{1}{n} \ \log \, \nu(\{ y \in \Omega : \Big| \int g \, d\Pi_n(\cdot \mid y) - \int g \, d\Pi_* \Big| > \delta \})$$

we will estimate the set $\Big\{ \Big| \int g \, d\Pi_n(\cdot \mid y) - \int g \, d\Pi_* \Big| > \delta \Big\}$ as in Lemma 4.2. For that purpose, fix $0 < \xi < \min_{\theta \in \Theta} \psi_*(\theta)$. For each fixed $\theta \in \Theta$ the family $\Psi^\theta := (\psi_n(\theta, \cdot))_n$ is almost-additive. Hence Theorem 3.6 implies that

$$\limsup_{n \to \infty} \frac{1}{n} \ \log \, \nu(\{ y \in \Omega : \Big| \frac{\psi_n(\theta, y)}{n} - \psi_*(\theta) \Big| \geqslant \xi \}) \leqslant \sup_{\mathcal{P}^1_{\theta, \xi, \delta}} \Big\{ - P(\sigma, \varphi) + h_\eta(\sigma) + \int \varphi \, d\eta \Big\}$$

where $\mathcal{P}^1_{\theta, \xi, \delta} \subset \mathcal{M}_\sigma(\Omega)$ is the space of invariant probability measures η such that $|\mathcal{F}(\eta, \Psi^\theta) - \psi_*(\theta)| \geqslant \xi$. In consequence,

$$\begin{aligned}
\limsup_{n \to \infty} & \frac{1}{n} \log \, \nu\Big(\Big\{ y \in \Omega : \Big| \int_\Theta \frac{\psi_n(\theta, y)}{n} \, d\Pi_0(\theta) - \int_\Theta \psi_*(\theta) \, d\Pi_0(\theta) \Big| \geqslant \xi \Big\} \Big) \\
& \leqslant \limsup_{n \to \infty} \frac{1}{n} \ \log \, \nu\Big(\Big\{ y \in \Omega : \int_\Theta \Big| \frac{\psi_n(\theta, y)}{n} - \psi_*(\theta) \Big| \, d\Pi_0(\theta) \geqslant \xi \Big\} \Big) \\
& \leqslant \limsup_{n \to \infty} \frac{1}{n} \ \log \, \nu\Big(\Big\{ y \in \Omega : \Big| \frac{\psi_n(\theta, y)}{n} - \psi_*(\theta) \Big| \ \geqslant \xi, \text{ for some } \theta \in \Theta \Big\} \Big) \\
& \leqslant \sup_{\theta \in \Theta} \sup_{\mathcal{P}^1_{\theta, \xi, \delta}} \Big\{ - P(\sigma, \varphi) + h_\eta(\sigma) + \int \varphi \, d\eta \Big\}.
\end{aligned} \tag{47}$$

Analogously,

$$\begin{aligned}
\limsup_{n \to \infty} & \frac{1}{n} \log \, \nu\Big(\Big\{ y \in \Omega : \frac{1}{\int_\Theta \psi_*(\theta) \, d\Pi_0(\theta) - \xi} \Big| \int g(\theta) \frac{\psi_n(\theta, y)}{n} \, d\Pi_0(\theta) - \int_\Theta g(\theta) \psi_*(\theta) \, d\Pi_0(\theta) \Big| \geqslant \frac{\delta}{2} \Big\} \Big) \\
& \leqslant \limsup_{n \to \infty} \frac{1}{n} \log \, \nu\Big(\Big\{ y \in \Omega : \Big| \int \frac{\psi_n(\theta, y)}{n} \, d\Pi_0(\theta) - \int_\Theta \psi_*(\theta) \, d\Pi_0(\theta) \Big| \geqslant \frac{\int_\Theta \psi_*(\theta) \, d\Pi_0(\theta) - \xi}{2 \|g\|_\infty} \delta \Big\} \Big) \\
& \leqslant \sup_{\theta \in \Theta} \sup_{\mathcal{P}^2_{\theta, \xi, \delta}} \Big\{ - P(\sigma, \varphi) + h_\eta(\sigma) + \int \varphi \, d\eta \Big\},
\end{aligned} \tag{48}$$

where $\eta \in \mathcal{P}^2_{\theta,\xi,\delta} \subset \mathcal{M}_\sigma(\Omega)$ if and only if $|\mathcal{F}(\eta,\Psi^\theta) - \psi_*(\theta)| \geqslant \frac{\int_\Theta \psi_*(\theta)\, d\Pi_0(\theta) - \xi}{2\|g\|_\infty}\delta$. The third term in the decomposition of Lemma 4.2 is identical to the estimate of (47) and we have

$$\limsup_{n\to\infty} \frac{1}{n} \log \nu\Big(\Big\{ y \in \Omega : \Big| \int_\Theta \frac{\psi_n(\theta, y)}{n}\, d\Pi_0(\theta) - \int_\Theta \psi_*(\theta)\, d\Pi_0(\theta)\Big| \geqslant \frac{\big(\int_\Theta \psi_*(\theta)\, d\Pi_0(\theta) - \xi\big)^2}{2\int_\Theta \psi_*(\theta)\, d\Pi_0(\theta)}\delta\Big\}\Big)$$
$$\leqslant \sup_{\theta\in\Theta} \sup_{\mathcal{P}^3_{\theta,\xi,\delta}} \Big\{ -P(\sigma,\varphi) + h_\eta(\sigma) + \int \varphi\, d\eta\Big\}, \tag{49}$$

where $\eta \in \mathcal{P}^3_{\theta,\xi,\delta} \subset \mathcal{M}_\sigma(\Omega)$ if and only if $|\mathcal{F}(\eta,\Psi^\theta) - \psi_*(\theta)| \geqslant \frac{\big(\int_\Theta \psi_*(\theta)\, d\Pi_0(\theta) - \xi\big)^2}{2\int_\Theta \psi_*(\theta)\, d\Pi_0(\theta)}\delta$. Altogether, if $0 < \delta < 1$ and $\xi = \delta \cdot \min\{\inf_{\theta\in\Theta} \psi_*(\theta), \int_\Theta \psi_*(\theta)\, d\Pi_0(\theta)\} > 0$, estimates (47)-(49) imply that there exists $c > 0$ so that

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n}\ \log \nu\Big(\Big\{ y \in \Omega : \Big| \int g\, d\Pi_n(\cdot \mid y) - \int g\, d\Pi_*\Big| \geqslant \delta\Big\}\Big)
&\leqslant \sup_{\theta\in\Theta} \max_{1\leqslant i\leqslant 3} \sup_{\mathcal{P}^i_{\theta,\xi,\delta}} \Big\{ -P(\sigma,\varphi) + h_\eta(\sigma) + \int \varphi\, d\eta\Big\} \\
&\leq \sup_{\theta\in\Theta} \sup_{\{\eta:\, |\mathcal{F}(\eta,\Psi^\theta) - \psi_*(\theta)| \geqslant c\delta\}} \Big\{ -P(\sigma,\varphi) + h_\eta(\sigma) + \int \varphi\, d\eta\Big\}
\end{aligned}$$

Finally, it remains to guarantee that the right hand-side above is strictly negative. Notice that as $\mathcal{F}(\nu,\Psi^\theta) = \psi_*(\theta)$, the uniqueness of the equilibrium state (which is an invariant Gibbs measure) for the potential φ and the continuity of the map $\eta \mapsto \mathcal{F}(\eta,\Psi^\theta)$ imply that the set $\mathcal{B}_\theta(\delta) := \{\eta \in \mathcal{M}_\sigma(\Omega) \colon |\mathcal{F}(\eta,\Psi^\theta) - \psi_*(\theta)| \geqslant c\delta\}$ is compact and disjoint from $\{\nu\}$, hence $d_{\mathcal{M}_\sigma(\Omega)}\big(\nu, \mathcal{B}_\theta(\delta)\big) > 0$, for each $\theta \in \Theta$. Hence, under the additional assumption that both maps $\theta \mapsto \mathcal{F}(\eta,\Psi^\theta) = \inf_{n\geq 1} \frac{1}{n}\int \psi_n(\theta,\cdot) d\eta$ and $\theta \mapsto \psi_*(\theta) = \mathcal{F}(\nu,\Psi^\theta)$ are continuous we conclude that

$$\min_{\theta\in\Theta} d_{\mathcal{M}_\sigma(\Omega)}\big(\nu, \mathcal{B}_\theta(\delta)\big) > 0$$

and, consequently,

$$\sup_{\theta\in\Theta} \sup_{\{\eta:\, |\mathcal{F}(\eta,\Psi^\theta) - \psi_*(\theta)| \geqslant c\delta\}} \Big\{ -P(\sigma,\varphi) + h_\eta(\sigma) + \int \varphi\, d\eta\Big\} < 0,$$

which finishes the proof of the theorem □

Acknowledgments. The authors are indebted to the anonymous referees for the careful reading of the manuscript and many suggestions that helped to improve the presentation of the paper. AOL and SRCL was partially supported by CNPq grant. PV was partially supported by CMUP (UID/MAT/00144/2019), which is funded by FCT with national (MCTES) and European structural funds through the programs FEDER, under the partnership agreement PT2020, and by Fundação para a Ciência e Tecnologia (FCT) - Portugal through the grant CEECIND/03721/2017 of the Stimulus of Scientific Employment, Individual Support 2017 Call.

References

[1] F. Abramovich and Y. Ritov, Statistical Theory, A Concise Introduction, CRC Press (2013) 1.1

[2] B. Altaner, Nonequilibrium thermodynamics and information theory: basic concepts and relaxing dynamics, Jour. of Phys. A: Math. and Theo. 50 (2017) 454001 1.1

[3] L. Barreira. Nonadditive thermodynamic formalism: equilibrium and Gibbs measures. Disc. Contin. Dyn. Syst., 16 (2006) 279–305. 3.2.1, 3.3

[4] L. Barreira, Thermodynamic formalism and applications to dimension theory, Birkhäuser, 2011. 1.1

[5] L. Barreira, Y. Cao and J. Wan, Multifractal analysis of asymptotically additive sequences, J. Stat. Phys. 153 (2013) 888–910. 1.1

[6] T.Benoist, M.Fraas, Y.Pautrat, and C.Pellegrini. Invariant measure for quantum trajectories. Prob. Theory and Related Fields, 174, no. 1–2, 307–334 (2019) 1.1

[7] P. Bissiri, C. Holmes and S. Walker. A general framework for updating belief distributions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78(5):1103–1130, 2016 (document), 1.1, 1.2

[8] A. Avila and J. Bochi, A formula with some applications to the theory of Lyapunov exponents. Israel J. Math. 131 (2002), 125–137. 1.1, 2.6

[9] R. Bowen, Equilibrium states and the ergodic theory of Anosov diffeomorphisms, Springer Lecture Notes in Math., vol. 470, 1975. 2.3

[10] J. E. Brasil, J. Knorst and A. O. Lopes, Lyapunov exponents for Quantum Channels: an entropy formula and generic properties, Journal of Dynamical Systems and Geometric Theories, Vol 19(2) 155-187 (2021)

[11] E. Cameron and A. Pettitt, Recursive pathways to marginal likelihood estimation with prior-sensitivity analysis. Statist. Sci. 29 (2014), no. 3, 397-419 1.1

[12] J.A. Buckle, Large Deviation Techniques in Decision, Simulation and Estimation. New York: Wiley, 1990. 1.1

[13] A. Caticha, Entropic Inference and the Foundations of Physics, Lecture Notes, Department of Physics, State University of New York http://dl.icdst.org/pdfs/files1/77964f05542451c01e8e420e975dd664.pdf 3.3 1.1, 1.1

[14] B. Cessac, B, H. Rostro, H, J. C. Vasquez and T. Viéville, How Gibbs distributions may naturally arise from synaptic adaptation mechanisms. A model-based argumentation. J. Stat. Phys. 136 (2009), no. 3, 565—602. 1.3

[15] P. Collet, A. Galves and A. O. Lopes, Maximum Likelihood and Minimum Entropy Estimation of Grammars, Random and Computational Dynamics, 3, pp-241–256 (1995) 1.3

[16] J-R. Chazottes, R. Floriani and R. Lima, Relative entropy and identification of Gibbs measures in dynamical systems, J. Statist. Phys. 90 (1998) no. 3–4, 697–725. 1.2, 1.3, 3.1, 3.1, 3.1

[17] J-R. Chazottes and E. Olivier, Relative entropy, dimensions and large deviations for g -measures, J. Phys. A: Math. Gen. 33 675 (2000) 1.3

[18] J-R Chazottes and D. Gabrielli, Large deviations for empirical entropies of g-measures, Nonlinearity 18 (2005) 2545-2563 1.3

[19] N. Cuneo, Additive, almost additive and asymptotically additive potential sequences are equivalent. Comm. Math. Phys. 377 (2020) 2579–2595. 3.4

[20] I. Daubechies and J. C. Lagarias. Two-scale difference equations, local regularity, infinite products of matrices and fractals. SIAM J. Math. Anal., 23:4 (1992) 1031–1079. 1.1

[21] M. H. DeGroot and M. J. Schervish, Probability and Statistics, Addison Wesley Pearson; 4th edition (January 6, 2011) 1.1

[22] M. Denker and W. Woyczynski, Introductory Statistics and Random Phenomena: Uncertainty, Complexity and Chaotic Behavior in Engineering and Science. New York: Birkhäuser, 2012. 1.1

[23] R. Douc, J. Olssonb, F. Roueff, Posterior consistency for partially observed Markov models, Stochastic Process. Appl. 130 (2020), no. 2, 733—759. 2.1

[24] R. Douc, E. Moulines, J. Olsson and R. van Handel, Consistency of the maximum likelihood estimator for general hidden Markov models. Ann. Statist. 39 (2011), no. 1, 474–513. 2.1

[25] R. Douc, F. Roueff and T. Sim, The maximizing set of the asymptotic normalized log-likelihood for partially observed Markov chains, Technical Report, Institut Mines-Telecom, 2015 2.1

[26] A.C.D. van Enter, A. O. Lopes, S. R. C. Lopes and J. K. Mengue, How to get the Bayesian *a posteriori* probability from an *a priori* probability via thermodynamic formalism for plans; the connection to Disordered Systems, Preprint 2020 1.1

[27] K. Falconer. A subadditive thermodynamic formalism for mixing repellers. J. Phys. A, 21 (1988) L737–L742. 1.1

[28] D.-J. Feng, Lyapunov exponent for products of matrices and Multifractal analysis. Part I: Positive matrices. *Israel J. of Math.*, 138 (2003), 353–376. 1.1, 3.2.3

[29] D-J. Feng and W. Huang. Lyapunov spectrum of asymptotically sub-additive potentials. Comm. Math. Phys. 297 (2010) 1–43. 1.1

[30] H. H. Ferreira, A. O. Lopes and S. R. C. Lopes, Decision Theory and Large Deviations for Dynamical Hypothesis Test: Neyman-Pearson, min max and Bayesian, Journal of Dynamics and Games (on line) 1.1, 2.5, 3.3

[31] H. Furstenberg, Noncommuting Random Products. Trans. Amer. Math. Soc. 108:3 (1963) 377–428. 3.2.3

[32] G. Gallavoti, Nonequilibrium and fluctuation relation. J. Stat. Phys. 180 (2020) 172–226. 1.1

[33] G. Gallavoti, Chaotic hypothesis: Onsager reciprocity and fluctuation-dissipation theorem, J. Stat. Phys., 84 (1996) 899–925. 1.1

[34] V. Girardin and P. Regnault, Escort distributions minimizing the Kullback-Leibler divergence for a large deviations principle and tests of entropy level. Ann. Inst. Statist. Math. 68 (2016), no. 2, 439-468. 3.1

[35] P. Giulietti, B. Kloeckner, A. O. Lopes and D. Marcon, The calculus of thermodynamical formalism, Journ. of the European Math Society, Vol 20, Issue 10, pages 2357–2412 (2018) 1.1

[36] C. Ji. Estimating functionals of one-dimensional Gibbs states. Probab. Theory Related Fields 82:2 (1989) 2, 155-175. 1.1

[37] W. Jiang and M. Tanner. Gibbs posterior for variable selection in high-dimensional classification and data mining. The Annals of Statistics, 2207–2231, 2008. (document), 1.2

[38] Y. Kifer, Large Deviations in Dynamical Systems and Stochastic processes, Trans. Amer. Math. Soc, 321:2 (1990) 505–524. 1.3

[39] A. O. Lopes, Entropy and Large Deviation, *Nonlinearity*, Vol. 3, N. 2, 527-546, 1990. (document), 1.3

[40] A. O. Lopes, Entropy, Pressure and Large Deviation, Cellular Automata, Dynamical Systems and Neural Networks, E. Goles e S. Martinez (eds.), Kluwer, Massachusets, pp. 79-146, 1994. (document)

[41] A. O. Lopes, Thermodynamic Formalism, Maximizing probability measures and Large Deviations, Preprint (2022) UFRGS 1.2, 2.1

[42] A. Lopes and J. Mengue. Duality Theorems in Ergodic Transport. Journal of Statistical Physics. Vol 149. issue 5. (2012), 921-942. 1.1

[43] A. Lopes, J. Mengue, J. Mohr and R. Souza, Entropy, Pressure and Duality for Gibbs plans in Ergodic Transport, Bull. of the Brazilian Math. Soc. Vol 46 - N 3 - 353–389 (2015) 1.1

[44] A. Lopes and J. Mengue, On information gain, Kullback-Leibler divergence, entropy production and the involution kernel, arXiv (2020) 3.1

[45] A. O. Lopes and R. Ruggiero The sectional curvature of the infinite dimensional manifold of Hölder equilibrium probabilities, Preprint arXiv:1811.07748v7 1.1

[46] A. O. Lopes and R. Ruggiero Nonequilibrium in Thermodynamic Formalism: the Second Law, gases and Information Geometry, Qualitative Theory of Dynamical Systems 21: 21 p 1-44 (2022) 1.1

[47] K. McGoff, S. Mukherjee and A. Nobel, Gibbs posterior convergence and Thermodynamic formalism, Annals of Applied Probability (to appear) (document), 1.1, 1.2, 1.2, 1.3, 2.5

[48] W. Parry and M. Pollicott. Zeta functions and the periodic orbit structure of hyperbolic dynamics, Asterisque, Vol 187-188 1990 1.2, 1.2, 1.3, 1.3

[49] V. K. Rohatgi, An Introduction to Probability Theory and Mathematical Statistics, Wiley (1976) 1.1

[50] D. Ruelle. Analyticity Properties of the Characteristic Exponents of Random Matrix Products, Adv. Math.,32 (1979) 68–80. 3.2.3

[51] M. J. Schervish, Theory of Statistics, Springer Series in Statistics, Springer-Verlag, New York, 1995. 1.1

[52] A. M. Shur, Growth properties of power-free languages, Computer Science Review 6: 5–6 (2012) 187–208. 1.1

[53] F. Spitzer. A Variational characterization of finite Markov chains. *The Annals of Mathematical Statistics.* (43): N.1 303-307, 1972. 2.1

[54] Y. Suhov and M. Kelbert, Probability and statistics by example. II, Cambridge Press (2014) 1.1

[55] L. Su and S. Mukherjee, A large deviation approach to posterior consistency in dynamical systems, Preprint ArXiv::2106.06894 1.1

[56] V. Tan, A. Anandkumar, L. Tong and A. Willsky, A large-deviation analysis of the maximum-likelihood learning of Markov tree structures. IEEE Trans. Inform. Theory 57 (2011), no. 3, 1714-–1735. 2.1, 3.3

[57] P. Varandas and Y. Zhao, Weak Gibbs measures: speed of convergence to entropy, topological and geometrical aspects, Ergodic Theory Dynam. Systems 37, no. 7, 2313–2336 (2017) (document), 1.1, 3.3

[58] W. von der Linden, V. Dose and U. von Toussaint. *Bayesian Probability Theory, Applications in the Physical Sciences*. Cambridge: Cambridge University Press, 2014 1.1

Universidade Federal do Rio Grande do Sul, 91509-900 MAT, Porto Alegre, Brasil
E-mail address: arturoscar.lopes@gmail.com

Universidade Federal do Rio Grande do Sul, 91509-900 MAT, Porto Alegre, Brasil
E-mail address: silviarc.lopes@gmail.com

Instituto de Matemática e Estatística, Universidade Federal da Bahia, 40170-110, Brazil
E-mail address: paulo.varandas@ufba.br