The Annals of Applied Probability
2022, Vol. 32, No. 1, 461–496
https://doi.org/10.1214/21-AAP1685

GIBBS POSTERIOR CONVERGENCE AND THE THERMODYNAMIC FORMALISM

BY KEVIN MCGOFF[1], SAYAN MUKHERJEE[2] AND ANDREW B. NOBEL[3]

[1]*Department of Mathematics and Statistics, University of North Carolina at Charlotte, kmcgoff1@uncc.edu*
[2]*Departments of Statistical Science, Mathematics, Computer Science, and Biostatistics & Bioinformatics, Duke University, sayan@stat.duke.edu*
[3]*Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, nobel@email.unc.edu*

In this paper we consider the posterior consistency of Bayesian inference procedures when the family of models consists of appropriate stochastic processes. Specifically, we suppose that one observes an unknown ergodic process and one has access to a family of models consisting of dependent processes. In this context, we consider Gibbs posterior inference, which is a loss-based generalization of standard Bayesian inference. Our main results characterize the asymptotic behavior of the Gibbs posterior distributions on the space of models. Furthermore, we show that in the case of properly specified models our convergence results may be used to establish posterior consistency. Our model processes are defined via the thermodynamic formalism for dynamical systems, and they allow for a large degree of dependence, including both Markov chains of unbounded orders and processes that are not Markov of any order. This work establishes close connections between Gibbs posterior inference and the thermodynamic formalism for dynamical systems, which we hope will lead to new questions and results in both nonparametric Bayesian analysis and the thermodynamic formalism.

1. Introduction. In this paper we study the posterior convergence of a generalized Bayes inference procedure for fitting an observed stochastic process to a model family consisting of a suitably parametrized collection of Gibbs measures. We focus on Gibbs measures for finite alphabet ergodic processes, which we will call Gibbs processes. Gibbs processes include irreducible Markov chains and processes with infinite order dependence. Standard approaches to Bayesian inference require full specification or substantial knowledge of the data generating process, which is usually assumed to belong to the family of models under study. However, when studying dynamical or other complex systems, this requirement is difficult to verify and often unrealistic. Instead, we assume the existence of a loss-function that can be used to assess how well a sequence generated by a model process fits the observed data sequence. We then study the resulting Gibbs posterior, which is a loss-based extension of the standard Bayesian posterior that does not require specification of the data generating model. The standard Bayesian posterior corresponds to the special case of negative log-likelihood loss.

We characterize the asymptotic behavior of the Gibbs posterior distribution on the parameter space as the number of observations tends to infinity. In particular, we establish that the limiting exponential growth rate of the normalizing constant (partition function) of the Gibbs posterior is characterized by a variational problem over the space of joinings of the observed and model systems. Moreover, we show that the Gibbs posterior distributions concentrate around the solution set of this variational problem. The variational problem is a generalization of the well-known variational principle for Gibbs measures. In the case of properly

Received August 2020; revised January 2021.
MSC2020 subject classifications. Primary 62M09; secondary 37D35.
Key words and phrases. Bayesian posterior consistency, thermodynamic formalism, Gibbs measure, Gibbs posterior principle.

specified models our convergence results may be used to establish posterior consistency. We apply the posterior convergence result to the direct observation of a Gibbs process as well as hidden Gibbs processes, generalizing previous posterior consistency results for Markov and hidden Markov models in Bayesian nonparametrics.

Both Gibbs posteriors and Gibbs measures have close connections with statistical physics; our analysis shows that they provide a natural framework for generalized Bayesian inference. Our work relies in part on ideas from the thermodynamic formalism and the theory of dynamical systems, and one of its primary contributions is demonstrating how these ideas can be brought to bear on problems of statistical inference.

1.1. *Connections to previous work.* The work in this paper lies at the intersection of several research areas, including Bayesian nonparametrics, inference for stochastic processes, generalized Bayes inference, and ergodic theory and dynamical systems. Here we discuss some related literature.

In the i.i.d. setting, Doob [13] established Bayesian posterior consistency for almost every parameter value in the support of the prior using martingale methods. Subsequently, Schwartz [48] gave necessary and sufficient conditions for posterior consistency at individual parameter values. The challenges of establishing posterior consistency for nonparametric models were highlighted by Diaconis and Freedman in [11] (see also [12]). Bayesian nonparametrics remains an active area of research: for a detailed review we refer the reader to the recent books of Ghosal and van der Vaart [19] and of Ghosh and Ramamoorthi [20].

Recent work on inference from stationary ergodic processes and dynamical systems includes denoising (filtering) [30, 31], consistency of maximum likelihood estimation [38], forecasting and density estimation [24, 50], empirical risk minimization [40, 41], as well data assimilation and uncertainty quantification [32]. More information can be found in the survey [39].

A number of researchers have been working on Bayesian posterior consistency for finite state hidden Markov chains [8, 14, 17, 51]. Our work generalizes these results, covering the setting of deterministic dynamics in the state transitions, as well as models with longer range dependencies.

Shalizi [49] considered posterior consistency for dependent processes, extending the testing-based approach of Schwartz [48]. His work describes a framework for establishing posterior consistency based on a set of general assumptions, among which are (i) the existence of a sieve-like structure for capacity control that must be compatible with the prior, the model family, and the observed process, (ii) finiteness of the relative entropy rate between the observed process and a set of models having positive prior measure, and (iii) the existence of a dominating measure for the model family. In the setting studied here, the first condition is difficult to verify and the latter two conditions will not hold in general.

Our approach and results differ from those in the existing literature in several respects. We do not adopt the testing-based (or sieve-based) approach of Schwartz and others. Instead, our arguments proceed in a direct way from the definition of the Gibbs model families and the Gibbs posterior construction, making only mild assumptions on the regularity of the model families and the integrability and modulus of continuity of the loss function.

The idea of a variational formulation of Bayesian inference was developed by Zellner [59], and the link between statistical mechanics and information theory with Bayesian inference was at the heart of the inference framework advocated by Edwin T. Jaynes [25]. The motivation for generalized Bayes is to have a coherent inferential procedure that quantifies uncertainty when the model may be misspecified, the likelihood is difficult to compute, or the data generating process is known modulo invariants or equivariance. Generalized Bayes can adapt classical conditional probability updating to these settings.

Generalized Bayes inference refers to loss-based procedures for updating prior beliefs. These procedures remain valid when one does not have access to a true likelihood; standard Bayesian inference corresponds to the negative log likelihood loss. The idea of using loss functions to update beliefs goes back at least to Vovk [52]. It has played a central role in the PAC-Bayesian approach to statistical learning [6, 37] and has been adopted by the mainstream Bayesian community [4, 42]. In [27] the term Gibbs posterior was introduced, and the advantages of the Gibbs posterior over standard Bayesian approaches in some settings was demonstrated. In [22] the authors provide a generalized Bayesian framework (closely related to the Gibbs posterior) that is consistent under misspecification. In [23] consistency and rates of convergence are obtained for generalized Bayesian methods including the Gibbs posterior as well as PAC-Bayes procedures. In [2] a general inference procedure called data dependent measures is introduced, of which Gibbs posteriors are a special case, along with consistency and rates of convergence in the i.i.d. setting. The Gibbs posterior framework was adapted in [62] for uncertainty quantification for inverse problems involving the solution of partial differential equations.

Bayesian inference for infinite dimensional problems has been explored in the control theory and inverse problems literatures [36, 43, 45]. Whereas our work considers posterior consistency for discrete, deterministic dynamics, the generating process in [36] is a stochastic differential equation. In [43, 45] the authors consider nonlinear estimation and communication from a variational Bayesian point of view, establishing close connections between information theoretic quantities and associated primary and secondary Bayes problems. While the objectives of this work are different than ours, the underlying models are similar to those in this paper, and exploring connections in more detail is clearly of interest.

The setting and results of [40] and [41] may be considered frequentist analogues of the present work. They consider empirical risk minimization based on observation of an unknown ergodic process, a model family determined by a continuous self-map of a compact space, and a loss function relating observations and model trajectories. The asymptotic behavior of empirical risk minimization is determined by a variational problem that is similar in spirit, but substantially simpler than that arising in the Bayesian setting studied here.

The thermodynamic formalism in dynamical systems, originally pioneered by Sinai, Ruelle, and Bowen, arises from the study of statistical physics, and has played a large role in the development of dynamical systems over many years. For an introduction to the area and some connections to statistical physics, see the books by Bowen [5], Ruelle [46], or Walters [53]. Let us mention that connections to Markov chains and other stochastic processes have a long history in this area [3, 57, 58].

1.2. *Overview.* The next section describes our general framework for the observed system and model families. In Section 3, we describe Gibbs posterior inference and introduce loss functions in our setting. Section 4 contains statements and discussion of our main results, and Section 5 describes our main application of these results, posterior consistency for hidden Gibbs processes. In Section 6, we present a detailed discussion of the rate function that arises in our main results. Finally, after establishing a technical preliminary in Section 7, we present the proofs of our main results in Sections 8–10. Note that we also provide several appendices containing background material and routine technical results that are used throughout our proofs.

2. Observed system and model family.

2.1. *Observed system.* Our inference framework consists of two main components. The first component is an observed stochastic process, which we formulate in terms of dynamical

systems as follows. Let $\mathcal{Y}$ be a complete separable metric space. Here and throughout this work we assume that all such spaces are endowed with their Borel σ-algebras, and we suppress this choice in our notation. Let $T : \mathcal{Y} \to \mathcal{Y}$ be a Borel measurable map. We let $\mathcal{M}(\mathcal{Y})$ denote the set of Borel probability measures on $\mathcal{Y}$, endowed with the weak* topology on measures. A measure $\nu \in \mathcal{M}(\mathcal{Y})$ is said to be invariant under T if $\nu(T^{-1}E) = \nu(E)$ for all Borel sets $E \subset \mathcal{Y}$. The set of T-invariant measures in $\mathcal{M}(\mathcal{Y})$ is denoted by $\mathcal{M}(\mathcal{Y}, T)$. Furthermore, $\nu \in \mathcal{M}(\mathcal{Y}, T)$ is said to be ergodic if $\nu(E) \in \{0, 1\}$ for all Borel sets E satisfying $T^{-1}(E) = E$. We assume in what follows that we observe the trajectory of a system $(\mathcal{Y}, T, \nu)$ where $\nu \in \mathcal{M}(\mathcal{Y}, T)$ is ergodic.

2.2. *Gibbs measures.* The second component of our inference framework is a collection of models. In order to model dependence in the standard statistical setting, one typically considers (hidden) Markov models or more complex state space models. In our analysis we would like to be able to handle model processes with long range dependencies, and so we consider a general class of processes arising from so-called Gibbs measures in dynamical systems. This class of processes strictly generalizes the class of finite state Markov models with arbitrarily large order (see Example 1). Furthermore, as we discuss below, the thermodynamic formalism guarantees that these model families can be continuously parametrized and that they admit strong exponential estimates. These properties ensure that our model families are suitable for Gibbs posterior inference.

Let us also remark that Gibbs measures and their associated processes have strong connections to lattice models in statistical physics, such as the Ising model or the Potts model. Lattice models have been used in statistics for problems like image segmentation [18] and Bayesian variable selection [34]. The use of lattice models for inference requires that a unique probability measure exists and can be specified for the configuration space on the lattice: if a unique measure cannot be specified, then quantities such as the posterior and the likelihood would not be well defined. As we describe below, the thermodynamic formalism provides general conditions under which the configuration space on the lattice $\mathbb{Z}^d$ has a distinguished measure. For stochastic processes, the natural lattice to consider is $\mathbb{Z}$.

Before giving a precise definition of a Gibbs measure and its corresponding stochastic process, we must first introduce the underlying state space for such models, which is called a mixing shift of finite type (SFT). A shift of finite type is a dynamical system that is the topological analogue of a finite state aperiodic and irreducible Markov chain. SFTs have been widely studied in the dynamical systems literature, both for their own sake [35] and as model systems for some smooth systems such as Axiom A diffeomorphisms [5]. Furthermore, SFTs have substantial connections to statistical physics and other fields such as coding and information theory [35, 46].

Here we give a proper definition for a mixing SFT. Let $\mathcal{A}$ be a finite set, known as an alphabet, and let $\Sigma = \mathcal{A}^{\mathbb{Z}}$ be the set of bi-infinite sequences $x = (x_n)$ with values in $\mathcal{A}$. For $i \le j$ in $\mathbb{Z}$, we set $x_i^j = x_i \dots x_j$. Define the left-shift map $\sigma : \Sigma \to \Sigma$ by $\sigma(x)_n = x_{n+1}$. A set $\mathcal{X}$ is called an SFT if there exists $n \ge 0$ and a collection of words $\mathcal{W} \subset \mathcal{A}^n$ such that $\mathcal{X}$ is exactly the set of sequences in Σ that contain no words from $\mathcal{W}$:

$$\mathcal{X} = \{x \in \Sigma : \forall i \in \mathbb{Z},\ x_{i+1} \dots x_{i+n} \notin \mathcal{W}\}.$$

Here $\mathcal{W}$ is called a set of *forbidden words* for $\mathcal{X}$. Note that by choosing $\mathcal{W} = \varnothing$, one obtains the full sequence space Σ, which is known as the full shift (on the alphabet $\mathcal{A}$). Also, we endow $\mathcal{A}$ with the discrete topology and Σ with the product topology, which makes any such $\mathcal{X}$ closed and compact. We define the map $S : \mathcal{X} \to \mathcal{X}$ to be the restriction of the left shift σ to $\mathcal{X}$. Let $\mathcal{L}_m$ denote the set of words of length m (i.e., elements of $\mathcal{A}^m$) that appear in at least one point of $\mathcal{X}$, and let $\mathcal{L} = \bigcup_{m \ge 0} \mathcal{L}_m$. An SFT $\mathcal{X}$ is said to be mixing if for any

two words $u, v \in \mathcal{L}$, there exists N such that for all $m \geq N$, there exists a word $w \in \mathcal{L}_m$ such that $uwv \in \mathcal{L}$. The following equivalent definition is perhaps more intuitive to readers familiar with Markov chains. Let A be the square matrix indexed by $\mathcal{A}^n$ defined for for words $u, v \in \mathcal{A}^n$ by the rule

$$A_{uv} = \begin{cases} 1, & \text{if } \exists x \in \mathcal{X} \text{ such that } x_0^{n-1} = u \text{ and } x_1^n = v \\ 0, & \text{otherwise.} \end{cases}$$

Then $\mathcal{X}$ is mixing if and only if there exists $N \geq 1$ such that A^N contains all positive entries. Our standard assumption on $\mathcal{X}$ is that it is a mixing SFT.

To model stochastic behavior on the topological system $(\mathcal{X}, S)$, we consider a family of stochastic processes defined by S-invariant probability measures on $\mathcal{X}$, called Gibbs measures. To introduce Gibbs measures, one begins with a function $f : \mathcal{X} \to \mathbb{R}$, which is called a potential function (or just a potential). A Borel probability measure μ on $\mathcal{X}$ is said to be a Gibbs measure corresponding to the potential function $f : \mathcal{X} \to \mathbb{R}$ if there exists constants $\mathcal{P} \in \mathbb{R}$ and $K > 0$ such that for all $x \in \mathcal{X}$ and $m \geq 1$,

$$K^{-1} \leq \frac{\mu(x[0, m-1]))}{\exp(-\mathcal{P}m + \sum_{k=0}^{m-1} f(S^k(x)))} \leq K, \tag{1}$$

where $x[0, m-1]$ is the cylinder set of points y in $\mathcal{X}$ such that $x_i = y_i$ for all $i = 0, \ldots, m-1$. The property in (1) is called the Gibbs property. By a celebrated result of Bowen [5], under mild regularity conditions on f, there is a unique Gibbs measure $\mu \in \mathcal{M}(\mathcal{X}, S)$ with potential function f, and furthermore the measure μ is ergodic. The constant $\mathcal{P} = \mathcal{P}(f)$ is called the *pressure* of f.

DEFINITION 1. Let $f : \mathcal{X} \to \mathbb{R}$ be a potential with unique Gibbs measure μ_f. Then the *Gibbs process* with potential f is the process $\mathbb{X}_f = \{X_i\}_{i \in \mathbb{Z}}$ with $X_i \in \mathcal{A}$ and distribution μ_f.

The Gibbs measure is a generalization of the canonical ensemble in statistical physics to infinite systems. Potential functions have natural connections with Hamiltonians in the study of lattice systems in statistical physics. In considering inference, we will think of loss functions as potential functions. We remark (again) that the class of Gibbs measures strictly generalizes the class of Markov chains, allowing for arbitrarily long dependencies. Indeed, Example 1 shows that any Markov chain of order k on the alphabet $\mathcal{A}$ can be realized as a Gibbs measure by an appropriate choice of a potential function that depends on only k coordinates. On the other hand, when the potential function f depends on infinitely many coordinates, the corresponding Gibbs measure is not Markov of any order. In this way, our model families may include Markov chains with unbounded orders, which highlights the degree of dependence allowed by our framework.

EXAMPLE 1. Suppose P is the transition matrix for an irreducible aperiodic Markov chain on the state space $\mathcal{A}$. Let p be the unique stationary distribution for P, and let $\mathbb{X}'$ be the corresponding Markov process. In this example we show how to construct a potential function f so that the associated Gibbs process $\mathbb{X}_f$ is equal in distribution to $\mathbb{X}'$. First, consider the $0-1$ matrix A such that $A_{uv} = 1$ if and only if $P_{uv} > 0$. Then let $\mathcal{X}$ be the SFT on alphabet $\mathcal{A}$ defined by the matrix A. Equivalently, one may take the set of forbidden words $\mathcal{W}$ to be the set of words uv of length two such that $P_{uv} = 0$. Note that $\mathcal{X}$ is mixing since P is irreducible and aperiodic. Next, define the potential function $f : \mathcal{X} \to \mathbb{R}$ by the rule $f(x) = -\log P(x_0, x_1)$. Then the corresponding Gibbs process $\mathbb{X}_f$ is equal in distribution to the Markov process $\mathbb{X}'$

(defined by stationary distribution p and transition matrix P). Similarly, one may check that if $\mathbb{Y} = \{Y_i\}_{i\in\mathbb{Z}}$ is a stationary irreducible aperiodic k-step Markov chain on $\mathcal{A}$, and if $\mathcal{X}$ is the SFT with forbidden words $x_0^k \in \mathcal{A}^{k+1}$ such that $\mathbb{P}(Y_0^k = x_0^k) = 0$ and $f : \mathcal{X} \to \mathbb{R}$ is the potential function $f(x) = -\log \mathbb{P}(Y_k = x_k | Y_0^{k-1} = x_0^{k-1})$, then the associated Gibbs process $\mathbb{X}_f$ is equal in distribution to $\mathbb{Y}$.

2.3. *Model families.* In order to perform statistical estimation or inference, we require not just a single model process, but rather a family of processes. In this section we specify general conditions under which a family of Gibbs processes (see Definition 1) is suitable for Gibbs posterior inference. As Gibbs processes are uniquely defined by their potential functions, we specify a *family* of Gibbs processes by parametrizing a family of potential functions. In order to state our regularity condition on families of potential functions, we require some additional definitions.

For points x, y in $\mathcal{X}$, we let $n(x, y)$ denote the infimum of all $|m|$ such that $x_m \neq y_m$. We then set $d_{\mathcal{X}}(x, y) = 2^{-n(x,y)}$, and we remark that $d_{\mathcal{X}}(\cdot, \cdot)$ is a metric on $\mathcal{X}$ (see [35], p. 174). For $r > 0$, we let $C^r(\mathcal{X})$ denote the set of continuous functions from $\mathcal{X}$ to $\mathbb{R}$ with Hölder exponent r, that is, the set of functions $f : \mathcal{X} \to \mathbb{R}$ for which there exists a constant c such that for all $x, y \in \mathcal{X}$,

$$\big|f(x) - f(y)\big| \leq c\, d_{\mathcal{X}}(x, y)^r.$$

Furthermore, we endow $C^r(\mathcal{X})$ with the topology induced by the norm $\|\cdot\|_r$, where

$$\|f\|_r = \sup_{x\in\mathcal{X}} \big|f(x)\big| + \sup_{x\neq y} \frac{|f(x) - f(y)|}{d_{\mathcal{X}}(x, y)^r}.$$

Now we define the regularity condition necessary for our model families.

DEFINITION 2. Let Θ be a compact metric space with metric d_Θ. A parametrized family of potential functions $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ will be called a *regular family* if there exists $r > 0$ such that $\mathcal{F} \subset C^r(\mathcal{X})$ and the map $\theta \to f_\theta$ is continuous in the topology induced by the norm $\|\cdot\|_r$.

If a family $\{f_\theta : \theta \in \Theta\}$ is a regular family, then the map $\theta \mapsto \mu_\theta$ is continuous in the weak* topology on measures, and the constants $K(f_\theta)$ and $\mathcal{P}(f_\theta)$ that appear in (1) depend continuously on θ (see [1]). Furthermore, since Θ is compact, a uniform Gibbs property holds: there exists a uniform constant K and a continuous function $\theta \mapsto \mathcal{P}(f_\theta)$ such that for all $\theta \in \Theta$, $x \in \mathcal{X}$, and $m \geq 1$,

$$K^{-1} \leq \frac{\mu_\theta(x[0, m-1])}{\exp(-\mathcal{P}(f_\theta)m + \sum_{k=0}^{m-1} f_\theta(S^k x))} \leq K. \tag{2}$$

We assume throughout that $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ is a regular family of potential functions, and that our model class consists of the corresponding Gibbs measures $\{\mu_\theta : \theta \in \Theta\}$, or equivalently, Gibbs processes $\{\mathbb{X}_\theta : \theta \in \Theta\}$, where we define $\mathbb{X}_\theta = \mathbb{X}_{f_\theta}$. In this way we obtain a continuously parametrized family of measures, or equivalently dependent processes, that are characterized by the potential functions f_θ. These model families substantially generalize parameterized families of finite order Markov chains.

3. Gibbs posterior inference. The inference paradigm we consider is known as Gibbs posterior inference, which is a generalization of the standard Bayesian inference framework. The basic idea behind the Gibbs posterior [4, 27] is to replace the likelihood with an exponentiated loss or utility function in the standard Bayesian procedure for updating beliefs about an unknown parameter of interest θ. Whereas the standard Bayes posterior takes the form

$$\pi(\theta|\text{data}) = \frac{\text{Likelihood}(\text{data}|\theta) \times \pi(\theta)}{\int_\Theta \text{Likelihood}(\text{data}|\theta') \times \pi(\theta')\, d\theta'},$$

the Gibbs posterior has the form

$$\pi(\theta|\text{data}) = \frac{\exp(-\ell(\text{data}, \theta)) \times \pi(\theta)}{\int_\Theta \exp(-\ell(\text{data}, \theta')) \times \pi(\theta')\, d\theta'},$$

where $\ell(\text{data}, \theta)$ is the loss associated with θ based on the observed data. When the loss function is the negative log-likelihood then the two paradigms are identical. The original motivation for the Gibbs posterior was to specify a coherent procedure for Bayesian inference when the parameter of interest is connected to observations via a loss function, rather than the classical setting where the likelihood or true sampling distribution is known; see [4] for more arguments in favor of the Gibbs posterior and discussion about how the Gibbs posterior framework addresses model misspecification and robustness to nuisance parameters. Note that in the general Gibbs posterior framework without a likelihood, there is no generative model assumed for the observations.

Recall that as part of our standard assumptions, the model class $\{\mu_\theta : \theta \in \Theta\}$ is a family of Gibbs measures on $\mathcal{X}$ corresponding to a regular family of potential functions (Definition 2) indexed by parameters from a compact metric space Θ with metric d_Θ. The elements of Θ will also be used to parametrize the relationship between states and observations, for example, emission probabilities in hidden Gibbs processes. Define a metric on $\Theta \times \mathcal{X}$ by

$$d\big((\theta, x), (\theta', x')\big) = \max\big(d_\Theta(\theta, \theta'), d_{\mathcal{X}}(x, x')\big).$$

Recall that the observed system has a Polish state space $\mathcal{Y}$ with invariant measure ν. Here and throughout this work, we assume that we have a loss function $\ell : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ satisfying the following conditions:

(i) ℓ is continuous;

(ii) there exists a measurable function $\ell^* : \mathcal{Y} \to \mathbb{R}$ such that for all $y \in \mathcal{Y}$, $\sup_{\theta,x} |\ell(\theta, x, y)| \le \ell^*(y)$, and $\int \ell^*\, d\nu < \infty$;

(iii) for each $\delta > 0$ there exists a measurable function $\rho_\delta : \mathcal{Y} \to (0, \infty)$ such that for each $y \in \mathcal{Y}$,

$$\sup\big\{\big|\ell(\theta, x, y) - \ell\big(\theta', x', y\big)\big| : d\big((\theta, x), (\theta', x')\big) \le \delta\big\} \le \rho_\delta(y),$$

and $\lim_{\delta \to 0^+} \int \rho_\delta\, d\nu = 0$.

Condition (ii) is an integrability condition on the loss, while condition (iii) is a requirement on the modulus of continuity of the loss. In Section 4.1 we provide examples of loss functions satisfying these conditions. Note that including the parameter θ in the loss function may be considered nonstandard in statistics. However, this formulation will simplify notation throughout the paper, and in Section 4.1 we establish that this setting is equivalent to the standard one. Also note that the dependence of the loss on Θ and on the uncountable space $\mathcal{X}$ allows us to model continuous observations and emission probabilities.

With the loss function and parameter $\theta \in \Theta$ fixed, we define the loss of the finite sequence $x_0^{n-1} \in \mathcal{X}^n$ with respect to a finite sequence of observations $y_0^{n-1} \in \mathcal{Y}^n$ to be the sum of the

per-state losses:

$$\ell_n\big(\theta; x_0^{n-1}, y_0^{n-1}\big) = \sum_{k=0}^{n-1} \ell(\theta, x_k, y_k). \tag{3}$$

When $x_0^{n-1} = (x, Sx, \ldots, S^{n-1}x)$ and $y_0^{n-1} = (y, Ty, \ldots, T^{n-1}y)$ are initial segments of trajectories of S and T (defined in Section 2), respectively, we write $\ell_n(\theta, x, y)$ instead of $\ell_n(\theta; x_0^{n-1}, y_0^{n-1})$.

Let us now give the definition of Gibbs posterior distributions on Θ. Here we consider the (subjective) case, in which one begins with a fully supported prior probability measure π_0 on Θ. As a first step in obtaining the Gibbs posterior, we extend π_0 to a prior distribution on $\Theta \times \mathcal{X}$. In detail, given the family $\{\mu_\theta : \theta \in \Theta\}$ of Gibbs measures on $\mathcal{X}$, define the induced prior distribution P_0 on $\Theta \times \mathcal{X}$ by

$$P_0(E) = \int\int \mathbf{1}_E(\theta, x)\, d\mu_\theta(x)\, d\pi_0(\theta) \tag{4}$$

for any Borel set $E \subset \Theta \times \mathcal{X}$. Under the Gibbs posterior paradigm [4, 27], given observations $(y, Ty, \ldots, T^{n-1}y) \in \mathcal{Y}^n$, our updated beliefs are represented by the Gibbs posterior distribution $P_n(\cdot|y)$ on $\Theta \times \mathcal{X}$ defined for Borel sets $E \subset \Theta \times \mathcal{X}$ by

$$P_n(E|y) = \frac{1}{Z_n(y)} \int_E \exp\big(-\ell_n(\theta, x, y)\big)\, dP_0(\theta, x) \tag{5}$$

$$= \frac{1}{Z_n(y)} \int\int \mathbf{1}_E(\theta, x) \exp\big(-\ell_n(\theta, x, y)\big)\, d\mu_\theta(x)\, d\pi_0(\theta). \tag{6}$$

Here $Z_n(y)$ is the normalizing constant (also known as the partition function in statistical physics terminology), given by

$$Z_n(y) = \int \exp\big(-\ell_n(\theta, x, y)\big)\, dP_0(\theta, x).$$

The Gibbs posterior distribution $\pi_n(\cdot|y)$ on Θ is simply the Θ-marginal of $P_n(\cdot|y)$, defined for Borel sets $A \subset \Theta$ by

$$\pi_n(A|y) = P_n(A \times \mathcal{X}|y).$$

We are interested in the asymptotic behavior of the posterior distributions $\pi_n(\cdot|y)$. As the observed process need not be in our model family, standard notions of posterior consistency are not appropriate. Instead, we establish that the posteriors $\pi_n(\cdot|y)$ concentrate on the set of parameters that minimize a lower semicontinuous rate function. In this sense, our inferential focus is on parameters θ, and not the initial states x of the models, as the latter is known to be impossible for many dynamical systems, including shifts of finite type [30, 31]. Let us summarize our framework.

- We begin with a fully supported prior π_0 on a compact set Θ that smoothly parametrizes a family of Gibbs measures $\{\mu_\theta : \theta \in \Theta\}$ on $\mathcal{X}$.
- From π_0 and $\{\mu_\theta : \theta \in \Theta\}$, we obtain an extended prior P_0 on $\Theta \times \mathcal{X}$.
- We observe the initial trajectories $y, \ldots, T^{n-1}y$ of a stationary ergodic system $(\mathcal{Y}, T, \nu)$.
- From P_0, the observed initial trajectory, and the loss function ℓ we obtain the Gibbs posterior P_n on $\Theta \times \mathcal{X}$.
- Finally, we marginalize P_n to get the posterior π_n on Θ.

4. Main results. Before stating our main results, let us briefly summarize the standard assumptions that we make here and throughout this work. We assume that Θ is a compact metric space, π_0 is a fully supported prior distribution on Θ, $\{f_\theta : \theta \in \Theta\}$ is a regular family of potential functions on a mixing SFT $\mathcal{X}$ (as in Definition 2), $\{\mu_\theta : \theta \in \Theta\}$ is the corresponding family of Gibbs process measures, $(\mathcal{Y}, T, \nu)$ is a stationary ergodic probability-preserving system on a complete separable metric space equipped with the Borel σ-algebra, and $\ell : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a loss function satisfying conditions (i)–(iii). With these objects in place, we consider Gibbs posterior inference, as described in Section 3 above. Our analysis begins with the exponential growth rate of the (random) partition function Z_n for large n. In particular, we establish a variational principle for the almost sure limit of $n^{-1} \log Z_n$ as n tends to infinity.

THEOREM 1. *Under the standard assumptions above there exists a lower semicontinuous function $V : \Theta \to \mathbb{R}$ such that for ν-almost every $y \in \mathcal{Y}$,*

$$\lim_n -\frac{1}{n} \log Z_n(y) = \inf_{\theta \in \Theta} V(\theta).$$

REMARK 1. The compactness of Θ and lower semicontinuity of V ensure that the infimum in Theorem 1 is obtained. The conclusion of the theorem is similar to a large deviations principle (see, e.g., [15]), with $V : \Theta \to \mathbb{R}$ playing the role of the rate function. For this reason, we refer to V as the rate function in this setting. A detailed discussion of V appears in Section 6, where we show that V can be expressed as the sum of an expected loss term and a divergence term.

The variational expression that appears in Theorem 1 suggests that we focus on the (nonempty, compact) set of parameters θ that minimize the rate function,

$$\Theta_{\min} = \operatorname*{argmin}_{\theta \in \Theta} V(\theta).$$

The Gibbs posterior distribution is asymptotically concentrated on this set.

THEOREM 2. *For each open neighborhood U of $\Theta_{\min}$ and ν-almost every $y \in \mathcal{Y}$,*

$$\lim_n \pi_n(\Theta \setminus U | y) = 0.$$

In light of this result, it is possible to answer questions about Gibbs posterior consistency by analyzing the variational problem defining $\Theta_{\min}$. We illustrate this approach to posterior consistency in our main application, hidden Gibbs processes (see Section 5).

REMARK 2 (Optimality of $\Theta_{\min}$). One may wonder whether $\pi_n(\cdot|y)$ in fact concentrates around a strict subset of $\Theta_{\min}$. Proposition 8 (in Appendix B) addresses this question on the exponential scale. It states that if $U \subset \Theta$ is open and intersects $\Theta_{\min}$, then the posterior probability of U cannot be exponentially small as n tends to infinity, that is, for ν-almost every y the quantity $n^{-1} \log \pi_n(U|y)$ tends to zero as n tends to infinity.

REMARK 3 (Support of the prior). Recall that as part of our standard assumptions, we assume that the prior π_0 is fully supported on Θ. In general, the topological support of π_0 will be closed (by definition) and therefore compact, as Θ is compact by assumption. Thus, if the support of π_0 is a strict subset $\Theta' \subsetneq \Theta$, then Theorems 1 and 2 continue to hold with Θ replaced by Θ'.

REMARK 4 (Functions of Markov chains and differentiable state space models). Although we present our results here in the context of Gibbs processes, we note that analogous results may be immediately obtained for functions of Gibbs processes. In particular, our results hold for appropriate model families consisting of functions of mixing Markov chains, which may not be Markov of any order. Additionally, results similar to those here may be established for certain state space models in which the underlying dynamics are governed by families of differentiable dynamical systems on manifolds (as in [38]). In particular, using our results and the well-known connections between SFTs and Axiom A systems (see [5]), it is possible to establish analogous conclusions for Axiom A diffeomorphisms with Gibbs measures.

REMARK 5 (Ground states and MAP). From a thermodynamic perspective, it is natural to introduce an inverse temperature parameter $\beta \in \mathbb{R}$ and consider the new loss function $\ell_\beta(\theta, x, y) = \beta \cdot \ell(\theta, x, y)$. In this setting, one would like to understand what happens as β tends to infinity. In Section 6.7, we identify the limit of both V and $\Theta_{\min}$ as β tends to infinity in terms of variational problems considered in previous work [40].

The use of an inverse temperature parameter has also been used in practice to perform maximum a posteriori (MAP) estimation. MAP estimation is a common alternative to fully Bayesian inference that is used in both statistics and machine learning. It involves finding the parameter that is the posterior mode. The motivation for MAP estimation is often computational efficiency and the lack of a need for uncertainty quantification. The idea of adding an inverse temperature parameter (β) to a Gibbs distribution for MAP estimation was introduced for Bayesian models in a seminal paper by Geman and Geman [18], who also gave an annealing schedule to increase the inverse temperature with a provable guarantee for finding the posterior mode.

REMARK 6 (Connections to penalization). The formulation of Bayesian updating as a variational problem with an entropic penalty has been previously explored [4, 59], and these ideas are related to Jaynes' maximum entropy formulation of Bayesian inference [26]. In both [4] and [59], posterior inference was formulated as follows: given a loss function $\ell(\theta, x)$ and a prior π, the posterior distribution is

$$\pi(\theta|x) = \arg\min_{\mu} \left\{ \int_\theta \ell(\theta, x)\, d\mu(\theta) + d_{KL}(\mu, \pi) \right\},$$

where $d_{KL}(\mu, \pi)$ is the relative entropy between μ and π. The function being minimized above has close connections to the function $V(\theta)$ in Theorem 1; see Definition 4 below.

REMARK 7 (Convergence of full Gibbs posteriors). Our main results establish the concentration of the Θ-marginal posterior distributions $\pi_n(\cdot|y)$ around the limit set $\Theta_{\min}$. In contrast, the $\mathcal{X}$-marginal of the full posterior distribution $P_n(\cdot|y)$ need not concentrate around any particular subset of $\mathcal{X}$ (according to the negative results of [30, 31]). Nonetheless, Proposition 9 (in Appendix C) gives a characterization of any Cesàro limit of the full posteriors.

REMARK 8 (Posterior contraction rate). In light of the posterior convergence guaranteed by Theorem 2, it is natural to ask about the rate of this convergence. We expect that the thermodynamic formalism could be used to obtain a posterior contraction rate under the assumptions of this paper, but our arguments cannot be easily adapted to yield such a rate. We mention it here as an interesting question for future research.

REMARK 9 (Importance of the Gibbs property). The Gibbs property (1) of the measures μ_θ makes them particularly suitable as model distributions for the purposes of Gibbs posterior inference. Broadly speaking, the existence of such exponential estimates for the model distributions renders them amenable to analysis, and similar estimates appear elsewhere in the Bayesian nonparametrics literature [19, 55]. At a technical level, this property ensures that the divergence term in the rate function V is well behaved. Finding generalizations of this property that are satisfied by additional model classes represents an interesting avenue for future research.

4.1. *Examples of inference settings and associated loss functions.* While the inference framework above allows very general observations, the model families we consider are restricted to Gibbs measures on shift spaces, which correspond to families of dependent finite-valued processes. The latter may seem limited when compared to real-valued or more abstract-valued processes. However, as the following examples illustrate, this is not the case. The sequence space $\mathcal{X}$ is typically uncountable, and, in conjunction with the left shift and appropriate state-observable maps, it can be used to generate real or more general valued processes from the finite state processes in the model family. This flexibility in modeling arises in part from the generality of the loss function, and its potential dependence on θ. Each of the examples below yield loss functions that satisfy conditions (i)–(iii).

EXAMPLE 2 (Continuous, deterministic observations). Suppose that the state space $\mathcal{Y}$ of the observed system is a subset of the real line, so that the observations $y, Ty, T^2y, \ldots$ are real-valued and deterministic. In this case, we may fit the observations to a family of continuous models generated by the Gibbs measures $\{\mu_\theta : \theta \in \Theta\}$ on $\mathcal{X}$ using a continuously parametrized family $\{\varphi_\theta : \mathcal{X} \to \mathbb{R}\}$ of continuous observation functions. Given θ and x, the initial part of the real-valued sequence $\{\varphi_\theta(S^k x)\}_{k\geq 0}$ can be fit to the observations. Models of this sort are called dynamical models, and they have been studied in the context of empirical risk minimization in [41]. If the measure ν has finite second moment, and ℓ is the squared loss $\ell(\theta, x, y) = |\varphi_\theta(x) - y|^2$, then conditions (i)–(iii) on the loss are satisfied.

EXAMPLE 3 (Discrete observations). Let $\mathcal{A}$ and $\mathcal{B}$ be finite sets. Suppose that we make $\mathcal{B}$-valued observations, that is, $\mathcal{Y} \subset \mathcal{B}^{\mathbb{Z}}$, and we wish to model these observations with a family of Gibbs measures $\{\mu_\theta : \theta \in \Theta\}$ on $\mathcal{X}$ with $\mathcal{X} \subset \mathcal{A}^{\mathbb{Z}}$. Let $\varphi : \mathcal{A} \to \mathcal{B}$ be an observation function, so that a point x in $\mathcal{X}$ gives rise to the $\mathcal{B}$-valued sequence $\{\varphi(x_k)\}_{k\geq 0}$. Let ℓ be the discrete loss, $\ell(\theta, x, y) = \mathbf{1}(\varphi(x_0) \neq y_0)$. Then the conditions (i)–(iii) on the loss are satisfied.

EXAMPLE 4 (Family of conditional likelihoods). Suppose that $\{p(\cdot|x, \theta) : \theta \in \Theta, x \in \mathcal{X}\}$ is a family of conditional densities on $\mathcal{Y}$ with respect to a common Borel measure m on $\mathcal{Y}$. Here $p(\cdot|x, \theta)$ is the conditional likelihood of a single observation given the parameter θ and system state x. Under appropriate continuity and integrability conditions on the family of likelihoods, the negative log-likelihood function, $\ell(\theta, x, y) = -\log p(y|x, \theta)$, satisfies conditions (i)–(iii). In this situation, the Gibbs posterior is the same as the standard Bayes posterior. Furthermore, the dependence of the loss on the parameter θ allows one to parametrize the conditional observation densities, as in the parametrization of emission densities in the study of hidden Markov models. Note that in the Gibbs posterior framework, the true observation system $(\mathcal{Y}, T, \nu)$ may be fully misspecified—it need not be related to any of the generative processes implied by the family of Gibbs measures and conditional likelihoods.

5. Application. In this section we present an application of our main results on Gibbs posterior consistency to standard posterior consistency for families of properly specified dependent processes. In particular, we establish Bayesian posterior consistency for hidden Gibbs processes. This result generalizes previous results on posterior consistency for hidden Markov models by allowing substantially more dependence in the hidden processes, including families of Markov chains with unbounded orders.

5.1. *Hidden Gibbs processes.* In this section we consider posterior consistency for more general observation processes. Let $\mathcal{X}$ be a mixing SFT, $\{f_\theta : \theta \in \Theta\} \subset C^r(\mathcal{X})$ a regular family of Hölder potential functions (as in Definition 2), and $\{\mu_\theta : \theta \in \Theta\}$ the corresponding family of Gibbs measures. Let Π_0 be any fully supported prior distribution on Θ.

The novel feature of the present setting is that we allow for general observations of the underlying family. Suppose that m is a σ-finite Borel measure on a complete separable metric space $\mathcal{U}$ and that $\varphi : \Theta \times \mathcal{X} \times \mathcal{U} \to [0, \infty)$ is a jointly continuous function such that for all $\theta \in \Theta$ and $x \in \mathcal{X}$,

$$\int \varphi_\theta(u|x)\, dm(u) = 1.$$

We regard $\{\varphi_\theta(\cdot|x) : \theta \in \Theta,\ x \in \mathcal{X}\}$ as a family of conditional likelihoods for $u \in \mathcal{U}$ given θ and x. We assume that the function $L : \Theta \times \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ given by $L(\theta, x, u) = -\log \varphi_\theta(u|x)$ satisfies the integrability and regularity conditions (i)–(iii) from Section 3. Furthermore, we require condition (L2) from [38], which stipulates that there exists $\alpha > 0$ and a Borel measurable function $C : \Theta \times \mathcal{U} \to [0, \infty)$ such that for each $(\theta, u) \in \Theta \times \mathcal{U}$, the function $L(\theta, \cdot, u) : \mathcal{X} \to \mathbb{R}$ is α-Hölder continuous with constant $C(\theta, u)$, and for each $\beta > 0$,

$$\sup_{(\theta,x)\in\Theta\times\mathcal{X}} \int \exp\bigl(\beta C(\theta, u)\bigr)\varphi_\theta(u|x)\, dm(u) < \infty.$$

This condition may be viewed as a condition on the regularity of the conditional density functions; it is used in [38] to control the likelihood function in the large deviations regime.

With these conditions in place, we assume that the conditional likelihood of observing $u_0^{n-1} \in \mathcal{U}^n$ given $(\theta, x) \in \Theta \times \mathcal{X}$ is

$$p_\theta\bigl(u_0^{n-1}|x\bigr) = \prod_{k=0}^{n-1} \varphi_\theta\bigl(u_k|S^k x\bigr),$$

and that the likelihood of observing $u_0^{n-1} \in \mathcal{U}^n$ given $\theta \in \Theta$ is

$$p_\theta\bigl(u_0^{n-1}\bigr) = \int p_\theta\bigl(u_0^{n-1}|x\bigr)\, d\mu_\theta(x).$$

In other words, for each $\theta \in \Theta$, we have an observed sequence $U_0, U_1, \ldots$ generated as follows: select $X \in \mathcal{X}$ according to μ_θ and for each $k \geq 0$ let $U_k \in \mathcal{U}$ have density $\varphi_\theta(\cdot|S^k X)$ with respect to m. Denote by $\mathbb{P}_\theta^U$ the process measure for the process $\{U_k\}$, which has likelihood p_θ.

Now let $\Pi_n(\cdot|u_0^{n-1})$ be the standard Bayesian posterior distribution on Θ given observations u_0^{n-1} based on the prior Π_0 and the likelihood p_θ: for Borel sets $E \subset \Theta$,

$$\Pi_n\bigl(E|u_0^{n-1}\bigr) = \frac{\int_E p_\theta(u_0^{n-1})\, d\Pi_0(\theta)}{\int_\Theta p_\theta(u_0^{n-1})\, d\Pi_0(\theta)}.$$

We consider the properly specified case, in which there exists a parameter $\theta^* \in \Theta$ such that the observed process $\{Y_n\}$ is drawn from $\mathbb{P}_{\theta^*}^U$. In order to address posterior consistency, we define the identifiability class of θ^*, denoted $[\theta^*]$, to be the set of $\theta \in \Theta$ such that $\mathbb{P}_\theta^U = \mathbb{P}_{\theta^*}^U$; in other words, a parameter is in $[\theta^*]$ if its associated process has the same distribution as the process generated by θ^*. The following result establishes posterior consistency in this setting.

THEOREM 3. *Let $E \subset \Theta$ be an open neighborhood of $[\theta^*]$. Then*

$$\lim_n \Pi_n\big(\Theta \setminus E | Y_0^{n-1}\big) = 0, \quad \mathbb{P}_{\theta^*}^U\text{-a.s.}$$

The proof of Theorem 3 is based on the principal results above. In particular, we use these results to establish convergence of the posterior distribution, and then we use problem-specific arguments to prove that the limit set $\Theta_{\min}$ is equal to the identifiability set $[\theta^*]$. These problem-specific arguments rely on previously studied connections between large deviations for Gibbs measures and identifiability of observed systems [38].

5.2. *Example*: *Ising model*, *Potts model*, *and nearest neighbor spin systems*. In this example we consider models from statistical physics that have been used for statistical inference in the presence of dependence and verify that they satisfy the hypotheses of our results. In the simplest version of these models the random variables $\ldots, X_{-1}, X_0, X_1, \ldots$ correspond to ordered sites in a one-dimensional lattice system. Each site can be in one of $q \geq 1$ states (or "spins") that we label $\mathcal{A} = \{1, \ldots, q\}$. The energy of the system and the probability of state configurations are determined by a function $h : \mathcal{A} \to \mathbb{R}$ that captures the baseline energy of individual states and a function $g : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ that captures the interactions between neighboring pairs of states. For $n \in \mathbb{N}$ the model assumes that the configuration $x_0^n = x_0 \ldots x_n \in \mathcal{A}^{n+1}$ appears with probability proportional to

$$\exp\left(\sum_{k=0}^{n-1} g(x_k, x_{k+1}) + \sum_{k=0}^{n} h(x_k)\right).$$

If $q \geq 2$, h is identically zero, and $g(a, b) = 2\beta \cdot \mathbf{1}(a = b)$, then this is the well-known Potts model with inverse temperature β. The special case $q = 2$ corresponds to the standard Ising model. Gibbs processes can be viewed as generalizations of such models in which nontrivial interactions between states may occur at arbitrary distances with appropriate decay in the strength of interaction as the distance increases [46].

Ising and Potts models have been used in statistical contexts [18, 44], where the underlying states x_0^n are considered hidden (or latent) variables, and the observations $y_0, \ldots, y_n \in \mathbb{R}$ are conditionally independent given these latent states. More precisely, the conditional likelihood of observing $y_0^n \in \mathbb{R}^{n+1}$ given $x_0^n \in \mathcal{A}^{n+1}$ is assumed to be

$$p\big(y_0^n | x_0^n\big) = \prod_{k=0}^{n} p(y_k | x_k),$$

where $p(\cdot|a)$ is a probability density for each hidden state $a \in \mathcal{A}$. Following [44], we suppose that for each hidden state $a \in \mathcal{A}$, the corresponding observation is conditionally normal with mean $m(a)$ and variance $\sigma(a)^2 > 0$, that is,

$$y_k | x_k \sim \mathcal{N}\big(m(x_k), \sigma(x_k)^2\big).$$

Such models have been used as dependent versions of finite mixture models.

As long as the dependence of h, g, m, and σ is continuous in the parametrization, these models satisfy all of the conditions for our consistency results. More precisely, we will establish the following result.

PROPOSITION 4. *Suppose that Θ is a compact metrizable space, and for each $a, b \in \mathcal{A}$, the functions $\theta \mapsto h_\theta(a)$, $\theta \mapsto g_\theta(a, b)$, $\theta \mapsto m_\theta(a)$, and $\theta \mapsto \sigma_\theta(a)$ are all continuous. Then the corresponding family of hidden Gibbs models satisfies all of the conditions stated in Section 5.1.*

PROOF. In this example, the state space $\mathcal{X}$ is simply $\mathcal{A}^{\mathbb{Z}}$ (since finite configurations have positive probability), which is trivially a mixing SFT. Now, for each parameter value $\theta \in \Theta$, define the potential $f_\theta : \mathcal{X} \to \mathbb{R}$ by setting $f_\theta(x) = g_\theta(x_0, x_1) + h_\theta(x_0)$. As these functions depend on only two coordinates (x_0 and x_1), they are all α-Hölder continuous on $\mathcal{X}$ for any $\alpha < 1$ (we fix one for the sake of concreteness). Hence, the latent process measure μ_θ satisfies the Gibbs property (1). Furthermore, the continuous dependence of g_θ and h_θ on θ imply that the family $\{f_\theta : \theta \in \Theta\}$ is a regular family (as in Definition 2).

As our conditional likelihood function is normal, the function $L : \Theta \times \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ is given by

$$L(\theta, x, y) = -\log p_\theta(y|x) = \log\bigl(\sigma_\theta(x_0)\bigr) + \frac{1}{2\sigma_\theta(x_0)^2}\bigl(m_\theta(x_0) - y\bigr)^2 + \log(\sqrt{2\pi}).$$

Using this explicit expression, the fact that the observations have finite second moments, and the continuous dependence of m_θ and σ_θ on θ, one may easily check that L satisfies conditions (i)–(iii). Additionally, for each $\theta \in \Theta$ and $y \in \mathbb{R}$, the function $L(\theta, \cdot, y) : \mathcal{X} \to \mathbb{R}$ depends only on one coordinate (x_0), and one may check that it is α-Hölder with $C(\theta, y) = C_0 + C_1|y|$ for uniform constants $C_0, C_1 > 0$. Then it satisfies condition (L2) from [38] (which one may check using the exponential moments of the normal distribution). Thus we have verified all of the conditions in Section 5.1. □

6. Joinings, divergence, and the rate function. In this section we discuss the rate function $V : \Theta \to \mathbb{R}$, whose existence is asserted by Theorem 1. In order to provide a thorough discussion, we first recall some background material from ergodic theory, including joinings and fiber entropy.

6.1. *Joinings.* Joinings were introduced by Furstenberg [16], and they have played an important role in the development of ergodic theory (see [10, 21]). Suppose $(\mathcal{U}_0, R_0, \eta_0)$ and $(\mathcal{U}_1, R_1, \eta_1)$ are two probability measure-preserving Borel systems with $R_i : \mathcal{U}_i \to \mathcal{U}_i$ and $\eta_i \in \mathcal{M}(\mathcal{U}_i, R_i)$. The product transformation $R_0 \times R_1 : \mathcal{U}_0 \times \mathcal{U}_1 \to \mathcal{U}_0 \times \mathcal{U}_1$ is defined by $(R_0 \times R_1)(u, v) = (R_0(u), R_1(v))$. A *joining* of these two systems is a Borel probability measure λ on $\mathcal{U}_0 \times \mathcal{U}_1$ with marginal distributions η_0 and η_1 that is invariant under the product transformation $R_0 \times R_1$. Thus, a joining is a coupling of the measures η_0 and η_1 that is also invariant under the joint action of the transformations R_0 and R_1; the former condition concerns the invariant measures of the two systems, while the latter concerns their dynamics. Let $\mathcal{J}(\eta_0, \eta_1)$ denote the set of all joinings of $(\mathcal{U}_0, R_0, \eta_0)$ and $(\mathcal{U}_1, R_1, \eta_1)$. Note that this set is nonempty, since the product measure $\eta_0 \otimes \eta_1$ is always a joining. When the transformation $R_0 : \mathcal{U}_0 \to \mathcal{U}_0$ is fixed but we have not associated any invariant measure with it, we set

$$\mathcal{J}(R_0 : \eta_1) = \bigcup_{\eta_0 \in \mathcal{M}(\mathcal{U}_0, R_0)} \mathcal{J}(\eta_0, \eta_1),$$

which is the family of joinings of $(\mathcal{U}_1, R_1, \eta_1)$ with all systems of the form $(\mathcal{U}_0, R_0, \eta_0)$, with $\eta_0 \in \mathcal{M}(\mathcal{U}_0, R_0)$.

6.2. *Entropy.* Our statements and proofs also require us to introduce some notions from the entropy theory of dynamical systems. Let $\mathcal{U}$ be a compact metric space, $R : \mathcal{U} \to \mathcal{U}$ continuous, and $\eta \in \mathcal{M}(\mathcal{U}, R)$. For any finite measurable partition α of $\mathcal{U}$, we define

$$H(\eta, \alpha) = -\sum_{C \in \alpha} \eta(C) \log \eta(C),$$

where $0 \cdot \log 0 = 0$ by convention. For $k \geq 0$, let $R^{-k}\alpha = \{R^{-k}A : A \in \alpha\}$, and for any partitions $\alpha^0, \ldots, \alpha^n$, define their join to be the mutual refinement

$$\bigvee_{k=0}^{n} \alpha^k = \{A_0 \cap \cdots \cap A_n : A_i \in \alpha^i\}.$$

For $n \geq 0$, let $\alpha_n = \bigvee_{k=0}^{n-1} R^{-k}\alpha$. By standard subadditivity arguments, the following limit exists:

$$h_R(\eta, \alpha) := \lim_n \frac{1}{n} H(\eta, \alpha_n) = \inf_n \frac{1}{n} H(\eta, \alpha_n).$$

The measure-theoretic or Kolmogorov–Sinai entropy of $(\mathcal{U}, R)$ with respect to η is given by $h_R(\eta) = \sup_\alpha h_R(\eta, \alpha)$, where the supremum is taken over all finite measurable partitions α of $\mathcal{U}$. We note for future reference that for any $\epsilon > 0$, the value $h_R(\eta)$ remains the same if the supremum is instead taken over all finite measurable partitions with diameter less than ϵ. When the transformation R is clear from context, we may omit the subscript.

6.3. *The variational principle for pressure.* Let $\mathcal{X}$ be a mixing SFT, and let $f : \mathcal{X} \to \mathbb{R}$ be a Hölder continuous potential. The variational principle [5] for the pressure $\mathcal{P}(f)$ states that

$$\mathcal{P}(f) = \sup\left\{\int f\, d\mu + h(\mu) : \mu \in \mathcal{M}(\mathcal{X}, S)\right\}, \tag{7}$$

and furthermore, the supremum is achieved by the measure $\mu \in \mathcal{M}(\mathcal{X}, S)$ if and only if μ is the Gibbs measures associated with f.

6.4. *Disintegration of measure.* The following result is a special case of standard results on disintegration of Borel measures (see [21]).

THEOREM ([21]). *Let $\mathcal{U}$ and $\mathcal{Y}$ be standard Borel spaces, and $\phi : \mathcal{U} \times \mathcal{Y} \to \mathcal{Y}$ be the natural projection. Let $\lambda \in \mathcal{M}(\mathcal{U} \times \mathcal{Y})$, and let $\nu = \lambda \circ \phi^{-1}$ be its image in $\mathcal{M}(\mathcal{Y})$. Then there is a Borel map $y \mapsto \lambda_y$, from $\mathcal{Y}$ to $\mathcal{M}(\mathcal{U})$ such that for every bounded Borel function $f : \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$,*

$$\int f d\lambda = \int \left(\int f\, d[\lambda_y \otimes \delta_y]\right) d\nu(y).$$

Moreover, such a map is unique in the following sense: if $y \mapsto \lambda'_y$ is another such map, then $\lambda_y = \lambda'_y$ for ν-almost every y.

Note that if λ is a joining, then the family $\{\lambda_y\}_{y \in \mathcal{Y}}$ satisfies an important invariance property, which we state as Lemma 10 in Appendix A.3.

6.5. *Fiber entropy.* Now we give a definition of fiber entropy, along with statements of some properties relevant to this work; for a thorough introduction, see [29]. Let $\mathcal{U}$ be a compact metric space and $\mathcal{Y}$ be a separable complete metric space. Further, let $R : \mathcal{U} \to \mathcal{U}$ be continuous and $T : \mathcal{Y} \to \mathcal{Y}$ be Borel measurable. For any Borel probability measure λ on $\mathcal{U} \times \mathcal{Y}$ with $\mathcal{Y}$-marginal ν, let $\lambda = \int \lambda_y \otimes \delta_y\, d\nu(y)$ be its disintegration over $\mathcal{Y}$. Then for any finite measurable partition α of $\mathcal{U}$, we define

$$H(\lambda, \alpha|\mathcal{Y}) = \int H(\lambda_y, \alpha)\, d\nu(y). \tag{8}$$

Now suppose $\nu \in \mathcal{M}(\mathcal{Y}, T)$. It's possible to show (see, e.g., [28]) that if $\lambda \in \mathcal{J}(R : \nu)$ and $\lambda = \int \lambda_y \otimes \delta_y \, d\nu(y)$ is its disintegration over ν, then for every finite measurable partition α of $\mathcal{U}$ the following limit exists:

$$h^{\nu}(\lambda, \alpha) := \lim_n \frac{1}{n} H(\lambda, \alpha_n | \mathcal{Y}) = \inf_n \frac{1}{n} H(\lambda, \alpha_n | \mathcal{Y}) \, d\nu(y),$$

where $\alpha_n = \bigvee_{k=0}^{n-1} R^{-k}\alpha$. Furthermore, when λ is ergodic, it can be shown (again see [28]) that for ν-almost every y,

$$h^{\nu}(\lambda, \alpha) = \lim_n \frac{1}{n} H(\lambda_y, \alpha_n).$$

The *fiber entropy* of λ over ν is defined as $h^{\nu}(\lambda) = \sup_{\alpha} h^{\nu}(\lambda, \alpha)$, where the supremum is taken over all finite measurable partitions α of $\mathcal{U}$. Note that the supremum may also be taken over partitions with diameter less than any $\epsilon > 0$. The fiber entropy $h^{\nu}(\lambda)$ quantifies the relative entropy of λ over ν.

6.6. *Divergence terms.* Consider a parameter $\theta \in \Theta$ and a joining $\lambda \in \mathcal{J}(S : \nu)$. We would like to quantify the divergence of the joining λ to the product measure $\mu_\theta \otimes \nu$, as it will play a role in the rate function V. (Note that the measure $\mu_\theta \otimes \nu$ may be interpreted as a prior distribution on $\mathcal{X} \times \mathcal{Y}$ given θ, as the prior on $\mathcal{X}$ is assumed to be independent of the observations.) However, the standard KL-divergence is insufficient for our purposes, since any two ergodic measures for a given system are known to be mutually singular, and hence their KL-divergence will be infinite. Instead, we make the following definitions, which are more suitable for dynamical systems.

Given two Borel probability measures η and γ on a compact metric space $\mathcal{U}$ and a finite measurable partition α of $\mathcal{U}$, we write $\eta \prec_\alpha \gamma$ whenever $\gamma(C) = 0$ implies that $\eta(C) = 0$ for $C \in \alpha$. Let

$$KL(\eta : \gamma | \alpha) = \begin{cases} \displaystyle\sum_{C \in \alpha} \eta(C) \log \frac{\eta(C)}{\gamma(C)}, & \text{if } \eta \prec_\alpha \gamma \\ +\infty, & \text{otherwise,} \end{cases}$$

where $0 \cdot \log \frac{0}{x} = 0$ for any x by convention. Note that $KL(\eta : \gamma | \alpha)$ is the KL-divergence from γ to η with respect to the partition α, which is nonnegative.

Now consider a Hölder continuous potential $f : \mathcal{X} \to \mathbb{R}$ on a mixing SFT $\mathcal{X}$ with associated Gibbs measure $\mu \in \mathcal{M}(\mathcal{X}, S)$. Let α be the partition of $\mathcal{X}$ into cylinder sets of the form $x[0]$ for some $x \in \mathcal{X}$, and let $\eta \in \mathcal{M}(\mathcal{X}, S)$ be ergodic. In this situation, it is known [7] that

$$\lim_n \frac{1}{n} KL(\eta : \mu | \alpha_n) = \mathcal{P}(f) - \left(h(\eta) + \int f \, d\eta \right),$$

where we recall that $\mathcal{P}(f)$ is the pressure of f, the partition α_n is defined to be $\bigvee_{k=0}^{n-1} S^{-k}\alpha$, and $h(\eta)$ is the entropy of η with respect to S. Next we generalize this result to handle the relative situation, which involves joinings and relative entropy.

LEMMA 1. *Let $f : \mathcal{X} \to \mathbb{R}$ be a Hölder continuous potential on a mixing SFT $\mathcal{X}$ with associated Gibbs measure μ. Let α be the partition of $\mathcal{X}$ into cylinder sets of the form $x[0]$, and let $\lambda \in \mathcal{J}(S : \nu)$ be ergodic. Then for ν-almost every $y \in \mathcal{Y}$,*

$$\lim_n \frac{1}{n} KL(\lambda_y : \mu | \alpha_n) = \mathcal{P}(f) - \left(h^{\nu}(\lambda) + \int f \, d\lambda \right).$$

We defer the proof of Lemma 1 to Appendix A.6. Based on this lemma, we make the following definition.

DEFINITION 3. Let $\mathcal{X}$ be a mixing SFT, $f : \mathcal{X} \to \mathbb{R}$ a Hölder continuous function, and μ the associated Gibbs measure. Further, let $(\mathcal{Y}, T, \nu)$ be an ergodic system. Then define the relative divergence rate of $\lambda \in \mathcal{J}(S : \nu)$ to μ to be

$$D(\lambda : \mu) = \mathcal{P}(f) - \left(h^{\nu}(\lambda) + \int f(x)\, d\lambda(x, y) \right). \tag{9}$$

In the present setting, $D(\lambda : \mu)$ is always finite, and one may check that it is also nonnegative (see Lemma 15 in Appendix A.6).

6.7. *The rate function.* In this section we define and discuss the rate function $V : \Theta \to \mathbb{R}$ whose existence is guaranteed by Theorem 1.

DEFINITION 4. For $\theta \in \Theta$, let

$$V(\theta) = \inf \left\{ \int \ell\, d\lambda + D(\lambda : \mu_\theta) : \lambda \in \mathcal{J}(S : \nu) \right\}.$$

Note that the variational expression defining V contains the sum of an expected loss term and a divergence term. It is known that Bayesian posterior distributions satisfy a similar variational principle in the finite sample setting (see [27, 60, 61]). Our results show that this interpretation passes to the limit as the number of samples tends to infinity.

By Proposition 7, which appears in Appendix A.7, we have that V is lower semi-continuous. Since the loss function is continuous, the proof of Proposition 7 essentially follows from the upper semi-continuity of the fiber entropy on the space of joinings $\mathcal{J}(S : \nu)$.

REMARK 10. Consider the introduction of an inverse temperature parameter $\beta \in \mathbb{R}$, as discussed in Remark 5, and let $\ell_\beta = \beta \cdot \ell$ be the associated loss function. If we let V_β be the associated rate function, then we see from Definition 4 that

$$V_\beta(\theta) = \inf \left\{ \beta \cdot \int \ell\, d\lambda + D(\lambda : \mu_\theta) : \lambda \in \mathcal{J}(S : \nu) \right\}.$$

Dividing by β and letting β tend to infinity to investigate the ground state behavior, it is clear that the associated variational expression is

$$V_\infty(\theta) := \lim_{\beta \to \infty} \frac{V_\beta(\theta)}{\beta} = \inf \left\{ \int \ell\, d\lambda : \lambda \in \mathcal{J}(S : \nu) \right\}.$$

Interestingly, this variational expression has been studied recently as part of an asymptotic analysis of estimators based on empirical risk minimization for dynamical systems [40, 41]. Indeed, the solution set Θ_∞ of this ground state variational problem exactly characterizes the set of possible limits of parameter estimates that asymptotically minimize average empirical risk.

7. A technical preliminary. Define

$$L(\eta : \gamma | \alpha) = \begin{cases} \sum_{C \in \alpha} \eta(C) \log \gamma(C), & \text{if } \eta \prec_\alpha \gamma, \\ -\infty, & \text{otherwise,} \end{cases}$$

where $0 \cdot \log 0 = 0$ by convention. With these definitions, we always have

$$-KL(\eta : \gamma|\alpha) = H(\eta, \alpha) + L(\eta : \gamma|\alpha). \tag{10}$$

We now establish a lemma that is used in the proof of Theorem 1. This result allows us to approximate the expected information in the prior P_0, where the expectation is with respect to an arbitrary measure, in terms of an average of a continuous function. These types of estimates are available precisely because our model class consists of Gibbs measures: indeed, they do not hold for arbitrary invariant measures for dynamical systems.

For any Borel probability measure η on $\Theta \times \mathcal{X}$, let η_n denote its time-average up to time n:

$$\eta_n(E) = \frac{1}{n} \sum_{k=0}^{n-1} \eta\big((I_\Theta \times S)^{-k} E\big),$$

where $I_\Theta : \Theta \to \Theta$ is the identity.

LEMMA 2. *Let K be the constant in the uniform Gibbs property (2). For any $\epsilon > 0$ there exists $\delta > 0$ such that if the diameter of α is less than δ and the prior π_0 assigns positive measure to each element of α, and if β is the partition of $\mathcal{X}$ into cylinder sets of the form $x[0]$, then for any Borel probability measure η on $\Theta \times \mathcal{X}$, and any $n \geq 0$,*

$$\left| \frac{1}{n} L(\eta : P_0|\alpha \times \beta_n) - \frac{1}{n} \int \big(f_\theta(x) - \mathcal{P}(f_\theta)\big)\, d\eta_n(\theta, x) \right| \leq \epsilon + \frac{\log K}{n}.$$

PROOF. Let $\epsilon > 0$. By the uniform continuity of f_θ and $\mathcal{P}(f_\theta)$ in θ and the uniform Gibbs property, there exists $\delta > 0$ such that if the diameter of α is less than δ and β is the partition of $\mathcal{X}$ into sets of the form $x[0]$, then for all $\theta \in \Theta$, $x \in \mathcal{X}$, and $n \geq 1$,

$$K^{-1} \exp(-\epsilon n) \leq \frac{P_0((\alpha \times \beta_n)(\theta, x))}{\exp(-n\,\mathcal{P}(f_\theta) + \sum_{k=0}^{n-1} f_\theta(S^k x))} \leq K \exp(\epsilon n).$$

Taking logarithms and dividing by n, we obtain the inequality

$$\left| \frac{1}{n} \log P_0\big((\alpha \times \beta_n)(\theta, x)\big) - \frac{1}{n} \sum_{k=0}^{n-1} f_\theta\big(S^k x\big) + \mathcal{P}(f_\theta) \right| \leq \epsilon + \frac{\log K}{n},$$

which is uniform over $(\theta, x) \in \Theta \times \mathcal{X}$. Now let η be any Borel probability measure on $\Theta \times \mathcal{X}$. Then by integrating with respect to η, we see that

$$\left| \frac{1}{n} L(\eta : P_0|\alpha \times \beta_n) - \frac{1}{n} \int \big(f_\theta(x) - \mathcal{P}(f_\theta)\big)\, d\eta_n(\theta, x) \right| \leq \epsilon + \frac{\log K}{n}. \qquad \square$$

8. Proof of Theorem 1. In this section, we prove Theorem 1, which concerns the convergence of the average log normalizing constant (partition function) $n^{-1} \log Z_n$. In an attempt to keep the paper mostly self-contained, we have included an appendix (Appendix A) containing several routine technical results that we use in the proof. The starting point of the proof, which is an application of the Pressure Lemma (in Appendix A.1), allows us to express the main statistical object, the Gibbs posterior distribution, as the solution of a variational problem involving information theoretic notions such as entropy and average information, which have long been studied in dynamics. The proof of Theorem 1 follows.

To ease notation slightly in this section, we let $g = -\ell$ and $g_n = -\ell_n$, where ℓ_n is defined in (3). We also set $\mathcal{U} = \Theta \times \mathcal{X}$ and $R(\theta, x) = (\theta, S(x))$. For $\lambda \in \mathcal{J}(R : \nu)$, we will have use for the notation

$$G(\lambda) = \int \big(\mathcal{P}(f_\theta) - f_\theta(x)\big)\, d\lambda(\theta, x, y) - h^\nu(\lambda).$$

Although we do not use this fact, we note that $G(\lambda)$ can be written as an integral over θ of terms of the form $D(\lambda_\theta, \mu_\theta)$ (as in Definition 3). Lemma 12 (in Appendix A.5) ensures that $h^\nu(\cdot)$ is harmonic (see Appendix A.2 for precise definition), and therefore the same is true of $G : \mathcal{J}(R : \nu) \to \mathbb{R}$. In this notation, our goal is to prove

$$\lim_n \frac{1}{n} \log Z_n = \sup\left\{ \int g \, d\lambda - G(\lambda) : \lambda \in \mathcal{J}(R : \nu) \right\}.$$

We present the proof in two stages: first we establish that the expression in the right-hand side is a lower bound for $\lim_n n^{-1} \log Z_n$, and then we prove that the same expression provides an upper bound.

8.1. *Lower bound.* The goal of this section is to prove the following result.

PROPOSITION 5. *For ν-almost every $y \in \mathcal{Y}$,*

$$\lim_n \frac{1}{n} \log Z_n \geq \sup\left\{ \int g \, d\lambda - G(\lambda) : \lambda \in \mathcal{J}(R : \nu) \right\},$$

where $Z_n = Z_n(y)$.

Before proving this proposition, we first establish a lemma. If η is a Borel probability measure on $\Theta \times \mathcal{X}$ and $\eta(C) > 0$, then let η_C denote the conditional distribution $\eta(\cdot|C)$. Also, we say that β is a partition of $\mathcal{X}$ according to central words whenever $\beta = \{[x_{-m}^m] : x \in \mathcal{X}\}$ for some $m \geq 0$.

LEMMA 3. *Let α be a finite measurable partition of Θ with* $\operatorname{diam}(\alpha) < \delta$, *and let β be a partition of $\mathcal{X}$ according to central words such that* $\operatorname{diam}(\beta) < \delta$. *Then for any Borel probability measure η on $\Theta \times \mathcal{X}$, any $y \in \mathcal{Y}$, and any $n \geq 1$,*

$$\int g_n(\theta, x, y) \, d\eta(\theta, x) - KL(\eta : P_0 | \alpha \times \beta_n)$$

$$\leq \log\left[\int \exp\bigl(g_n(\theta, x, y)\bigr) \, dP_0(\theta, x) \right] + \sum_{k=0}^{n-1} \rho_\delta\bigl(T^k y\bigr),$$

where ρ_δ is the local difference function appearing in property (iii) of the loss.

PROOF. If $\eta \not\prec_{\alpha \times \beta_n} P_0$, then the inequality holds trivially. Now suppose $\eta \prec_{\alpha \times \beta_n} P_0$, and let $\xi = \{C \in \alpha \times \beta_n : \eta(C) > 0\}$. For $C \in \xi$ and $(\theta, x), (\theta', x') \in C$, property (iii) of the loss function, and our hypotheses on α and β yield that

$$g_n\bigl(\theta', x', y\bigr) \leq g_n(\theta, x, y) + \sum_{k=0}^{n-1} \rho_\delta\bigl(T^k y\bigr).$$

Integrating out (θ', x') with respect to the conditional distribution η_C gives

$$\int_C g_n\bigl(\theta', x', y\bigr) \, d\eta_C\bigl(\theta', x'\bigr) \leq g_n(\theta, x, y) + \sum_{k=0}^{n-1} \rho_\delta\bigl(T^k y\bigr).$$

After exponentiation and integration with respect to the $P_{0,C}$, we get

$$\exp\left(\int_C g_n\bigl(\theta', x', y\bigr) \, d\eta_C\bigl(\theta', x'\bigr) \right)$$

$$\leq \exp\left(\sum_{k=0}^{n-1} \rho_\delta\bigl(T^k y\bigr) \right) \int_C \exp\bigl(g_n(\theta, x, y)\bigr) \, dP_{0,C}(\theta, x).$$

Invoking the Pressure Lemma (Lemma 9 in Appendix A.1) and the inequality above, we find that

$$\begin{aligned}
&\int g_n(\theta', x', y)\, d\eta(\theta', x') - KL(\eta : P_0|\alpha \times \beta_n) \\
&\quad = \sum_{C\in\xi} \eta(C)\Big[\log P_0(C) + \int_C g_n(\theta', x', y)\, d\eta_C(\theta', x') - \log \eta(C)\Big] \\
&\quad \le \log \sum_{C\in\xi} \exp\Big(\log P_0(C) + \int_C g_n(\theta', x', y)\, d\eta_C(\theta', x')\Big) \\
&\quad\quad = \log \sum_{C\in\xi} \exp\Big(\int_C g_n(\theta', x', y)\, d\eta_C(\theta', x')\Big) P_0(C) \\
&\quad \le \log \sum_{C\in\xi} \Big(\int_C \exp\big(g_n(\theta, x, y)\big)\, dP_{0,C}(\theta, x)\Big) P_0(C) + \sum_{k=0}^{n-1} \rho_\delta(T^k y) \\
&\quad = \log \int \exp\big(g_n(\theta, x, y)\big)\, dP_0(\theta, x) + \sum_{k=0}^{n-1} \rho_\delta(T^k y),
\end{aligned}$$

as was to be shown. □

PROOF OF PROPOSITION 5. Fix an ergodic joining $\lambda \in \mathcal{J}(R : \nu)$ and $\epsilon > 0$. Let $\delta > 0$ be sufficiently small that the bound of Lemma 2 holds and that $\int \rho_\delta\, d\nu < \epsilon$ (using property (iii) of the loss). Fix a finite measurable partition α of Θ such that $\operatorname{diam}(\alpha) < \delta$ and such that the prior π_0 assigns positive measure to all elements of α (which can be done since Θ is compact and metrizable and π_0 is fully supported), and select m large enough so that the partition β of $\mathcal{X}$ generated by central words of length m satisfies $\operatorname{diam}(\beta) < \delta$. Then for ν-almost every y,

$$\begin{aligned}
&H(\lambda_y, \alpha \times \beta_n) + L(\lambda_y : P_0|\alpha \times \beta_n) + \int g_n\, d\lambda_y \\
&\quad = \int g_n\, d\lambda_y - KL(\lambda_y : P_0|\alpha \times \beta_n) \\
&\quad \le \log \int \exp\big(g_n(\theta, x, y)\big)\, dP_0(\theta, x) + \sum_{k=0}^{n-1} \rho_\delta(T^k y),
\end{aligned}$$

where the inequality follows from Lemma 3. Dividing each side of the inequality above by n, and then letting n tend to infinity, we have that for ν-almost every $y \in \mathcal{Y}$,

$$h^\nu(\lambda, \alpha \times \beta) + \int \big(f_\theta(x) - \mathcal{P}(f_\theta)\big)\, d\lambda + \int g\, d\lambda \le \liminf_n \frac{1}{n} Z_n(y) + 2\epsilon.$$

(The limit of the entropic term is part of the definition of fiber entropy. The limit of the middle term is obtained by an application of Lemma 2. The limit of the average loss term is a consequence of the invariance property of the decomposition of λ, stated formally as Lemma 11 in Appendix A.4, and the limit of the term containing ρ_δ is given by the ergodic theorem, using hypothesis (iii) on the loss function.) Taking the supremum over all partitions α of Θ with diameter less than δ and all partitions β of $\mathcal{X}$ generated by central words of length at least m, we obtain the inequality

$$\int g\, d\lambda - G(\lambda) \le \liminf_n \frac{1}{n} \log Z_n(y) + 2\epsilon.$$

Since $\epsilon > 0$ was arbitrary,

$$\int g\,d\lambda - G(\lambda) \le \liminf_n \frac{1}{n}\log Z_n. \tag{11}$$

Notice that the left-hand side is harmonic in λ (see Appendix A.2 for details), and therefore

$$\sup_{\lambda\in\mathcal{J}(R:\nu)} \int g\,d\lambda - G(\lambda) = \sup_{\substack{\lambda\in\mathcal{J}(R:\nu)\\ \lambda \text{ ergodic}}} \int g\,d\lambda - G(\lambda).$$

Hence there exists a sequence $\{\lambda_k\}_{k=1}^{\infty}$ of ergodic joinings in $\mathcal{J}(R:\nu)$ that achieves the supremum. For each $k \ge 1$, we have shown that there is a measurable set $E_k \subset \mathcal{Y}$ such that $\nu(E_k) = 1$ and equation (11) holds with $\lambda = \lambda_k$ for all $y \in E_k$. Then for all $y \in \bigcap_k E_k$, we have that

$$\sup_{\lambda\in\mathcal{J}(R:\nu)} \left\{\int g\,d\lambda - G(\lambda)\right\} \le \liminf_n \frac{1}{n}\log Z_n.$$

Since $\nu(\bigcap_k E_k) = 1$, this completes the proof. □

8.2. *Upper bound.* In Proposition 6 below we establish an almost sure upper bound on the limiting behavior of $n^{-1}\log Z_n(y)$. Together with the lower bound in Proposition 5, this completes the proof of Theorem 1.

PROPOSITION 6. *For ν-almost every $y \in \mathcal{Y}$,*

$$\limsup_n \frac{1}{n}\log Z_n(y) \le \sup_{\lambda\in\mathcal{J}(R:\nu)} \left\{\int g\,d\lambda - G(\lambda)\right\}.$$

We begin with a preliminary lemma. Recall that P_0 is the prior distribution on $\Theta\times\mathcal{X}$ generated by the prior π_0 (defined in (4)) and the family $\{\mu_\theta : \theta\in\Theta\}$, while $P_n(\cdot|y)$ is the Gibbs posterior distribution associated with $y, Ty, \ldots, T^{n-1}y$ (defined in (5)). To simplify notation, in what follows $P_n(\cdot|y)$ is denoted by P_n^y.

LEMMA 4. *If α is a finite measurable partition of $\Theta\times\mathcal{X}$ with diameter less than δ, then for $y\in\mathcal{Y}$ and $n\ge 1$,*

$$\log \int \exp\bigl(g_n(\theta,x,y)\bigr)\,dP_0(\theta,x)$$

$$\le H\bigl(P_n^y,\alpha_n\bigr) + L\bigl(P_n^y : P_0|\alpha_n\bigr) + \int g_n(\theta,x,y)\,dP_n^y(\theta,x) + \sum_{k=0}^{n-1}\rho_\delta\bigl(T^k y\bigr).$$

PROOF. Let α be a finite measurable partition of $\Theta\times\mathcal{X}$ with $\operatorname{diam}(\alpha) < \delta$, and let $y\in\mathcal{Y}$. By definition P_n^y and P_0 are equivalent measures, and hence $P_n^y \prec_{\alpha_n} P_0$ and $P_0 \prec_{\alpha_n} P_n^y$. Let $\xi = \{C\in\alpha_n : P_0(C) > 0\} = \{C\in\alpha_n : P_n^y(C) > 0\}$.

Fix $C\in\xi$ for the moment. For points $(\theta,x),(\theta',x')\in C$ the hypothesis on α ensures that

$$g_n(\theta,x,y) \le g_n\bigl(\theta',x',y\bigr) + \sum_{k=0}^{n-1}\rho_\delta\bigl(T^k y\bigr),$$

where ρ_δ is defined in condition (iii) of the loss. Exponentiating both sides of the inequality and integrating (θ,x) with respect to the prior $P_{0,C}$ conditioned on being in C yields

$$\int_C \exp\bigl(g_n(\theta,x,y)\bigr)\,dP_{0,C} \le \exp\bigl(g_n\bigl(\theta',x',y\bigr)\bigr)\exp\left(\sum_{k=0}^{n-1}\rho_\delta\bigl(T^k y\bigr)\right).$$

Taking logarithms and integrating (θ', x') with respect to the posterior $P^y_{n,C}$ conditioned on being in C yields

$$\log \int_C \exp\bigl(g_n(\theta, x, y)\bigr)\, dP_{0,C} \le \int_C g_n(\theta', x', y)\, dP^y_{n,C} + \sum_{k=0}^{n-1} \rho_\delta(T^k y). \tag{12}$$

By the definition of P_n and the pressure lemma (Lemma 9 in Appendix A.1) we have

$$\begin{aligned}
&\log \int \exp\bigl(g_n(\theta, x, y)\bigr)\, dP_0 \\
&\quad = \log \sum_{C\in\xi} \exp\biggl[\log \int_C \exp\bigl(g_n(\theta, x, y)\bigr)\, dP_0\biggr] \\
&\quad = \sum_{C\in\xi} P^y_n(C)\biggl[-\log P^y_n(C) + \log P_0(C) + \log \int_C \exp\bigl(g_n(\theta, x, y)\bigr)\, dP_{0,C}\biggr] \\
&\quad = H\bigl(P^y_n, \alpha_n\bigr) + L\bigl(P^y_n : P_0|\alpha_n\bigr) + \sum_{C\in\xi} P^y_n(C) \log \int_C \exp\bigl(g_n(\theta, x, y)\bigr)\, dP_{0,C}.
\end{aligned}$$

Applying inequality (12) to the terms of the final sum above, we see that

$$\begin{aligned}
&\log \int \exp\bigl(g_n(\theta, x, y)\bigr)\, dP_0(\theta, x) \\
&\quad \le H\bigl(P^y_n, \alpha_n\bigr) + L\bigl(P^y_n : P_0|\alpha_n\bigr) + \int g_n(\theta, x, y)\, dP^y_n(\theta, x) + \sum_{k=0}^{n-1} \rho_\delta(T^k y)
\end{aligned}$$

as desired. □

PROOF OF PROPOSITION 6. First, let

$$\nu_n = \frac{1}{n} \sum_{k=0}^{n-1} \delta_{y_k}.$$

Additionally, for $y \in \mathcal{Y}$, we define

$$\eta^y_n = \frac{1}{n} \sum_{k=0}^{n-1} \bigl(P^y_n \circ R^{-k}\bigr) \otimes \delta_{y_k}.$$

By [28], Lemma 2.1, for ν-almost every y, the sequence $\{\eta^y_n\}_n$ is tight and all of its limit points are contained in $\mathcal{J}(R : \nu)$. Using this fact and the ergodic theorem, there exists a set E such that $\nu(E) = 1$ and the following relations hold:

- the sequence $\{\eta^y_n\}_n$ is tight and all of its limit points are in $\mathcal{J}(R : \nu)$;
- $\{\nu_n\}_{n=1}^\infty$ converges weakly to ν;
- for each $m \ge 1$, we have $\int_{\ell^*>m} \ell^*\, d\nu_n \to \int_{\ell^*>m} \ell^*\, d\nu$.

We refer to elements of E as *generic* points. Fix a generic point $y \in E$. Let λ and $\{n_k\}$ be such that $\eta^y_{n_k} \to \lambda \in \mathcal{J}(R : \nu)$.

Let $\epsilon > 0$, and choose $\delta > 0$ such that $\int \rho_\delta\, d\nu < \epsilon$. Choose a finite measurable partition α of $\Theta \times \mathcal{X}$ such that $\operatorname{diam}(\alpha) < \delta$, the prior π_0 assigns positive measure to each element of α, and $\operatorname{proj}_{\Theta\times\mathcal{X}}(\lambda)(\partial\alpha) = 0$ (which exists since $\Theta \times \mathcal{X}$ is compact [53], Lemma 8.5, and

π_0 is fully supported). By adapting an argument from [53], p. 190, involving subadditivity of measure-theoretic entropy, we obtain that for each $q \geq 1$, for $n \geq q$,

$$\begin{aligned}\frac{1}{n} H\big(P_n^y, \alpha_n\big) &\leq \frac{1}{n} \sum_{k=0}^{n-1} \frac{1}{q} H\big(P_n^y \circ R^{-k}, \alpha_q\big) + o(1) \\ &= \frac{1}{q} H\big(\eta_n^y, \alpha_q | \mathcal{Y}\big) + o(1),\end{aligned}$$

where $o(1)$ refers to a term that tends to 0 as n tends to infinity (for fixed q). Then by letting n tend to infinity and applying [28], Lemma 2.1, again, we see that

$$\limsup_n \frac{1}{n} H\big(P_n^y, \alpha_n\big) \leq \frac{1}{q} H(\lambda, \alpha_q | \mathcal{Y}),$$

where the conditional entropy $H(\cdot|\mathcal{Y})$ is defined in (8). To proceed with the proof, we require the following lemma. Recall that at the beginning of this section, we set $g = -\ell$ and $g_n = -\ell_n$.

LEMMA 5. *Let y, $\{n_k\}$, and λ be as above. Then*

$$\lim_k \frac{1}{n_k} \int g_{n_k}(\theta, x, y)\, dP_{n_k}^y(\theta, x) = \int g\, d\lambda. \tag{13}$$

PROOF. Let $\eta_n = \eta_n^y$. By definition of η_n,

$$\begin{aligned}\frac{1}{n} \int g_n(\theta, x, y)\, dP_n^y(\theta, x) &= \frac{1}{n} \int \sum_{k=0}^{n-1} g\big(\theta, S^k x, y_k\big)\, dP_n^y(\theta, x) \\ &= \int g\, d\eta_n.\end{aligned}$$

Note that g is continuous, since we have assumed that ℓ is continuous. If g were bounded, then the desired limit would follow directly from the Portmanteau theorem for weak convergence. In general we rely on a truncation argument; although the details are routine, we include them here for completeness.

For $m \in \mathbb{N}$, define the truncated function

$$g_m(x, y) = \begin{cases} g(\theta, x, y), & \text{if } \big|g(\theta, x, y)\big| \leq m, \\ m, & \text{if } g(\theta, x, y) \geq m, \\ -m, & \text{if } g(\theta, x, y) \leq -m. \end{cases}$$

Note that $|g_m| \leq |g|$ and that $g_m \to g$ as m tends to infinity. The integrability of g with respect to λ follows from that of ℓ^* with respect to ν, and the dominated convergence theorem then ensures that $\int g_m\, d\lambda \to \int g\, d\lambda$. Moreover, with ν_n defined as above, it follows from the choice of y that

$$\limsup_n \int |g - g_m|\, d\eta_n \leq \limsup_n \int_{\ell^* > m} \ell^*\, d\nu_n = \int_{\ell^* > m} \ell^*\, d\nu.$$

In order to establish (13), let $\epsilon > 0$ be fixed. By virtue of the results in the previous paragraph, there exist integers m and n_1 sufficiently large that for each $n \geq n_1$

$$\left| \int g_m\, d\lambda - \int g\, d\lambda \right| < \epsilon/3 \quad \text{and} \quad \int |g - g_m|\, d\eta_n < \epsilon/3.$$

Moreover, as $\eta_n \Rightarrow \lambda$ and g_m is continuous and bounded, there exists $n_2 \geq n_1$ such that for each $n \geq n_2$,

$$\left| \int g_m \, d\lambda - \int g_m \, d\eta_n \right| < \epsilon/3.$$

Combining the inequalities above, a straightforward bound shows that

$$\left| \int g \, d\lambda - \int g \, d\eta_n \right| < \epsilon$$

for $n > n_2$. As $\epsilon > 0$ was arbitrary, we conclude that (13) holds. $\square$

Combining Lemma 5 with Lemmas 2 and 4, we find that for ν-almost every $y \in \mathcal{Y}$

$$\begin{aligned} &\limsup_n \frac{1}{n} \log \int \exp\bigl(g_n(\theta, x, y)\bigr) \, dP_0(\theta, x) \\ &\quad \leq \frac{1}{q} H(\lambda, \alpha_q | \mathcal{Y}) + \int \bigl(f_\theta(x) - \mathcal{P}(f_\theta)\bigr) \, d\lambda(\theta, x, y) + \int g \, d\lambda + 2\epsilon. \end{aligned}$$

Letting q tend to infinity, we get

$$\begin{aligned} &\limsup_n \frac{1}{n} \log \int \exp\bigl(g_n(\theta, x, y)\bigr) \, dP_0(\theta, x) \\ &\quad \leq h^\nu(\lambda, \alpha) + \int \bigl(f_\theta(x) - \mathcal{P}(f_\theta)\bigr) \, d\lambda(\theta, x, y) + \int g \, d\lambda + 2\epsilon \\ &\quad \leq h^\nu(\lambda) + \int \bigl(f_\theta(x) - \mathcal{P}(f_\theta)\bigr) \, d\lambda(\theta, x, y) + \int g \, d\lambda + 2\epsilon. \end{aligned}$$

Since ϵ was arbitrary, we obtain

$$\begin{aligned} &\limsup_n \frac{1}{n} \log \int \exp\bigl(g_n(\theta, x, y)\bigr) \, dP_0(\theta, x) \\ &\quad \leq \int g \, d\lambda - G(\lambda) \\ &\quad \leq \sup \left\{ \int g \, d\lambda - G(\lambda) : \lambda \in \mathcal{J}(R : \nu) \right\}. \end{aligned}$$

This concludes the proof of Proposition 6. $\square$

9. Convergence of Gibbs posterior distributions. The purpose of this section is to establish Theorem 2 concerning convergence of the Gibbs posterior distributions to the solution set of a variational problem. From the dynamics point of view, this convergence highlights the role of the variational problem and the associated equilibirum joinings. We believe these objects to be worthy of further study. From the statistical point of view, this result describes the concentration of posterior distributions, which is of interest in any frequentist analysis of Bayesian methods. The proof follows somewhat directly from Theorem 1.

PROOF OF THEOREM 2. Let U be an open neighborhood of $\Theta_{\min}$. Let $F = \Theta \setminus U$, which is closed and therefore compact. If $\pi_0(F) = 0$, then $\pi_n(F|y) = 0$ for all n. Now suppose $\pi_0(F) > 0$, and let $\tilde{\pi}_0 = \pi_0(\cdot|F)$ be the conditional prior on F. Let V_* be the common value of $V(\theta)$ for $\theta \in \Theta_{\min}$. As $V : \Theta \to \mathbb{R}$ is lower semi-continuous and F is compact and disjoint from $\Theta_{\min}$, there exists $\epsilon > 0$ such that $\inf_{\theta \in F} V(\theta) \geq V_* + \epsilon$. Now we apply Theorem 1 in two ways: first, with the full parameter set Θ and prior π_0, and second, with F in place of

Θ and the conditional prior $\tilde{\pi}_0$ in place of π_0. Let Z_n^F denote the normalizing constant in the second case. Then for ν-almost every $y \in \mathcal{Y}$, there exists $N_1 = N_1(y)$ and $N_2 = N_2(y)$ such that for all $n \geq N_1$,

$$-\frac{1}{n}\log Z_n^F(y) \geq V_* + 2\epsilon/3,$$

and for all $n \geq N_2$,

$$-\frac{1}{n}\log Z_n(y) \leq V_* + \epsilon/3.$$

Then for all $n \geq \max(N_1, N_2)$, we have

$$\begin{aligned}
\pi_n(F|y) &= P_n(F \times \mathcal{X}|y) \\
&= \frac{1}{Z_n(y)} \int_{F\times\mathcal{X}} \exp\bigl(-\ell_n(\theta, x, y)\bigr)\, dP_0(\theta, x) \\
&= \frac{\pi_0(F) Z_n^F(y)}{Z_n(y)} \\
&\leq \exp\bigl(-V_* n - (2\epsilon/3)n + V_* n + (\epsilon/3)n\bigr) \\
&\leq \exp\bigl(-(\epsilon/3)n\bigr).
\end{aligned}$$

Thus, for ν-almost every $y \in \mathcal{Y}$, we see that $\pi_n(F|y)$ tends to 0. $\square$

10. Posterior consistency for hidden Gibbs processes. In this section we establish posterior consistency for hidden Gibbs processes, as described in Section 5.1. In addition to modeling substantial dependence with the underlying Gibbs processes, this setting also allows for quite general observational noise models. Note that hidden Markov models with arbitrarily large order appear as a special case in this framework. Here the first part of the proof involves an application of our main results to show that the posterior converges to the set $\Theta_{\min}$. However, the second part of the proof begins with the well-known fact that the Gibbs measures μ_θ satisfy large deviations principles (see [56]), and then relies on some recent results from [38] connecting these large deviations properties to the likelihood function in our general observational framework.

PROOF OF THEOREM 3. We begin by placing the setting of Section 5.1 within the general framework of Section 1. Let $\mathcal{X}$, $\{f_\theta : \theta \in \Theta\}$, $\{\mu_\theta : \theta \in \Theta\}$, Π_0, $\mathcal{U}$, m, and $\{\varphi_\theta(\cdot|x) : \theta \in \Theta, x \in \mathcal{X}\}$ be as in Section 5.1. To define the observation space in our general framework, we let $\mathcal{Y} = \mathcal{U}^{\mathbb{N}}$. We define the map $T : \mathcal{Y} \to \mathcal{Y}$ to be the left-shift, that is, if $y = \{y_k\} \in \mathcal{Y}$, then $T(y)$ is the sequence whose kth coordinate is y_{k+1}. Furthermore, we define $\nu = \mathbb{P}_{\theta^*}^U$, which is the process measure on $\mathcal{Y}$ described in Section 5.1. Then $(\mathcal{Y}, T, \nu)$ is an ergodic measure preserving system (see [38], Proposition 6.1, for ergodicity). Now define $\ell : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ by $\ell(\theta, x, \{y_k\}) = -\log \varphi_\theta(y_0|x)$. Note that the conditions (i)–(iii) on ℓ are satisfied by our assumptions on φ. Define $\pi_n(\cdot|y)$ to be the Gibbs posterior defined as in Section 1. Note that in this setting, if $y = \{y_k\}$, then the Gibbs posterior $\pi_n(\cdot|y)$ is equal to the standard posterior $\Pi_n(\cdot|y_0^{n-1})$. We require a few lemmas before finishing the proof of the theorem. Before we state the first such lemma, recall that ℓ^* denotes the ν-integrable function on $\mathcal{Y}$ appearing in property (ii) in Section 3.

LEMMA 6. *Let $\theta \in \Theta$. Then for each $n \geq 1$ and $y \in \mathcal{Y}$,*

$$\left|\frac{1}{n}\log \int_{\mathcal{X}} \exp\bigl(-\ell_n(\theta, x, y)\bigr)\, d\mu_\theta(x)\right| \leq \frac{1}{n}\sum_{k=0}^{n-1} \ell^*\bigl(T^k y\bigr).$$

PROOF. For notation, let

$$I_n(y) = \int_{\mathcal{X}} \exp\bigl(-\ell_n(\theta, x, y)\bigr)\, d\mu_\theta(x).$$

First suppose that $I_n(y) \leq 1$. Then by Jensen's inequality and the definition of ℓ^*,

$$\begin{aligned}\left|\frac{1}{n}\log I_n(y)\right| &= -\frac{1}{n}\log \int_{\mathcal{X}} \exp\bigl(-\ell_n(\theta, x, y)\bigr)\, d\mu_\theta(x) \\ &\leq \frac{1}{n}\int \ell_n(\theta, x, y)\, d\mu_\theta(x) \\ &= \frac{1}{n}\sum_{k=0}^{n-1}\int \ell\bigl(\theta, S^k x, T^k y\bigr)\, d\mu_\theta(x) \\ &\leq \frac{1}{n}\sum_{k=0}^{n-1} \ell^*\bigl(T^k y\bigr).\end{aligned}$$

Now suppose that $I_n(y) > 1$. Then

$$\begin{aligned}\left|\frac{1}{n}\log I_n(y)\right| &= \frac{1}{n}\log \int_{\mathcal{X}} \exp\bigl(-\ell_n(\theta, x, y)\bigr)\, d\mu_\theta(x) \\ &\leq \frac{1}{n}\log \sup_{x\in\mathcal{X}} \exp\bigl(-\ell_n(\theta, x, y)\bigr) \\ &\leq \frac{1}{n}\sum_{k=0}^{n-1} \sup_{x\in\mathcal{X}} \bigl|\ell\bigl(\theta, S^k x, T^k(y)\bigr)\bigr| \\ &\leq \frac{1}{n}\sum_{k=0}^{n-1} \ell^*\bigl(T^k y\bigr),\end{aligned}$$

where we have used that both the logarithm and the exponential are increasing. □

LEMMA 7. *Let $\theta \in \Theta$. Then*

$$\lim_n -\frac{1}{n}\mathbb{E}_{\theta^*}\bigl[\log p_\theta\bigl(Y_0^{n-1}\bigr)\bigr] = V(\theta).$$

PROOF. For each $n \geq 1$, let

$$f_n(y) = -\frac{1}{n}\log \int_{\mathcal{X}} \exp\bigl(-\ell_n(\theta, x, y)\bigr)\, d\mu_\theta(x),$$

and let $F_n(y) = n^{-1}\sum_{k=0}^{n-1} \ell^*(T^k y)$. By property (ii), ℓ^* is ν-integrable and thus the pointwise ergodic theorem ensures that $F_n(y)$ converges for ν-almost every y to the constant $\mathbb{E}_{\theta^*}[\ell^*]$. Furthermore, $\lim_n \mathbb{E}_{\theta^*}[F_n] = \mathbb{E}_{\theta^*}[\ell^*]$. By Lemma 6, $|f_n| \leq F_n$ for each $n \geq 1$. Therefore, by the generalized Lebesgue dominated convergence theorem and the definition of the loss,

$$\begin{aligned}\lim_n -\frac{1}{n}\mathbb{E}_{\theta^*}\bigl[\log p_\theta\bigl(Y_0^{n-1}\bigr)\bigr] &= \lim_n \mathbb{E}_{\theta^*}[f_n] \\ &= \mathbb{E}_{\theta^*}\Bigl[\lim_n f_n\Bigr].\end{aligned}$$

By Theorem 1, the $\mathbb{P}_{\theta^*}$-almost sure limit of $\{f_n\}$ is equal to $V(\theta)$. Combining these facts, we obtain the desired equality. □

LEMMA 8. *Suppose* $\theta \in \Theta \setminus [\theta^*]$. *Then*

$$V(\theta^*) < \lim_n -\frac{1}{n}\mathbb{E}_{\theta^*}[\log p_\theta(Y_0^{n-1})].$$

PROOF. The well-known large deviations principles for the Gibbs measures μ_θ [56] imply that they satisfy property (L1) from [38]. By hypothesis, g satisfies the regularity of observations property (L2) from [38]. Then results from [38] (in particular Propositions 4.3 and 6.4) yield the desired inequality. □

We now proceed with the proof of Theorem 3. Recall that for $y = \{y_k\} \in \mathcal{Y}$ our choice of loss function ensures that the Bayesian posterior $\Pi_n(\cdot|y_0^{n-1})$ is equal to the Gibbs posterior $\pi_n(\cdot|y)$. By Theorem 2, the Gibbs posterior $\pi_n(\cdot|Y)$ concentrates ν-almost surely around the set $\Theta_{\min}$, defined as the set of $\theta \in \Theta$ such that $V(\theta) = \inf\{V(\theta') : \theta' \in \Theta\}$. Hence $\Pi_n(\cdot|Y_0^{n-1})$ concentrates $\mathbb{P}_{\theta^*}^U$-almost surely around $\Theta_{\min}$. It remains to show that $\Theta_{\min} = [\theta^*]$.

Suppose $\theta \in \Theta \setminus [\theta^*]$. Then by Lemmas 7 and 8, we have

$$V(\theta) = \lim_n -\frac{1}{n}\mathbb{E}_{\theta^*}[\log p_\theta(Y_0^{n-1})] > V(\theta^*).$$

It follows immediately that $\Theta_{\min} \subset [\theta^*]$. For the reverse inclusion, note that if $\theta \in [\theta^*]$, then $\mathbb{P}_\theta^U = \mathbb{P}_{\theta^*}^U$, and thus for each n,

$$\mathbb{E}_{\theta^*}[\log p_\theta(Y_0^{n-1})] = \mathbb{E}_{\theta^*}[\log p_{\theta^*}(Y_0^{n-1})].$$

Then Lemma 7 gives that $V(\theta) = V(\theta^*)$ for each $\theta \in [\theta^*]$. This concludes the proof of Theorem 3. □

11. Discussion. The two main contributions of this paper are as follows: 1) showing that ideas developed in the thermodynamic formalism of dynamical systems can be used to provide guarantees on loss-based Bayesian inferential procedures, and 2) proving posterior consistency for inference of deterministic dynamical systems with observational noise and long range dependencies. Our results show that ideas from dynamical systems and ergodic theory are relevant for statistical inference, and that problems arising in statistical inference can suggest new and interesting questions in dynamical systems. It is of interest to explore how the variational formulation of the Gibbs posterior developed here may affect estimation procedures in the dynamical setting of this paper. Additionally, it is of interest to bridge the work on uncertainty quantification in stochastic differential equations with the results in this paper on inference for discrete dynamical systems. Last, we consider relaxation of the compactness assumptions in this work to be an interesting avenue for further investigation.

APPENDIX A: BACKGROUND AND TECHNICAL LEMMAS

A.1. Pressure lemma. We refer to the following elementary fact, which is an easy consequence of Jensen's inequality, as the Pressure Lemma; see [53], Lemma 9.9.

LEMMA 9. *Let* $a_1, \ldots, a_k$ *be real numbers. If* $p_i \geq 0$ *and* $\sum_{i=1}^k p_i = 1$, *then*

$$\sum_{i=1}^k p_i(a_i - \log p_i) \leq \log\left(\sum_{i=1}^k \exp(a_i)\right),$$

with equality if and only if

$$p_i = \frac{\exp(a_i)}{\sum_{j=1}^k \exp(a_i)}.$$

A.2. The space of joinings and the ergodic decomposition. Our proofs rely on a general version of the ergodic decomposition for invariant probability measures. The following version, a restatement of [47], Theorem 2.5, is sufficient for our purposes.

THEOREM ([47]). *Suppose that $R : \mathcal{U} \to \mathcal{U}$ is a Borel measurable map of a Polish space $\mathcal{U}$ and that $\mu \in \mathcal{M}(\mathcal{U}, R)$. Then there exists a Borel probability measure Q on $\mathcal{M}(\mathcal{U})$ such that*

(1) $Q(\{\eta \text{ is invariant and ergodic for } R\}) = 1$
(2) *If $f \in L^1(\mu)$, then $f \in L^1(\eta)$ for Q-almost every η, and*

$$\int f \, d\mu = \int \left(\int f \, d\eta \right) dQ(\eta).$$

Whenever (2) *holds, we write $\mu = \int \eta \, dQ$.*

Additionally, we require the following results about the structure of $\mathcal{J}(S : \nu)$ from [40].

THEOREM ([40]). *Suppose $R : \mathcal{U} \to \mathcal{U}$ is a continuous map of a compact metrizable space and $(\mathcal{Y}, T, \nu)$ is an ergodic measure-preserving system as in Section 1. Then $\mathcal{J}(R : \nu)$ is nonempty, compact, and convex. Furthermore, a joining $\lambda \in \mathcal{J}(R : \nu)$ is an extreme point of $\mathcal{J}(R : \nu)$ if and only if λ is ergodic for $R \times T$. Lastly, if $\lambda \in \mathcal{J}(R : \nu)$ and $\lambda = \int \eta \, dQ$ is its ergodic decomposition, then Q-almost every η is in $\mathcal{J}(R : \nu)$.*

Let $\lambda \in \mathcal{J}(R : \nu)$. By the above theorem, the ergodic decomposition of λ is a representation of λ as an integral combination of the extreme points of $\mathcal{J}(R : \nu)$. A function $F : \mathcal{J}(R : \nu) \to \mathbb{R}$ is called *harmonic* if for each $\lambda \in \mathcal{J}(R : \nu)$,

$$F(\lambda) = \int F(\eta) \, dQ(\eta),$$

where $\lambda = \int \eta \, dQ$ is the ergodic decomposition of λ.

A.3. Disintegration results. Suppose $R : \mathcal{U} \to \mathcal{U}$ is a continuous map of a compact metric space and $(\mathcal{Y}, T, \nu)$ is an ergodic system. It is well known in ergodic theory (see [21]) that for any joining $\lambda \in \mathcal{J}(R : \nu)$, if $\lambda = \int \lambda_y \otimes \delta_y \, d\nu(y)$ is its disintegration over ν, then the family of measures $\{\lambda_y\}_{y \in \mathcal{Y}}$ satisfies an additional invariance property, which we state in the following lemma.

LEMMA 10. *Let $\lambda \in \mathcal{J}(R : \nu)$, and let $\lambda = \int \lambda_y \otimes \delta_y \, d\nu(y)$ be its disintegration over ν. Then $(\lambda_y \otimes \delta_y) \circ (R \times T)^{-1} = \lambda_{Ty} \otimes \delta_{Ty}$ for ν-almost every $y \in \mathcal{Y}$, and hence, for every $f \in L^1(\lambda)$ and ν-almost every $y \in \mathcal{Y}$,*

$$\int f(Ru, Ty) \, d\lambda_y(u) = \int f(u, Ty) \, d\lambda_{Ty}(u).$$

A.4. Limiting average loss. The following lemma will be applied to the limiting average loss. Recall that when $R : \mathcal{U} \to \mathcal{U}$ is a continuous map of a compact metric space, the space $\mathcal{J}(R : \nu)$ of joinings is nonempty. For notation, if $f : \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$, then we let $f_n(u, y) = \sum_{k=0}^{n-1} f(R^k u, T^k y)$.

LEMMA 11. *Suppose that $R : \mathcal{U} \to \mathcal{U}$ is a continuous map of a compact metric space $\mathcal{U}$, and that $f : \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$ is a Borel function for which there exists $f^* : \mathcal{Y} \to \mathbb{R}$ in $L^1(\nu)$*

such that $\sup_{u\in U}|f(u,y)|\le f^*(y)$ *for each* $y\in\mathcal{Y}$. *Then for any joining* $\lambda\in\mathcal{J}(R:\nu)$ *with disintegration* $\lambda=\int\lambda_y\otimes\delta_y\,d\nu(y)$ *over* ν, *for* ν*-almost every* $y\in\mathcal{Y}$,

$$\lim_n\frac{1}{n}\int f_n(u,y)\,d\lambda_y(u)=\int f\,d\lambda.$$

PROOF. For $y\in\mathcal{Y}$ define $\tilde f(y)=\int f(u,y)\,d\lambda_y(u)$. Then $\tilde f\in L^1(\nu)$, since $f\in L^1(\lambda)$ (using the hypotheses involving f^*). Now Lemma 10, together with the pointwise ergodic theorem, yields that for ν almost every y,

$$\lim_n\frac{1}{n}\int f_n(u,y)\,d\lambda_y(u)=\lim_n\frac{1}{n}\sum_{k=0}^{n-1}\tilde f(T^k y)=\int\tilde f\,d\nu=\int f\,d\lambda.$$

□

A.5. Fiber entropy. We require two additional properties of the fiber entropy in our setting. The first property is that fiber entropy is harmonic. This fact appears with proof as Lemma 3.2 (iii) in [33] in a setting under which $T:\mathcal{Y}\to\mathcal{Y}$ is a continuous map of a compact space, but careful inspection shows that the proof does not depend on this hypothesis.

LEMMA 12. *The map* $\lambda\mapsto h^\nu(\lambda)$ *from* $\mathcal{J}(R:\nu)$ *to the nonnegative extended reals satisfies the following property*: *if* $\lambda=\int\eta\,dQ(\eta)$ *is the ergodic decomposition of* λ, *then*

$$h^\nu(\lambda)=\int h^\nu(\eta)\,dQ(\eta).$$

Next, we note that fiber entropy function is upper semi-continuous in our setting. The proof of Lemma 2.2 in [54] establishes upper semi-continuity of fiber entropy in a setting closely related to ours. By making only minor modifications of that proof, one may adapt it to our setting and prove the following lemma.

LEMMA 13. *Let* Θ, $(\mathcal{X},S)$, *and* $(\mathcal{Y},T,\nu)$ *be as in the introduction, and let* $R=I_\Theta\times S$ *act on the product space* $\mathcal{U}=\Theta\times\mathcal{X}$. *Then the map* $\lambda\mapsto h^\nu(\lambda)$ *from* $\mathcal{J}(R:\nu)$ *to* $\mathbb{R}$ *is upper semi-continuous.*

A.6. Divergence terms and average information. Recall that we have defined

$$L(\eta:\gamma|\alpha)=\begin{cases}\displaystyle\sum_{C\in\alpha}\eta(C)\log\gamma(C), & \text{if } \eta\prec_\alpha\gamma\\ -\infty, & \text{otherwise,}\end{cases}$$

where $0\cdot\log 0=0$ by convention. With these definitions, we always have

$$-KL(\eta:\gamma|\alpha)=H(\eta,\alpha)+L(\eta:\gamma|\alpha). \tag{14}$$

Recall that $H(\eta,\alpha)$ may be interpreted as the expected information of η under the partition α, where the expectation is with respect to η. In contrast, $-L(\eta:\gamma|\alpha)$ may be interpreted as the expected information of γ under the partition α, where the expectation is again taken with respect to η. In what follows, if α is a partition of a space $\mathcal{U}$ and $u\in\mathcal{U}$, we let $\alpha(u)$ denote the partition element containing u. Here we restate and then prove Lemma 1.

LEMMA 14. *Let* $f:\mathcal{X}\to\mathbb{R}$ *be a Hölder continuous potential on a mixing SFT* $\mathcal{X}$ *with associated Gibbs measure* μ. *Let* α *be the partition of* $\mathcal{X}$ *into cylinder sets of the form* $x[0]$, *and let* $\lambda\in\mathcal{J}(S:\nu)$ *be ergodic. Then for* ν*-almost every* $y\in\mathcal{Y}$,

$$\lim_n\frac{1}{n}KL(\lambda_y:\mu|\alpha_n)=\mathcal{P}(f)-\left(h^\nu(\lambda)+\int f\,d\lambda\right).$$

PROOF. Recall that by the Gibbs property for μ, for any $n \geq 1$ and x in $\mathcal{X}$, we have

$$K^{-1} \leq \frac{\mu(\alpha_n(x))}{-n\,\mathcal{P}(f) + \sum_{k=0}^{n-1} f \circ S^k(x)} \leq K.$$

Taking logarithms yields the bound

$$\left|\log\big(\mu(\alpha_n(x))\big) - \big(-n\,\mathcal{P}(f) + f_n(x)\big)\right| \leq \log K.$$

As this inequality is uniform in x, we may integrate with respect to λ_y to obtain

$$\left|L(\lambda_y : \mu|\alpha_n) - \left(-n\,\mathcal{P}(f) + \int f_n\, d\lambda_y\right)\right| \leq \log K.$$

Dividing by n and applying Lemma 11 gives

$$\lim_n \frac{1}{n} L(\lambda_y : \mu|\alpha_n) = -\mathcal{P}(f) + \int f\, d\lambda. \tag{15}$$

It follows from (10) that $K\,L(\lambda_y : \mu|\alpha_n) = -H(\lambda_y, \alpha_n) - L(\lambda_y : \mu|\alpha_n)$. Since λ is ergodic, for ν-almost every y, we have $n^{-1}H(\lambda_y, \alpha_n) \to h^\nu(\lambda, \alpha) = h^\nu(\lambda)$, where the equality is a result of the fact that α is a generating partition for $(\mathcal{X}, S)$. Combining this fact with (15), we find that for ν-almost every y,

$$\lim_n \frac{1}{n} K\,L(\lambda_y : \mu|\alpha_n) = -h^\nu(\lambda) - \left(-\mathcal{P}(f) + \int f\, d\lambda\right)$$

as desired. □

Now we prove a lemma that guarantees that $D(\lambda : \mu_\theta) \geq 0$.

LEMMA 15. *For each $\theta \in \Theta$ and $\lambda \in \mathcal{J}(S, \nu)$,*

$$\int f_\theta\, d\lambda + h^\nu(\lambda) \leq \mathcal{P}(f_\theta).$$

PROOF. Let μ be the $\mathcal{X}$-marginal of λ. Then $h^\nu(\lambda) \leq h^\nu(\mu \otimes \nu) = h(\mu)$, where the inequality follows from elementary information theoretic facts concerning conditional entropy (see [9]) and the equality is a basic property of fiber entropy. Then by the variational principle for pressure (7),

$$\int f_\theta\, d\lambda + h^\nu(\lambda) \leq \int f_\theta\, d\mu + h(\mu) \leq \mathcal{P}(f_\theta),$$

as desired. □

A.7. Lower semi-continuity of the rate function.

PROPOSITION 7. *The map $V : \Theta \to \mathbb{R}$ defined in Definition 4 is lower semi-continuous, and hence the set $\Theta_{\min}$ is compact and nonempty.*

PROOF. Let $\mathcal{U} = \Theta \times \mathcal{X}$ and let $R : \mathcal{U} \to \mathcal{U}$ be given by $R = I_\Theta \times S$, where I_Θ is the identity on Θ. Define $\psi : \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$ by

$$\psi(\theta, x, y) = -\ell(\theta, x, y) + f_\theta(x) - \mathcal{P}(f_\theta),$$

which is continuous and satisfies $\sup_{u \in \mathcal{U}} |\psi(u, y)| \leq \psi^* \in L^1(\nu)$. Finally, define $F : \mathcal{J}(R : \nu) \to \mathbb{R}$ by

$$F(\lambda) = \int \psi\, d\lambda + h^\nu(\lambda).$$

Since ψ is continuous and h^ν is upper semi-continuous (by Lemma 13), F is upper semi-continuous. Let $\mathrm{proj}_\Theta : \mathcal{J}(R:\nu) \to \mathcal{M}(\Theta)$ be defined by setting $\mathrm{proj}_\Theta(\lambda)$ to be the Θ-marginal of λ, which is a continuous surjection of compact spaces. One may easily check from the definition of upper semicontinuity that the function

$$\theta \mapsto \sup\{F(\lambda) : \mathrm{proj}_\Theta(\lambda) = \delta_\theta\}$$

is also upper semicontinuous. Since $V(\theta)$ is the negative of this function, we conclude that V is lower semi-continuous.

For the second part of the proposition, we note that $\Theta_{\min}$ is the argmin of the lower semi-continuous function V on the compact set Θ, and hence it is nonempty and compact. □

APPENDIX B: CONVERSE STATEMENT

In this section we collect some auxiliary results about Gibbs posterior inference. We begin with a converse to Theorem 2 on the exponential scale: if U is an open set intersecting $\Theta_{\min}$, then the Gibbs posterior measure of U cannot be exponentially small as n tends to infinity. This result appears as Proposition 8 below. First we establish a few preliminary lemmas.

LEMMA 16. *For all $\theta, \theta' \in \Theta$, for all $x \in \mathcal{X}$, and for all $n \geq 1$, we have*

$$K^{-2} e^{-n(|P(\theta)-P(\theta')|+\|f_\theta - f_{\theta'}\|)} \leq \frac{\mu_\theta[x_0^{n-1}]}{\mu_{\theta'}[x_0^{n-1}]} \leq K^2 e^{n(|P(\theta)-P(\theta')|+\|f_\theta - f_{\theta'}\|)}.$$

PROOF. Let θ, θ', x, and n be as above. By applying the Gibbs property (1) to both μ_θ and $\mu_{\theta'}$ and then simplifying, we have

$$\begin{aligned}\frac{\mu_\theta[x_0^{n-1}]}{\mu_{\theta'}[x_0^{n-1}]} &\leq \frac{K e^{-nP(\theta)+\sum_{k=0}^{n-1} f_\theta(S^k(x))}}{K^{-1} e^{-nP(\theta')+\sum_{k=0}^{n-1} f_{\theta'}(S^k(x))}} \\ &= K^2 e^{n(P(\theta')-P(\theta))+\sum_{k=0}^{n-1} f_{\theta'}(S^k(x))-f_\theta(S^k(x))}.\end{aligned}$$

By elementary estimates, we then have that

$$\begin{aligned}\frac{\mu_\theta[x_0^{n-1}]}{\mu_{\theta'}[x_0^{n-1}]} &\leq K^2 e^{n|P(\theta)-P(\theta')|+\sum_{k=0}^{n-1}|f_\theta(S^k x)-f_{\theta'}(S^k x)|} \\ &\leq K^2 e^{n|P(\theta)-P(\theta')|+n\|f_\theta - f_{\theta'}\|} \\ &= K^2 e^{n(|P(\theta)-P(\theta')|+\|f_\theta - f_{\theta'}\|)}.\end{aligned}$$

This estimate gives one of the inequalities in the conclusion of the lemma, and then interchanging the roles of θ and θ' gives the other. □

LEMMA 17. *Let $\delta > 0$ and $y \in \mathcal{Y}$. Suppose $\theta, \theta' \in \Theta$ and $x, x' \in \mathcal{X}$ satisfy $d_\theta(\theta, \theta') < \delta$ and $d_{\mathcal{X}}(S^k x, S^k x') < \delta$ for all $k = 0, \ldots, n-1$. Then*

$$\frac{e^{-\ell_n(\theta,x,y)}}{e^{-\ell_n(\theta',x',y)}} \leq e^{\sum_{k=0}^{n-1} \rho_\delta(T^k y)}.$$

PROOF. Let $\delta > 0$, $y \in \mathcal{Y}$, $\theta, \theta' \in \Theta$, $x, x' \in \mathcal{X}$, and $n \geq 1$ be as above. Note that for each $k = 0, \ldots, n-1$, we have that $d((\theta, S^k x), (\theta', S^k x')) < \delta$. Then

$$\frac{e^{-\ell_n(\theta,x,y)}}{e^{-\ell_n(\theta',x',y)}} = \frac{e^{-\sum_{k=0}^{n-1} \ell(\theta, S^k x, T^k y)}}{e^{-\sum_{k=0}^{n-1} \ell(\theta', S^k x', T^k y)}}$$

$$= e^{-\sum_{k=0}^{n-1} \ell(\theta, S^k x, T^k y) - \ell(\theta', S^k x', T^k y)}$$
$$\leq e^{\sum_{k=0}^{n-1} |\ell(\theta, S^k x, T^k y) - \ell(\theta', S^k x', T^k y)|}$$
$$\leq e^{\sum_{k=0}^{n-1} \rho_\delta(T^k y)},$$

which concludes the proof of the lemma. □

LEMMA 18. *Let $\delta > 0$ and $y \in \mathcal{Y}$. Suppose $\theta, \theta' \in \Theta$ satisfy $d_\Theta(\theta, \theta') < \delta$. Let $m \geq 1$ be such that if $x \in \mathcal{X}$ and $x' \in [x_0^{m-1}]$, then $d_\mathcal{X}(x, x') < \delta$. Then for any $n \geq 1$, we have*

$$\frac{\int_\mathcal{X} e^{-\ell_n(\theta,x,y)}\, d\mu_\theta(x)}{\int_\mathcal{X} e^{-\ell_n(\theta',x',y)}\, d\mu_{\theta'}(x')} \leq K^2 e^{(m+n)(|P(\theta)-P(\theta')| + \|f_\theta - f_{\theta'}\|) + \sum_{k=0}^{n-1} \rho_\delta(T^k y)}.$$

PROOF. Let $\delta > 0$, $y \in \mathcal{Y}$, $\theta, \theta' \in \Theta$, and $m \geq 1$ be as above. For each $n \geq 1$, let $\mathcal{L}_n$ denote the set of words $w \in \mathcal{A}^n$ such that there exists $x \in \mathcal{X}$ with $x_0^{n-1} = w$. For any such w, we let $\mu_{\theta,w}$ denote the conditional measure $\mu_\theta(\cdot|[w])$ obtained from μ_θ by conditioning on the cylinder set $[w]$. Then we have

$$\int_\mathcal{X} e^{-\ell_n(\theta,x,y)}\, d\mu_\theta(x) = \sum_{w \in \mathcal{L}_{m+n}} \mu_\theta([w]) \int_{[w]} e^{-\ell_n(\theta,x,y)}\, d\mu_{\theta,w}(x)$$
$$\leq \sum_{w \in \mathcal{L}_{m+n}} \mu_\theta([w]) \max_{x \in [w]} e^{-\ell_n(\theta,x,y)}.$$

By Lemma 17 and our choice of m, we have that

$$\max_{x \in [w]} e^{-\ell_n(\theta,x,y)} \leq e^{\sum_{k=0}^{n-1} \rho_\delta(T^k y)} \min_{x' \in [w]} e^{-\ell_n(\theta',x',y)}.$$

Combining the above inequalities and also applying Lemma 16, we obtain

$$\int_\mathcal{X} e^{-\ell_n(\theta,x,y)}\, d\mu_\theta(x) \leq \sum_{w \in \mathcal{L}_{m+n}} \mu_\theta([w]) \max_{x \in [w]} e^{-\ell_n(\theta,x,y)}$$
$$\leq e^{\sum_{k=0}^{n-1} \rho_\delta(T^k y)} \sum_{w \in \mathcal{L}_{m+n}} \mu_\theta([w]) \min_{x' \in [w]} e^{-\ell_n(\theta',x',y)}$$
$$\leq K^2 e^{(m+n)(|P(\theta)-P(\theta')| + \|f_\theta - f_{\theta'}\|) + \sum_{k=0}^{n-1} \rho_\delta(T^k y)}$$
$$\cdot \sum_{w \in \mathcal{L}_{m+n}} \mu_{\theta'}([w]) \int_{[w]} e^{-\ell_n(\theta',x',y)}\, d\mu_{\theta',w}(x')$$
$$= K^2 e^{(m+n)(|P(\theta)-P(\theta')| + \|f_\theta - f_{\theta'}\|) + \sum_{k=0}^{n-1} \rho_\delta(T^k y)}$$
$$\cdot \int_\mathcal{X} e^{-\ell_n(\theta',x',y)}\, d\mu_{\theta'}(x'),$$

as desired. □

PROPOSITION 8. *Suppose $U \subset \Theta$ is open and $U \cap \Theta_{\min} \neq \varnothing$. Then for ν-almost every $y \in \mathcal{Y}$,*

$$\lim_n \frac{1}{n} \log \pi_n(U|y) = 0.$$

PROOF. Let $\theta_0 \in U \cap \Theta_{\min}$. By definition of $\Theta_{\min}$ we have $V(\theta_0) = V_* = \inf_\theta V(\theta)$. Let $\epsilon > 0$ be arbitrary, and select $\delta > 0$ sufficiently small that

- $\int \rho_\delta \, d\nu < \epsilon/3$,
- if $d_\Theta(\theta, \theta') < \delta$, then $|P(\theta) - P(\theta)| < \epsilon/3$ and $\|f_\theta - f_{\theta'}\| < \epsilon/3$, and
- the ball U_0 of radius δ around θ_0 is contained in U.

Now choose $m \geq 1$ such that if $x \in \mathcal{X}$ and $x' \in [x_0^{m-1}]$, then $d_{\mathcal{X}}(x, x') < \delta$.

Since π_0 is fully supported, $\pi_0(U_0) > 0$. Note that for each $y \in \mathcal{Y}$ and $n \geq 1$,

$$\begin{aligned}\pi_n(U|y) &= \frac{1}{Z_n(y)} \int_U \int_{\mathcal{X}} \exp\bigl(-\ell_n(\theta, x, y)\bigr) \, d\mu_\theta(x) d\pi_0(\theta) \\ &\geq \frac{1}{Z_n(y)} \int_{U_0} \int_{\mathcal{X}} \exp\bigl(-\ell_n(\theta, x, y)\bigr) \, d\mu_\theta(x) d\pi_0(\theta).\end{aligned}$$

Applying Lemma 18 to θ_0 and any $\theta \in U_0$ and using our choice of δ, we see that

$$\begin{aligned}\int_{\mathcal{X}} \exp\bigl(-\ell_n(\theta, x, y)\bigr) \, d\mu_\theta(x) &\geq K^{-2} \exp\bigl(-(m+n)\bigl(|P(\theta) - P(\theta_0)| + \|f_\theta - f_{\theta_0}\|\bigr)\bigr) \\ &\quad \cdot \int_{\mathcal{X}} \exp\bigl(-\ell_n(\theta_0, x', y)\bigr) \, d\mu_{\theta_0}(x') \\ &\geq K^{-2} \exp\bigl(-(m+n)(\epsilon/3 + \epsilon/3)\bigr) \\ &\quad \cdot \int_{\mathcal{X}} \exp\bigl(-\ell_n(\theta_0, x', y)\bigr) \, d\mu_{\theta_0}(x').\end{aligned}$$

Combining the above inequalities, we have

$$\begin{aligned}\pi_n(U|y) &\geq \frac{1}{Z_n(y)} \int_{U_0} \int_{\mathcal{X}} \exp\bigl(-\ell_n(\theta, x, y)\bigr) \, d\mu_\theta(x) d\pi_0(\theta) \\ &\geq \frac{1}{Z_n(y)} K^{-2} \exp\left(-2(m+n)\epsilon/3 - \sum_{k=0}^{n-1} \rho_\delta(T^k y)\right) \\ &\quad \cdot \int_{U_0} \int_{\mathcal{X}} \exp\bigl(-\ell_n(\theta_0, x', y)\bigr) \, d\mu_{\theta_0}(x') \, d\pi_0(\theta) \\ &\geq \frac{1}{Z_n(y)} K^{-2} \exp\left(-2(m+n)\epsilon/3 - \sum_{k=0}^{n-1} \rho_\delta(T^k y)\right) \\ &\quad \cdot \int_{\mathcal{X}} \exp\bigl(-\ell_n(\theta_0, x', y)\bigr) \, d\mu_{\theta_0}(x') \cdot \pi_0(U_0).\end{aligned}$$

Taking logarithms, dividing by n, and letting n tend to infinity yields

$$\liminf_n \frac{1}{n} \log \pi_n(U|y) \;\geq\; V_* - 2\epsilon/3 - \int \rho_\delta \, d\nu - V_* \geq -\epsilon.$$

(Note that the above inequality holds for ν-almost every y.) As $\epsilon > 0$ was arbitrary, we obtain the desired result. □

APPENDIX C: CONVERGENCE OF THE POSTERIOR

We now address the Cesàro convergence of the full posterior P_n on $\Theta \times \mathcal{X}$. Recall that we let $I_\Theta : \Theta \to \Theta$ be the identity map on Θ. In the thermodynamic formalism, invariant measures that achieve the optimal value in the variational expression for pressure are called equilibrium measures. In our setting, we introduce terminology for joinings that achieve the optimal value in the variational expression for the rate function. We will call a joining $\lambda \in \mathcal{J}(I_\Theta \times S : \nu)$ an equilibrium joining if

$$\lambda \in \operatorname{argmin}\left\{\int \ell \, d\lambda' + G(\lambda') : \lambda' \in \mathcal{J}(I_\Theta \times S : \nu)\right\}.$$

PROPOSITION 9. *For each $y \in \mathcal{Y}$ and $n \geq 1$, let $Q_n(\cdot|y) \in \mathcal{M}(\Theta \times \mathcal{X})$ be defined for Borel sets $E \subset \Theta \times \mathcal{X}$ by*

$$Q_n(E|y) = \frac{1}{n} \sum_{k=0}^{n-1} P_n\big((I_\Theta \times S)^{-k} E|y\big).$$

Then for ν-almost every $y \in \mathcal{Y}$, all limit points of $\{Q_n(\cdot|y)\}_{n\geq 1}$ are $(\Theta \times \mathcal{X})$-marginals of equilibrium joinings.

PROOF. As in Section 8.2, let

$$\eta_n = \frac{1}{n} \sum_{k=0}^{n-1} \big(P_n^y \circ (I_\Theta \times S)^{-k}\big) \otimes \delta_{T^k y}.$$

By definition, $Q_n(\cdot|y)$ is the $(\Theta \times \mathcal{X})$-marginal of η_n. Let Q be a weak limit of the subsequence $\{Q_{n_k}(\cdot|y)\}_{k\geq 1}$. By repeating the arguments of Section 8.2, one may show that there is a subsequence $\{n_{k_j}\}_{j\geq 1}$ such that $\{\eta_{n_{k_j}}\}_{j\geq 1}$ converges weakly to an equilibrium joining λ. As Q is necessarily the $(\Theta \times \mathcal{X})$-marginal of the limit λ, the proof is complete. □

Funding. KM acknowledges the support of National Science Foundation grants DMS-1613261 and DMS-1847144. SM Acknowledges support from National Science Foundation grants DEB-1840223, DMS 17-13012, and DMS 16-13261, as well as National Institutes of Health grant R01 DK116187-01 and Human Frontier Science Program grant RGP0051/2017. ABN acknowledges support from National Science Foundation grants DMS-1613261 and NSF DMS-1613072, as well as National Institutes of Health grant R01 HG009125-01.

REFERENCES

[1] ALVES, J. F., RAMOS, V. and SIQUEIRA, J. (2019). Equilibrium stability for non-uniformly hyperbolic systems. *Ergodic Theory Dynam. Systems* **39** 2619–2642. MR4000509 https://doi.org/10.1017/etds.2017.138

[2] BELITSER, E. and NURUSHEV, N. (2019). General framework of projection structures. Preprint. Available at arXiv:1904.01003.

[3] BENEDICKS, M. and YOUNG, L.-S. (2000). Markov extensions and decay of correlations for certain Hénon maps. *Astérisque* 13–56. MR1755436

[4] BISSIRI, P. G., HOLMES, C. C. and WALKER, S. G. (2016). A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 1103–1130. MR3557191 https://doi.org/10.1111/rssb.12158

[5] BOWEN, R. (1975). *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms. Lecture Notes in Mathematics, Vol.* 470. Springer, Berlin. MR0442989

[6] CATONI, O. (2007). *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **56**. IMS, Beachwood, OH. MR2483528

[7] CHAZOTTES, J.-R., FLORIANI, E. and LIMA, R. (1998). Relative entropy and identification of Gibbs measures in dynamical systems. *J. Stat. Phys.* **90** 697–725. MR1616918 https://doi.org/10.1023/A:1023220802597

[8] CHOPIN, N., GADAT, S., GUEDJ, B., GUYADER, A. and VERNET, E. (2015). On some recent advances on high dimensional Bayesian statistics. In *Modélisation Aléatoire et Statistique—Journées MAS* 2014. *ESAIM Proc. Surveys* **51** 293–319. EDP Sci., Les Ulis. MR3440803 https://doi.org/10.1051/proc/201551016

[9] COVER, T. M. and THOMAS, J. A. (2012). *Elements of Information Theory*. Wiley Interscience, Hoboken, NJ.

[10] DE LA RUE, T. (2006). An introduction to joinings in ergodic theory. *Discrete Contin. Dyn. Syst.* **15** 121–142. MR2191388 https://doi.org/10.3934/dcds.2006.15.121

[11] DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1–67. MR0829555 https://doi.org/10.1214/aos/1176349830

[12] DIACONIS, P. W. and FREEDMAN, D. (1998). Consistency of Bayes estimates for nonparametric regression: Normal theory. *Bernoulli* **4** 411–444. MR1679791 https://doi.org/10.2307/3318659
[13] DOOB, J. L. (1949). Application of the theory of martingales. In *Le Calcul des Probabilités et Ses Applications*. *Colloques Internationaux du Centre National de la Recherche Scientifique*, *No*. 13 23–27. Centre National de la Recherche Scientifique, Paris. MR0033460
[14] DOUC, R., OLSSON, J. and ROUEFF, F. (2020). Posterior consistency for partially observed Markov models. *Stochastic Process. Appl.* **130** 733–759. MR4046518 https://doi.org/10.1016/j.spa.2019.03.012
[15] DUPUIS, P. and ELLIS, R. S. (1997). *A Weak Convergence Approach to the Theory of Large Deviations*. *Wiley Series in Probability and Statistics*: *Probability and Statistics*. Wiley, New York. MR1431744 https://doi.org/10.1002/9781118165904
[16] FURSTENBERG, H. (1967). Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation. *Math. Syst. Theory* **1** 1–49. MR0213508 https://doi.org/10.1007/BF01692494
[17] GASSIAT, E. and ROUSSEAU, J. (2014). About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli* **20** 2039–2075. MR3263098 https://doi.org/10.3150/13-BEJ550
[18] GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.
[19] GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782 https://doi.org/10.1017/9781139029834
[20] GHOSH, J. K. and RAMAMOORTHI, R. V. (2003). *Bayesian Nonparametrics*. *Springer Series in Statistics*. Springer, New York. MR1992245
[21] GLASNER, E. (2003). *Ergodic Theory via Joinings*. *Mathematical Surveys and Monographs* **101**. Amer. Math. Soc., Providence, RI. MR1958753 https://doi.org/10.1090/surv/101
[22] GRÜNWALD, P. and VAN OMMEN, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal.* **12** 1069–1103. MR3724979 https://doi.org/10.1214/17-BA1085
[23] GRÜNWALD, P. D. and MEHTA, N. A. (2020). Fast rates for general unbounded loss functions: From ERM to generalized Bayes. *J. Mach. Learn. Res.* **21** Paper No. 56, 80. MR4095335
[24] HANG, H. and STEINWART, I. (2017). A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *Ann. Statist.* **45** 708–743. MR3650398 https://doi.org/10.1214/16-AOS1465
[25] JAYNES, E. T. (1973). The well-posed problem. *Found. Phys.* **3** 477–492. MR0426227 https://doi.org/10.1007/BF00709116
[26] JAYNES, E. T. (1968). Prior probabilities. *IEEE Trans. Syst. Sci. Cybern.* **4** 227.
[27] JIANG, W. and TANNER, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist.* **36** 2207–2231. MR2458185 https://doi.org/10.1214/07-AOS547
[28] KIFER, Y. (2001). On the topological pressure for random bundle transformations. In *Topology, Ergodic Theory, Real Algebraic Geometry*. *Amer. Math. Soc. Transl. Ser.* 2 **202** 197–214. Amer. Math. Soc., Providence, RI. MR1819189 https://doi.org/10.1090/trans2/202/14
[29] KIFER, Y. and LIU, P.-D. (2006). Random dynamics. In *Handbook of Dynamical Systems*. *Vol.* 1*B* 379–499. Elsevier, Amsterdam. MR2186245 https://doi.org/10.1016/S1874-575X(06)80030-5
[30] LALLEY, S. P. (1999). Beneath the noise, chaos. *Ann. Statist.* **27** 461–479. MR1714721 https://doi.org/10.1214/aos/1018031203
[31] LALLEY, S. P. and NOBEL, A. B. (2006). Denoising deterministic time series. *Dyn. Partial Differ. Equ.* **3** 259–279. MR2271730 https://doi.org/10.4310/DPDE.2006.v3.n4.a1
[32] LAW, K., STUART, A. and ZYGALAKIS, K. (2015). *Data Assimilation*: *A Mathematical Introduction*. *Texts in Applied Mathematics* **62**. Springer, Cham. MR3363508 https://doi.org/10.1007/978-3-319-20325-6
[33] LEDRAPPIER, F. and WALTERS, P. (1977). A relativised variational principle for continuous transformations. *J. Lond. Math. Soc.* (2) **16** 568–576. MR0476995 https://doi.org/10.1112/jlms/s2-16.3.568
[34] LI, F. and ZHANG, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *J. Amer. Statist. Assoc.* **105** 1202–1214. MR2752615 https://doi.org/10.1198/jasa.2010.tm08177
[35] LIND, D. and MARCUS, B. (1995). *An Introduction to Symbolic Dynamics and Coding*. Cambridge Univ. Press, Cambridge. MR1369092 https://doi.org/10.1017/CBO9780511626302
[36] LU, Y., STUART, A. and WEBER, H. (2017). Gaussian approximations for transition paths in Brownian dynamics. *SIAM J. Math. Anal.* **49** 3005–3047. MR3684411 https://doi.org/10.1137/16M1071845
[37] MCALLESTER, D. A. (1999). Some PAC-Bayesian theorems. *Machine Learning* **37** 355–363.
[38] MCGOFF, K., MUKHERJEE, S., NOBEL, A. and PILLAI, N. (2015). Consistency of maximum likelihood estimation for some dynamical systems. *Ann. Statist.* **43** 1–29. MR3285598 https://doi.org/10.1214/14-AOS1259

[39] MCGOFF, K., MUKHERJEE, S. and PILLAI, N. (2015). Statistical inference for dynamical systems: A review. *Stat. Surv.* **9** 209–252. MR3422438 https://doi.org/10.1214/15-SS111

[40] MCGOFF, K. and NOBEL, A. B. Empirical risk minimization for dynamical systems and stationary processes. Information and Inference. Preprint. Available at arXiv:1601.05033.

[41] MCGOFF, K. and NOBEL, A. B. (2020). Empirical risk minimization and complexity of dynamical models. *Ann. Statist.* **48** 2031–2054. MR4134785 https://doi.org/10.1214/19-AOS1876

[42] MILLER, J. W. and DUNSON, D. B. (2019). Robust Bayesian inference via coarsening. *J. Amer. Statist. Assoc.* **114** 1113–1125. MR4011766 https://doi.org/10.1080/01621459.2018.1469995

[43] MITTER, S. K. and NEWTON, N. J. (2003). A variational approach to nonlinear estimation. *SIAM J. Control Optim.* **42** 1813–1833. MR2046387 https://doi.org/10.1137/S0363012901393894

[44] MOORES, M., NICHOLLS, G., PETTITT, A. N. and MENGERSEN, K. (2020). Scalable Bayesian inference for the inverse temperature of a hidden Potts model. *Bayesian Anal.* **15** 1–27. MR4050875 https://doi.org/10.1214/18-BA1130

[45] NEWTON, N. J. and MITTER, S. K. (2010). Variational Bayes and a problem of reliable communication: II. Infinite systems. *J. Stat. Mech. Theory Exp.* **2012** P11008.

[46] RUELLE, D. (2004). *Thermodynamic Formalism*: *The Mathematical Structures of Equilibrium Statistical Mechanics*, 2nd ed. *Cambridge Mathematical Library*. Cambridge Univ. Press, Cambridge. MR2129258 https://doi.org/10.1017/CBO9780511617546

[47] SARIG, O. (2008). Lecture Notes on Ergodic Theory.

[48] SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10–26. MR0184378 https://doi.org/10.1007/BF00535479

[49] SHALIZI, C. R. (2009). Dynamics of Bayesian updating with dependent data and misspecified models. *Electron. J. Stat.* **3** 1039–1074. MR2557128 https://doi.org/10.1214/09-EJS485

[50] STEINWART, I. and ANGHEL, M. (2009). Consistency of support vector machines for forecasting the evolution of an unknown ergodic dynamical system from observations with unknown noise. *Ann. Statist.* **37** 841–875. MR2502653 https://doi.org/10.1214/07-AOS562

[51] VERNET, E. (2015). Posterior consistency for nonparametric hidden Markov models with finite state space. *Electron. J. Stat.* **9** 717–752. MR3331855 https://doi.org/10.1214/15-EJS1017

[52] VOLODIMIR, G. V. (1990). Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, *COLT*'90 371–386. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[53] WALTERS, P. (1982). *An Introduction to Ergodic Theory*. *Graduate Texts in Mathematics* **79**. Springer, New York–Berlin. MR0648108

[54] WALTERS, P. (1986). Relative pressure, relative equilibrium states, compensation functions and many-to-one codes between subshifts. *Trans. Amer. Math. Soc.* **296** 1–31. MR0837796 https://doi.org/10.2307/2000558

[55] YANG, Y. and TOKDAR, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.* **43** 652–674. MR3319139 https://doi.org/10.1214/14-AOS1289

[56] YOUNG, L.-S. (1990). Large deviations in dynamical systems. *Trans. Amer. Math. Soc.* **318** 525–543. MR0975689 https://doi.org/10.2307/2001318

[57] YOUNG, L.-S. (1998). Statistical properties of dynamical systems with some hyperbolicity. *Ann. of Math.* (2) **147** 585–650. MR1637655 https://doi.org/10.2307/120960

[58] YOUNG, L.-S. (1999). Recurrence times and rates of mixing. *Israel J. Math.* **110** 153–188. MR1750438 https://doi.org/10.1007/BF02808180

[59] ZELLNER, A. (1988). Optimal information processing and Bayes's theorem. *Amer. Statist.* **42** 278–284. MR0971095 https://doi.org/10.2307/2685143

[60] ZHANG, T. (2006). From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *Ann. Statist.* **34** 2180–2210. MR2291497 https://doi.org/10.1214/009053606000000704

[61] ZHANG, T. (2006). Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans. Inf. Theory* **52** 1307–1321. MR2241190 https://doi.org/10.1109/TIT.2005.864439

[62] ZOU, Z., MUKHERJEE, S., ANTIL, H. and AQUINO, W. (2019). Adaptive particle-based approximations of the Gibbs posterior for inverse problems. Preprint. Available at arXiv:1904.01003.